ISSN 2995-5688 (Print)
ISSN 2995-570X (Online)

Global Academic Frontiers

Volume 3 · Issue 3 · September 2025

Published by Editorial Office of Global Academic Frontiers

Global Academic Frontiers

Volume 3 • Issue 3 • September 2025

(Quarterly, Published Since 2023)

Publisher	Editorial Office of Global Academic Frontiers
Place of publication	United States
Website	http://gafj.org
Email	office@gafj.org
Editor-in-Chief	Cunpeng Wu
Office number	+1 818-936-4444
Mailing Address	15617 NE Airport Way, Multnomah Mailbox SHKHVD Portland, OR, USA 97230
Printer	Editorial Office of Global Academic Frontiers self-prints on the Lulu Press
	A4 Paper, Version 1, Printing September 2025
Subscriptions	http://gafj.org/journal
Price	Free Copy
	ISSN: 2995-5688（Print）
	ISSN: 2995-570X（Online）
	International Standard Identifier for Libraries: OCLC-GAFEO
	ISNI: 0000000517857833

TABLE OF CONTENTS

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 1-13 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018216

Understanding Wage Disparities Through Educational Attainment

Jiayi Zhang [1], Xiaoran Han [1*]

[1] College of Art and Sciences, Boston University, Boston, MA, USA
*Corresponding author Email: chrishan0318@163.com
Received 20 May 2025; Accepted 11 July 2025; Published 1 September 2025

Abstract: This paper investigates the causal effect of education on wages using data from the 2018 American Community Survey (ACS) accessed through IPUMS. Our paper tests the causal effect of educational attainment on wages using individual-level data by using quarter of birth as an instrumental variable for education. We hypothesize that higher education leads to significantly higher wages. Based on our IV regression model, we find that each additional year of education is associated with approximately a 20.8% increase in log wages. These results support the importance of education in explaining wage differences, although other factors such as gender also play a notable role.

Keywords: Education, Wage Disparity, Instrumental Variables, American Community Survey, Labor Economics

I. Introduction

The relationship between educational attainment and labor market outcomes has been a central focus in labor economics. According to the foundational theory of human capital proposed by Becker (1964), individuals invest in education to increase their productivity, which in turn leads to higher earnings. Mincer (1974) further formalized this relationship through a widely adopted earnings function that regresses log wages on years of schooling and labor market experience. While the theoretical linkage is intuitive, empirical estimation of the causal effect of education on earnings is complicated by endogeneity concerns. Unobserved factors such as innate ability, family background, or motivation may influence both educational attainment and labor market success, biasing ordinary least squares (OLS) estimates.

To address this challenge, we adopt an instrumental variables (IV) approach using quarter of birth as an instrument for education, following the methodology of Angrist and Krueger (1991). The quarter-of-birth instrument exploits the exogenous variation created by compulsory schooling laws and school entry age cutoffs, which affect how long individuals stay in school without being correlated with innate ability or motivation. This approach allows us to better isolate the true causal impact of education on wages.

We use microdata from the 2018 American Community Survey (ACS) accessed via IPUMS to estimate the effect of educational attainment on log wages. To further reduce omitted variable bias, we include a range of controls such as age, gender, race, and hours worked per week. In our IV regression model, we find that each additional year of education is associated with a significant increase in log wages, suggesting a strong causal relationship. This finding contributes to the growing empirical literature on wage determination and underscores the policy relevance of education as a lever for income mobility and labor market equity. Moreover, we discuss how gender and other demographic factors intersect with education in explaining wage disparities.

II. Literature Review

A substantial body of literature has documented a positive relationship between education and earnings. Early empirical studies, including those by Mincer (1974) and Becker (1964), posited that each additional year of schooling increases an individual's productivity and thus their market wage. These models laid the foundation for estimating the returns to education using regression-based approaches. More recently, Card (1999) conducted a comprehensive review of the empirical evidence and concluded that the returns to education are both economically and statistically significant, though estimates vary depending on methods and data sources.

However, a key concern in estimating the return to education is endogeneity bias. Individuals with higher innate ability or better socioeconomic backgrounds may both attain more education and earn higher wages, leading to an upward bias in OLS estimates. To address this, several studies have turned to instrumental variables. Among the most influential is the work of Angrist and Krueger (1991), who used quarter of birth as an instrument for years of education. They argued that individuals born earlier in the year typically start school later and are therefore more likely to leave school earlier, leading to small but meaningful differences in educational attainment. Their IV estimates of the return to schooling—ranging from 8% to 13% per additional year—were higher than their OLS counterparts, emphasizing the importance of correcting for endogeneity.

Complementary research by Oreopoulos (2006) used changes in compulsory schooling laws in the United Kingdom and Canada as instruments and found large and robust returns to education. He also showed that these returns were particularly high for individuals from disadvantaged backgrounds, reinforcing the role of education in promoting economic mobility. Similarly, Sorel and Shinners (2019) analyzed data from Georgia using multiple regression models and found that each level of educational attainment was associated with a 12.6% increase in wages in a simple linear model, but the effect decreased to 5.7% after including demographic controls—highlighting the influence of confounding variables such as gender and race.

Furthermore, Heckman, Lochner, and Todd (2006) emphasized that while education is a strong predictor of earnings, non-cognitive skills, early childhood investments, and family environments also significantly contribute to labor market success. These findings suggest that policies aimed solely at increasing educational attainment may not fully address wage disparities unless they also consider broader social and economic factors.

Taken together, the literature demonstrates a consistent and substantial return to education, though the magnitude of the effect depends critically on the estimation strategy. Our study contributes to this ongoing discourse by applying a well-established IV methodology to recent ACS data, thereby providing updated evidence on the causal impact of education on earnings in the United States.

III. Data

A. Source of Data

The analysis uses data from the 2018 American Community Survey (ACS), accessed through IPUMS. The ACS is a nationally representative annual survey conducted by the U.S. Census Bureau, designed to collect detailed demographic, social, economic, and housing information. For our analysis, we use a 1% sample of the 2018 ACS and focus on variables relevant to wage determination, including age, sex, race, educational attainment (both general and detailed versions), usual hours worked per week, and wage and salary income. We also include the individual's quarter of birth as an instrumental variable to address potential endogeneity in education. This rich microdata allows for robust regression analysis of the relationship between educational attainment and earnings.

Table 1. Summary of the Variables

Variable	Obs	Mean	Std. Dev.	Min	Max

incwage	1,548,402	52,613.331	65,699.461	4	718,000
educyrs	1,548,402	14.000	2.955	0	21
age	1,548,402	43.18	15.061	18	96
female	1,548,402	0.483	0.500	0	1
uhrswork	1,548,402	38.774	12.567	1	99
NonWhite	1,548,402	0.798	0.402	0	1

To supplement the descriptive statistics table, we highlight a few key observations about the minimum and maximum values of our dependent variable, income (incwage), and key regressor, years of education (educyrs). The raw wage data ranges from $4 to $718,000, with a mean of $52,613.33 and a standard deviation of $65,699.46. This large variation and extreme upper bound suggest the presence of outliers or a skewed distribution, which is why we later transform income into a natural logarithm (lnwage)to improve interpretability and reduce the influence of extreme values. Similarly, educyrs ranges from 0 to 21, corresponding to the full educational spectrum from no schooling to doctoral degrees, as classified by the IPUMS detailed education codes. We excluded records with education codes labeled as "N/A" or "missing" and mapped each valid category into equivalent years of schooling to construct a continuous variable. The minimum of 0 reflects individuals with no formal education or only reached kindergarten, while the maximum of 21 corresponds to those holding doctoral degrees. Additionally, the age variable ranges from 18 to 96, with a mean of 43.18, ensuring our sample includes only working-age adults. These cleaned and transformed variables provide a more reliable basis for the regression analysis and ensure that extreme or non-informative values do not distort our results.

B. Models and Results

The scatter plot in Figure 1 depicts the relationship between educational attainment(educyrs) and the natural logarithm of wage (lnwage). Each dot represents an observation, and the red line represents the fitted values from a simple linear regression. From the plot, we observe a generally positive trend: as education level increases, the log of wages tends to increase as well. This supports the human capital theory that higher education leads to higher earnings. However, the relationship is not perfectly linear—there is considerable variation in wages within each education level, especially at lower education levels. Despite this dispersion, the upward slope of the fitted line indicates that on average, each additional level of education is associated with higher logged wages. The fitted line provides a reasonable linear approximation of the underlying pattern, justifying the use of a linear regression model for this analysis.

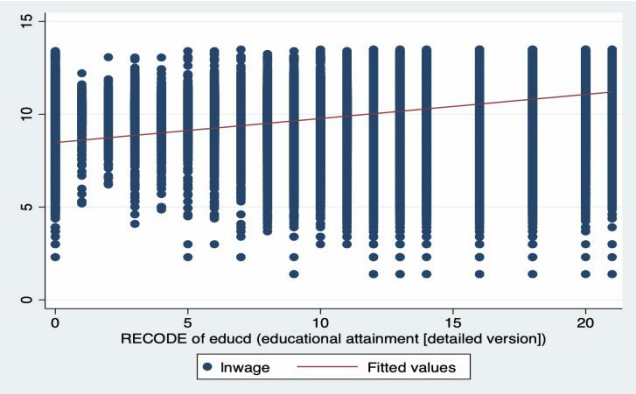


Figure 1. Scatter Plot of Logged-Wage versus Education

IV. Econometric Analysis

A. Simple Regression Model Selection

In both models, the slope coefficient on *educyrs* is highly statistically significant ($p < 0.001$), indicating a strong relationship between education and wages. The positive coefficients indicate that more years of education are associated with higher earnings, aligning with human capital theory. In the raw income model, the coefficient of 6984.07 means that each additional year of education is associated with an average increase of about $6,984 in annual income, holding other factors constant. In the log-linear model, the coefficient of 0.1252 suggests that a 1-year increase in education is associated with approximately a 12.52% increase in wages, interpreting it as a semi-elasticity since the dependent variable is in logs but the regressor is in levels. Psacharopoulos and Patrinos (2018) provided extensive cross-country evidence showing that returns to education remain consistently high, particularly in developing countries, and that each additional year of schooling significantly boosts income. Their findings reinforce the reliability of our result and its policy relevance.

Both models have relatively low R-squared values, though the log-linear model is slightly lower (0.0883 vs. 0.0996). However, after comparing the distribution of the raw wage variable (*incwage*) to its logarithmic transformation (*lnwage*), Figure 2 presents the histogram of raw wages, which is highly right skewed with a long tail extending toward higher income values.

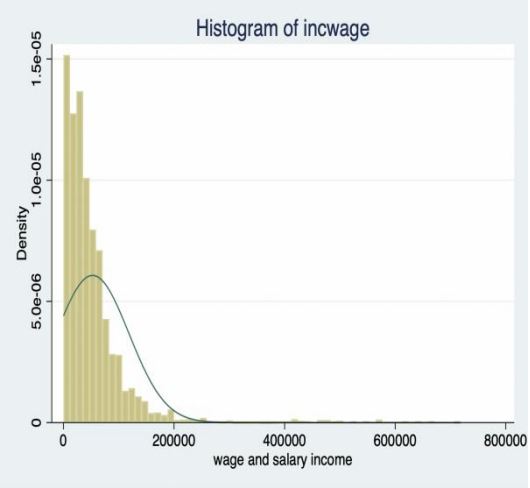


Figure 2. Histogram of Unlogged Wage

The distribution exhibits extreme values, with the maximum reaching $718,000, and a large proportion of observations concentrated at the lower end. Such skewness violates the normality assumption underpinning classical linear regression models and can lead to inefficient and biased estimates. Biewen and Fitzenberger (2005) emphasized that log-transforming skewed wage variables helps achieve a closer approximation to normality and homoscedasticity, leading to more efficient estimation in earnings regressions.

In contrast, Figure 3 displays the histogram of logged wages (lnwage).

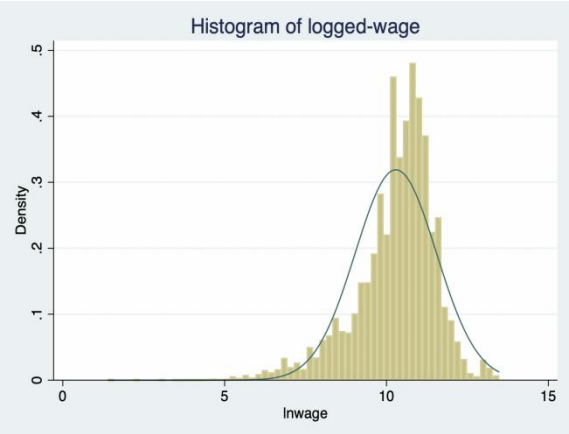


Figure 3. Histogram of Logged Wage

After log transformation, the distribution appears significantly more symmetric and bell-shaped, closely resembling the normal distribution. The transformation compresses the scale of wages and reduces the influence of outliers, resulting in a more homoscedastic variance structure. Additionally, using the log of wage facilitates elasticity-based interpretations of regression coefficients, which are common in labor economics research. Based on this comparison, we adopt the logged wage (lnwage) as our dependent variable in all subsequent regression analyses. Therefore, our equation of the simple regression model is shown as below:

$$\log(wage)=\beta 0 + \beta 1 \cdot educyrs + u, \quad (1)$$

$$\log(wage)=8.4734+0.1297 \cdot educyrs. \quad (2)$$

B. Other control variables

The inclusion of our controls can help with omitted variable bias. First off, age serves as a proxy for labor market experience, which tends to increase with time and often correlates positively with both educational attainment and income. Failing to control for age could lead to an inflated estimate of the education coefficient, as older individuals might command higher wages due to experience rather than schooling per se. Secondly, gender is included to capture sex-based wage disparities that, if unaccounted for, could confound the relationship between education and income. Similarly, the NonWhite variable helps account for racial differences in access to educational and occupational opportunities, which may systematically affect wage outcomes. Finally, uhrswork reflects the intensity of labor supply. Because wages are partially determined by the number of hours worked, controlling for this variable helps isolate the effect of education from variation in work effort. Collectively, these controls improve model specification and help ensure that the estimated return to education is not driven by omitted demographic or behavioral factors.

Appendix. Table 6 presents the pairwise correlation coefficients among the variables included in the wage regression model: log wages (lnwage), years of education (educyrs), age, race (NonWhite), gender (female), and usual hours worked per week (uhrswork). The matrix indicates that there is no evidence of perfect collinearity among any pair of variables. Even though we have a coefficient of 0.586 between lnwage and uhrswork, indicating that hours worked is a major determinant of income. However, in our VIF result (Appendix Table 7) all the control variables show VIF less than 5, which is an indication of non-multicollinearity in the model.

C. Multivariate Regression Model with Interaction Term

Below is the multivariate regression function we got from Appendix Table 8:

$$\log(wage)=\beta 0 + \beta 1\ educyrs + \beta 2\ age + \beta 3\ female + \beta 4\ NonWhite + \beta 5\ uhrswork + \beta 6\ femaleuhrswork + u, \quad (3)$$

$$\log(wage)=6.371+0.097\cdot educyrs+0.015\cdot age-0.489\cdot female+0.057\cdot NonWhite+0.049\cdot uhrswork+0.009\cdot femaleuhrswork, \quad (4)$$

In the empirical model presented, the interaction term femaleuhrswork captures how the effect of usual hours worked per week on log wages differs by gender. The estimated coefficient on this interaction is 0.0089, and it is statistically significant at the 1% level, indicating a robust relationship. This positive coefficient suggests that the marginal return to an additional hour worked per week is slightly higher for women compared to men. Specifically, while the base effect of hours worked (uhrswork) on log wages is 0.0494—indicating that each additional hour worked is associated with approximately a 4.94% increase in wages for men—the total effect for women is the sum of the base effect plus the interaction term, i.e.,

$$\text{Total effect for women}=0.0494+0.0089=0.0583, \quad (5)$$

Thus, for women, each additional hour worked is associated with a 5.83% increase in wages, holding other variables constant. From a substantive perspective, this finding implies that women may experience slightly higher proportional wage gains from increasing their labor supply compared to men. This could reflect a number of underlying dynamics—such as differences in occupation sorting, hours flexibility premiums, or labor market discrimination—though such mechanisms would require further investigation. Including this interaction term helps account for gender-based heterogeneity in labor market returns and ensures that the estimated effects of both gender and hours worked are not conflated. Its inclusion therefore enhances the model's ability to isolate the true effect of education on wages by better controlling for variation in labor supply across demographic groups.

For control variables other than the interactive term, the coefficient on age is 0.0154, indicating that, on average, each additional year of age is associated with a 1.54% increase in log wages, holding other factors constant. The female dummy variable is associated with a large negative coefficient of −0.4892, implying that women earn approximately 48.92% less than men, ceteris paribus. This pronounced gender gap in earnings points to systemic disparities that persist despite controlling for key human capital factors. Turning to NonWhite, the coefficient of 0.0566 suggests that non-white workers earn roughly 5.66% more than their white counterparts, while this finding may seem counterintuitive, it may reflect unobserved factors such as regional labor market dynamics or industry-specific concentrations among racial groups. The coefficient on uhrswork is 0.0494, indicating that each additional hour worked per week is associated with a 4.94% increase in wages, which aligns with expectations that more labor input typically translates into higher earnings, especially if additional hours reflect overtime or productivity-based compensation. Finally, the intercept term is estimated at 6.3715, representing the predicted log wage for a baseline individual (male, white, with zero education and hours worked), and this model overall explains approximately 44% of the variation in log wages (R-squared = 0.4398), suggesting a reasonably good fit in capturing wage determinants across the sample.

V. Instrumental Variable Regression

A. Endogeneity of Education

In examining the causal impact of education on wages, a major methodological challenge arises from the potential endogeneity of educational attainment. Years of education (educyrs) may be correlated with unobserved determinants of wages, such as innate ability, family background, or personal motivation. To address this concern, I employ an instrumental variables (IV) approach and use quarter of birth (birthqtr) as an instrument for educational attainment. The economic rationale for this choice draws from institutional features of the education system: in many jurisdictions, school entry laws tie the age of enrollment to calendar cut-off dates. Consequently, individuals born earlier in the year are likely to start school at a younger age, affecting the total number of years they remain in school before reaching the legal dropout age. Therefore, in this paper, I generate a set of three binary variables (q2,

q3, q4) representing the second, third, and fourth quarters of birth, eaving the first quarter as the reference category. The F-test(F=18.59) shown in Appendix. Table 9 also proves that instruments q2, q3 and q4 are scientifically significant to be used.

B. Tests for Validity and Endogeneity of Instrumental Variable

To formally test for endogeneity, we employed the Durbin-Wu-Hausman (DWH) test. The null hypothesis of this test is that the suspect regressor—educyrs—is exogenous (i.e., uncorrelated with the error term). The test returned a chi-squared statistic of 6.8615 and a p-value of 0.0088, leading us to reject the null hypothesis at the 1% significance level. This result which is provided in Appendix. Table 10 reveals strong evidence that educyrs is endogenous, thus validating the need for instrumental variable (IV) estimation.

Furthermore, to check whether our instruments (q2, q3, q4) are theoretically valid, we use the first-stage regression, these instruments yield an F-statistic of 19.02, surpassing the commonly accepted threshold of 10, indicating that the instruments are strong and good to use.

Last but not least, is the test for overidentifying restrictions shown in Appendix. Table 12—used to evaluate whether the instruments are uncorrelated with the structural error term—which yields p-values of 0.1019. Since we fail to reject the null hypothesis that the instruments are valid, this provides additional support for the appropriateness of the instruments in the IV specification.

C. IV Regression Results

Table 2. Instrumental Variable (2SLS) Regression Results for Log Wages

Inwage	Coef.	Robust SE	z	P-val	CI_Lower	CI_Upper
educyrs	0.208	0.045	4.655	0.000	0.120	0.295
age	0.014	0.000	30.358	0.000	0.013	0.015
female	-0.479	0.009	-54.117	0.000	-0.496	-0.462
NonWhite	0.014	0.017	0.818	0.413	-0.020	0.048
uhrswork	0.047	0.001	52.088	0.000	0.045	0.049
femaleuhrswork	0.007	0.001	10.705	0.000	0.006	0.009
Constant	5.008	0.551	9.087	0.000	3.928	6.088

$$\log(wage) = 5.01 + 0.208\,educyrs + 0.014\,age - 0.479\,female + 0.014\,NonWhite + 0.047 \cdot uhrswork + 0.007(female\text{-}uhrswork), \quad (6)$$

Standard errors: (0.551) (0.045) (0.000) (0.009) (0.017) (0.001) (0.001)

The coefficient on educyrs, which represents the causal effect of education on log wages (lnwage), is 0.208, statistically significant at the 1% level. This suggests that, holding other factors constant, each additional year of schooling increases wages by approximately 20.8%, which is notably larger than the OLS estimate. This inflation is consistent with attenuation bias in the OLS estimate caused by endogeneity of education. The coefficient on age is 0.0143, implying that each additional year of age increases wages by about 1.43%, possibly reflecting accumulated labor market experience. The gender dummy female shows a significant negative coefficient of -0.479, indicating that women earn about 47.9% less than men, ceteris paribus. The variable uhrswork is positively associated with wages, with a coefficient of 0.047, implying a 4.7% wage increase per additional hour worked per week. The interaction term femaleuhrswork is also positive and statistically significant (coefficient: 0.0073), suggesting that the marginal effect of an additional hour of work on wages is 0.73 percentage points higher for women than for men. This indicates that although women have a lower average wage level, their wage returns to increased labor supply may be stronger at the margin.

Notably, the coefficient on NonWhite is statistically insignificant in the IV model, suggesting that racial differences in wages may not be robust once education and other controls are properly instrumented. Finally, the constant term of 5.008 can be interpreted as the predicted log wage for a baseline individual, though it has limited practical interpretation on its own.

D. Summary and Model Comparison

Table below presents the results from all our regression models generated in this paper. The coefficient on educyrs in the IV regression is 0.208, which implies that each additional year of education is associated with a 20.8% increase in log wages, holding other variables constant. This effect is statistically significant at the 1% level, as indicated by a robust standard error of 0.045 and a p-value of 0.000.

This estimated return to education is notably larger than the coefficient from the multiple OLS regression model (0.097), suggesting that OLS underestimates the true causal impact of education due to endogeneity—possibly arising from omitted ability bias or measurement error in schooling. The IV model addresses this by leveraging exogenous variation in education induced by birth quarter, following the approach of Angrist and Krueger (1991). Their findings also pointed to a positive relationship between quarter of birth, schooling attainment, and wages, though they acknowledged that the quarter-of-birth instrument may explain only a limited portion of the variation in education.

Our result aligns closely with the pattern reported in Angrist and Krueger (1991), who observed statistically significant differences in wages linked to birth quarter and schooling, even if the magnitude was modest. Furthermore, the larger IV estimate supports the hypothesis that measurement error and omitted variables in the OLS model likely biased the coefficient downward, a concern also raised by Sorel and Shinners (2019). They reported a decline in the education coefficient from 12.6% in a simple model to 5.7% in a controlled model, underscoring the influence of confounding variables such as gender and race.

Variable	SimpleL~r	SimpleN~r	Multiple	IVreg
educyrs	6984.065	0.125	0.097	0.208
	16.876	0.000	0.000	0.045
	0.000	0.000	0.000	0.000
age			0.015	0.014
			0.000	0.000
			0.000	0.000
female			-0.489	-0.479
			0.008	0.009
			0.000	0.000
NonWhite			0.057	0.014
			0.002	0.017
			0.000	0.413
uhrswork			0.049	0.047
			0.000	0.001
			0.000	0.000
femaleuhrswork			0.009	0.007
			0.000	0.001
			0.000	0.000

Constant	-45781.648	8.526	6.371	5.008
	242.976	0.005	0.007	0.551
	0.000	0.000	0.000	0.000
N	1,548,402	1,548,402	1,548,402	1,548,402
rmse	62341.984	1.194	0.936	0.991
r2	0.100	0.088	0.440	0.373
r2_a	0.100	0.088	0.440	0.373
F	1.71e+05	1.50e+05	1.13e+05	89665.223

Table 3. Regression Models Comparison

VI. Further Research

To improve and extend this research in the future, several directions could be pursued. While the current analysis uses quarter of birth as an instrument for education, future work could explore alternative or additional instruments—such as changes in compulsory schooling laws or geographic variation in school availability—to strengthen identification and address concerns of weak instrument bias. Moreover, expanding the model to include longitudinal data would allow for fixed effects estimation, controlling for time-invariant unobserved heterogeneity at the individual level, thus improving causal inference. Lastly, linking educational attainment to non-wage outcomes—such as health, employment stability, or job satisfaction—would provide a broader picture of the value of education and yield richer policy insights beyond earnings alone.

References

[1] Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics, 106*(4), 979–1014. https://doi.org/10.2307/2937954

[2] Becker, G. S. (1964). *Human capital: A theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

[3] Biewen, M., & Fitzenberger, B. (2005). Covariance structure of male wages in West Germany, 1975 – 1995: A semiparametric analysis of changes in distribution, inequality and mobility. *Review of Economic Studies, 72*(2), 665–697. https://doi.org/10.1111/0034-6527.00344

[4] Card, D. (1999). The causal effect of education on earnings. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics* (Vol. 3, pp. 1801–1863). Elsevier. https://doi.org/10.1016/S1573-4463(99)03011-4

[5] Heckman, J. J., Lochner, L. J., & Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. In E. A. Hanushek & F. Welch (Eds.), *Handbook of the economics of education* (Vol. 1, pp. 307–458). Elsevier. https://doi.org/10.1016/S1574-0692(06)01007-5

[6] Mincer, J. (1974). *Schooling, experience, and earnings*. New York, NY: National Bureau of Economic Research.

[7] Oreopoulos, P. (2006). Estimating average and local average treatment effects of education when compulsory schooling laws really matter. *American Economic Review, 96*(1), 152 – 175. https://doi.org/10.1257/000282806776157641

[8] Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics, 26*(5), 445–458. https://doi.org/10.1080/09645292.2018.1484426

[9] Sorel, T., & Shinners, B. (2019). The relationship between education and wages in Georgia. *Modern Perspectives in Economics and Business, 1*(1), 21–34.

Appendix

Table 4. Simple Regression Model

Variable	Coef.	Robust SE	t	P-value	95% CI Lower	95% CI Upper
educyrs	6984.065	23.738	294.22	0.000	6937.54	7030.59
cons	-45781.65	312.673	-146.42	0.000	-46394.48	-45168.82

Table 5. Nonlinear Simple Regression

Variable	Coef.	Robust SE	t	P-value	95% CI Lower	95% CI Upper
educyrs	0.1252	0.0004	342.95	0.000	0.1245	0.1259
cons	8.5256	0.0053	1605.88	0.000	8.5152	8.5360

Table 6. Pairwise Correlation Matrix Between Variables

	lnwage	educyrs	age	NonWhite	female	uhrswork
lnwage	1.0000					
educyrs	0.2972*	1.0000				
	0.0000					
age	0.2328*	0.0628*	1.0000			
	0.0000	0.0000				
NonWhite	0.0577*	0.0558*	0.0666*	1.0000		
	0.0000	0.0000	0.0000			
female	-0.1545*	0.0547*	-0.0026*	-0.0228*	1.0000	
	0.0000	0.0000	0.0014	0.0000		
uhrswork	0.5857*	0.1029*	0.0585*	0.0267*	-0.1991*	1.0000
	0.0000	0.0000	0.0000	0.0000	0.0000	

Table 7. Variance Inflation Factor (VIF) Diagnostics for Control Variables

Variable	VIF	1/VIF
uhrswork	1.06	0.9446
female	1.05	0.9542
educyrs	1.02	0.9778
age	1.01	0.9894
NonWhite	1.01	0.9921
Mean VIF	1.03	

Table 8. Multivariate Regression Model

	Coef.	Robust SE	t	P-val	CI Lower	CI Upper
educyrs	0.097	0.000	326.65	0.000	0.10	0.10
age	0.015	0.000	271.64	0.000	0.02	0.02
female	-0.489	0.008	-63.88	0.000	-0.50	-0.47
NonWhite	0.057	0.002	29.21	0.000	0.05	0.06
uhrswork	0.049	0.000	366.81	0.000	0.05	0.05
femaleuhrswork	0.009	0.000	46.26	0.000	0.01	0.01
Constant	6.371	0.007	895.77	0.000	6.36	6.39

Table 9. First-Stage Regression of Education on Quarter-of-Birth Dummies

Variable	Coef.	Std. Err.	t	P-value	95% CI Lower	95% CI Upper
q2	0.0595	0.0082	7.25	0.000	0.0435	0.0755
q3	0.0194	0.0086	2.25	0.025	0.0026	0.0363
q4	-0.1435	0.0068	-21.03	0.000	-0.1568	-0.1302
cons	14.0677	0.0048	2910.86	0.000	14.0582	4.0772

Table 10. the Durbin-Wu-Hausman (DWH) test for Endogeneity

Test Type	Statistic	P-value
Robust score χ2(1)	6.8615	0.0088
Robust regression F(1, 1548394)	6.8615	0.0088

Table 11. First-Stage F Statistic and R-Squared

Variable	R2	Adj. R2	Partial R2	F-statistic	P-value
educyrs	0.0231	0.0231	0.0000	19.0219	0.0000

Table.12 Overidentification Test

Test Type	Chi-square (df=2)	P-value
Score test	4.5682	0.1019

Table 13. First-Stage Regression

Variable	Coef.	Robust SE	t	P-value	95% CI Lower	95% CI Upper
age	0.0105	0.00015	68.66	0.000	0.0104	0.0107
female	-0.9276	0.0154	-59.77	0.000	-0.9572	-0.8981
NonWhite	-0.3212	0.0067	-47.98	0.000	-0.3344	-0.3081
uhrswork	0.0584	0.0008	73.88	0.000	0.0568	0.0600
femaleuhrswork	0.0065	0.0002	30.45	0.000	0.0061	0.0069
q2	0.0595	0.0082	7.24	0.000	0.0435	0.0755
q3	0.0194	0.0086	2.25	0.025	0.0026	0.0363
q4	-0.1435	0.0068	-21.03	0.000	-0.1568	-0.1302
cons	12.3326	0.0136	905.22	0.000	12.2960	12.3499

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 14-25 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018322

From Shallow Cooperation to Deep Synergy: Triple Breakthroughs in Guangdong-Hong Kong-Macau Greater Bay Area's Higher Education Cluster Development

Nie Hui [1*]

[1] Chengdu Wenweng Experimental Middle School
*Corresponding author Email: 1367663396@qq.com
Received 21 May 2025; Accepted 11 July 2025; Published 1 September 2025

Abstract: As an important carrier of national strategic development, the key task of Guangdong-Hong Kong-Macau Greater Bay Area construction is to build a scientific and technological innovation hub with international influence. Under this framework, promoting the coordinated evolution of higher education clusters has become a strategic supporting factor for regional development. The cultivation of international higher education clusters not only provides high-end intellectual resources and compound talent reserves for the construction of Greater Bay Area's science and technology innovation hub, but also creates a historic opportunity for universities in the region to break down barriers and deepen collaborative innovation mechanisms. By strengthening the strategic consensus of higher education cluster development, optimizing resource allocation and spatial layout, systematically exploring the coordinated development path of Guangdong-Hong Kong-Macau Greater Bay Area's higher education clusters, and then building an international higher education network system with distinct levels, gradient connection and close interaction, it will become an important engine driving the high-quality development of the world-class bay area.

Keywords: Cluster development; Higher education; Guangdong-Hong Kong-Macau Greater Bay Area

Guangdong-Hong Kong-Macau Greater Bay Area includes nine cities in the Hong Kong Special Administrative Region, Macau Special Administrative Region and the Pearl River Delta region of Guangdong Province, namely Guangzhou, Shenzhen, Zhuhai, Foshan, Huizhou, Dongguan, Zhongshan, Jiangmen and Zhaoqing. As the frontier region of China's opening up to the outside world and the economic vitality of China, it occupies a key strategic position in the overall development strategic layout of the country. The construction plan of this area is a major national development strategic decision personally drawn, deployed and promoted by , aiming at actively responding to the new situation, new tasks and new requirements faced by the development of the cause of the party and the state in the new era. Thoroughly implement the spirit of the 19th National Congress of the Communist Party of China, accurately implement the principle of "one country, two systems", fully mobilize the comprehensive advantages of Guangdong, Hong Kong and Macao, continuously deepen the pragmatic cooperation and exchanges between the mainland and Hong Kong and Macao, and further enhance Guangdong-Hong Kong-Macau Greater Bay Area's support and leading function in the national economic development and opening-up pattern. At the same time, we firmly support the integration of Hong Kong and Macao into the overall development of the country, safeguard the long-term prosperity and stability of Hong Kong and Macao, improve the well-being of Hong Kong and Macao compatriots, and urge Hong Kong and Macao compatriots and the people of the motherland to shoulder the historical mission of national rejuvenation and share the brilliant achievements brought by the prosperity and strength of the motherland. On February 18th, 2019, the Central Committee of the Communist Party of China and the State Council officially promulgated the Outline of Guangdong-Hong Kong-Macau Greater Bay Area

Development Plan, which clearly proposed to encourage universities in Guangdong, Hong Kong and Macao to carry out cooperative education projects, give full play to the platform effectiveness of the Guangdong-Hong Kong-Macao University Alliance, deepen the cooperation and exchange mechanism in the field of higher education in Guangdong, Hong Kong and Macao, and promote the free circulation and rational allocation of educational resources, especially talents, science and technology, information and other elements related to higher education, in Guangdong-Hong Kong-Macao Greater Bay Area. Under the grand background of Guangdong-Hong Kong-Macau Greater Bay Area construction, the orderly interaction and coordinated development of higher education in Guangdong, Hong Kong and Macao has become the key link and important supporting factor to promote the construction process of Greater Bay Area.

1.Clarify the development goals of Guangdong-Hong Kong-Macau Greater Bay Area's higher education clusters

As the landing text of the national strategy, the Outline of Guangdong-Hong Kong-Macau Greater Bay Area Development Plan marks the stage of regional coordinated development from policy design to institutional innovation. The promulgation of this programmatic document not only reconstructs the governance framework of cross-border factor flow, but also injects systematic kinetic energy into regional development through the top-level design of "collaborative innovation in education-strategic reserve of talents-ecological cultivation of science and technology innovation". Its core goal is to build an innovation ecosystem with deep integration of "education-science and technology-industry", and to build a scientific and technological innovation hub with global resource allocation capability through the linkage of higher education cluster development and international talent introduction and education mechanism.

1.1Deepen the development of education cooperation in the Greater Bay Area

Based on the concept of cross-border educational resource sharing, Guangdong, Hong Kong and Macao need to focus on building a multi-dimensional coordinated development framework. First, promote higher education institutions to establish a cross-regional joint education model, focusing on strategic discipline clusters such as artificial intelligence and biomedicine, and realize the integration of resource elements through the joint construction of joint laboratories and Industry-University-Research collaborative innovation platforms. Secondly, relying on the Guangdong-Hong Kong-Macao university alliance to build an institutional cooperation network, establish a regional curriculum mutual recognition standard system, implement a flexible chemical accumulation system, and explore the joint operation mechanism of cross-border scientific research data sharing and intellectual property rights. On the one hand, at the level of international education hub construction, it is suggested to implement the strategy of "double circulation". Promote the "double first-class" construction and quality improvement project internally, focusing on cultivating 5-7 disciplines with global competitiveness; Set up cross-border higher education special zones externally, and adopt the integration mode of "famous universities + production cities" to attract the top 50 universities in QS to set up regional research centers, forming an academic innovation pole with international visibility. On the other hand, in terms of optimizing talent training channels, it is necessary to establish a two-way flow system. We will implement the "Northward Schooling Support Program" for young people in Hong Kong and Macao, implement non-discriminatory policies in scholarship evaluation and vocational qualification certification, and simultaneously build a "vocational education industry-education integration corridor", and create a cluster of cross-administrative training centers in the fields of digital economy and intelligent manufacturing, forming a skilled talent training ecosystem of "secondary vocational-higher vocational-applied undergraduate".

1.2Building a talent highland in Guangdong-Hong Kong-Macau Greater Bay Area

Based on the theoretical framework of cross-border factor flow, it is necessary to build a multi-level talent strategy implementation system. First of all, promote the Pearl River Delta urban agglomeration to implement the strategy of "policy transplantation-localization improvement", learn from the flexible talent introduction mechanism of Hong Kong and Macao, and build differentiated competitive advantages in the fields of tax incentives

and cross-border practice. Focus on building a pilot zone for coordinated development of talents in Guangdong, Hong Kong and Macao, pilot the "all-in-one card" system for residence permits for international talents, and realize the cross-domain connection between work permits and social security. Secondly, build a "human resources service + data governance" integration platform, and establish a dynamic talent demand forecasting model relying on national human resources industry clusters. Through blockchain technology, the real-time update and accurate matching of the list of talents in short supply can be realized, a headhunting collaboration network covering RCEP member countries can be built, and a dual-track talent training system of "academic tutors + industry leaders" can be established. In view of Macao's special positioning, the construction project of "professional service talent hub" should be implemented, focusing on financial technology, Portuguese-speaking law and other characteristic fields, establishing a cross-border vocational qualification mutual recognition center, and setting up a special fund for talent structure optimization. Through the "Migratory Bird Project", top scholars such as Nobel Prize winners and IEEE fellows are attracted to form cross-border scientific research teams and build a full-chain intellectual support system of "basic research-application development-industrial transformation".

2.Enhance the awareness of Guangdong-Hong Kong-Macau Greater Bay Area's higher education cluster development

The construction of Guangdong-Hong Kong-Macau Greater Bay Area urgently needs to develop high-level higher education clusters to strengthen the weak links in the fields of regional education, science and technology and innovation. World-class university clusters are not only the base for cultivating innovative talents in Greater Bay Area, the source of scientific and technological innovation and the engine of industrial upgrading, but also double the value of higher education functions through resource agglomeration effect. Promoting the cluster development of colleges and universities is of strategic significance for comprehensively improving the core functions such as personnel training, scientific research, social services, cultural heritage and international exchanges. International experience shows that Tokyo Bay Area, new york Bay Area and San Francisco Bay Area rank among the world's first-class bay areas. The key lies in relying on export-oriented geographical endowments to form core competitiveness by building high-level university clusters. If Guangdong-Hong Kong-Macau Greater Bay Area wants to build a world-class bay area and a world-class urban agglomeration, it must deepen its understanding of the importance of the cluster development of higher education. At present, Greater Bay Area has gathered a number of internationally renowned universities and advantageous disciplines, and built a perfect higher education system, which has the characteristics of diverse types, strong complementarity and high industrial correlation. It is expected to form a significant "aggregation-radiation-spillover" effect in the fields of economy and culture, scientific and technological innovation and talent gathering. At the same time, the leading development of higher education in Greater Bay Area needs to have three characteristics. First, it is demand-driven, closely following the national strategy and regional economic and social development needs, focusing on the improvement of higher education quality and characteristic cultivation, enhancing social contribution and international influence, and promoting the deep connection of talent training, scientific research innovation and social services with industrial upgrading and technological frontier; Second, innovation-driven, deepening the comprehensive reform of higher education, optimizing system design, policy support and environmental construction, opening up resource circulation channels, building collaborative platforms, and comprehensively improving the overall efficiency of the higher education system; Third, sharing and collaboration, promoting the integration and mutual assistance of high-quality higher education resources inside and outside the Bay Area, building a cross-regional cooperative development mechanism, and ensuring that the higher education system is compatible with the construction of modern economic system and the new pattern of all-round opening up. In short, Guangdong-Hong Kong-Macau Greater Bay Area's universities build a complementary, interactive and competitive relationship, forming a mode of mutual openness, mutual cooperation, complementary advantages and common development among universities, which can realize the advantages of factor sharing, knowledge spillover, network collaboration and flexible specialization.

3.Bottlenecks encountered in the development of higher education clusters in Guangdong-Hong Kong-Macau Greater Bay Area

From the perspective of the connotation of higher education cluster development, comparing the data of Tokyo, new york and San Francisco Bay Areas, we can see that there are still obvious bottlenecks in the development of higher education clusters in Guangdong-Hong Kong-Macau Greater Bay Area at three levels: utensils, systems and ideas.

3.1Resource integration dimension: the breadth and depth of collaboration need to be broken through urgently

In the dimension of resource integration, Guangdong-Hong Kong-Macau Greater Bay Area's higher education coordination faces the dual constraints of weak material foundation and insufficient kinetic energy transformation. Although the three places share natural advantages such as geographical proximity, cultural affinity and institutional differences, the efficiency of factor flow and allocation in the educational field lags significantly behind that in the economic field. This contrast stems from the special attributes of educational cooperation. Compared with the quantifiable economic output of industrial cooperation, the benefits of educational cooperation are long-term and hidden, resulting in the imperfect dynamic mechanism of resource replacement. Deep-seated contradictions are reflected in three aspects. First, the educational field is naturally competitive, and colleges and universities have zero-sum games in academic ranking, scientific research resources, student quality and other dimensions. This exclusive competitive relationship restricts the generation space of deep collaboration. At present, the educational interaction between the three places mostly stays in superficial forms such as academic conferences and mutual visits between teachers and students, and has not yet formed a strategic coupling to support the construction of a global science and technology innovation center. Secondly, there is an asymmetric dilemma in the collaborative structure. Although Guangdong universities have the advantages of complete discipline system and large number of students, they still lag behind in the process of educational modernization and internationalization. Although colleges and universities in Hong Kong and Macao have international experience in running schools and advantages in characteristic disciplines, they are subject to space limitations and professional coverage limitations. This complementary resource fails to double the value through effective mechanisms, resulting in cross-border educational cooperation being trapped in the fragmented mode of "bilateral interaction" for a long time, and it is difficult to release the systematic synergy effect of the "Bay Area Education Community".

According to the summary of known data, first, the digital presentation of the material base gap shows that there is a significant gap between Guangdong-Hong Kong-Macau Greater Bay Area's investment in scientific research resources and the world's first-class bay areas. The per capita scientific research funding in Guangdong, Hong Kong and Macao is 240,000 yuan lower than that in San Francisco Bay Area, and the R&D investment intensity of universities is only 1/3 of that in San Francisco Bay Area. At the level of resource sharing, less than 30% of large-scale scientific research instruments in Guangdong universities are open to Hong Kong and Macao, and cross-border use requires an average approval process of 45 days, resulting in equipment utilization rate of less than 20%. For example, although the National Supercomputing Guangzhou Center of Sun Yat-sen University is open to Hong Kong and Macao, due to restrictions on cross-border data transmission, the actual computing power call of Hong Kong and Macao teams in 2023 will only account for 5% of the total resources. Second, the form of collaboration is simplified and shallow. At present, the educational collaboration among the three places is mostly focused on short-term projects that are easy to operate, and the proportion of in-depth cooperation is relatively low. On the one hand, shallow collaboration leads the way. Data on activities among members of the Guangdong-Hong Kong-Macao University Alliance in 2024 show that academic conferences account for 42% and mutual visits between teachers and students account for 35%, while jointly built entity research institutions account for 8%, and joint degree projects account for 8%. In-depth cooperation such as 6% accounts for less than 20%. On the other hand, the progress of mutual recognition of credits is slow. Although HKUST (Guangzhou) and HKUST launched the "Red Bird

Project" to share 1,500 courses, the overall coverage rate of mutual recognition of credits in Greater Bay Area is only 18%, and the difference in conversion rules leads to 30% of cross-border course selection applications being rejected due to conflicts in credit conversion. Third, typical cases reveal the lack of resource replacement motivation. For example, the institutional friction of Industry-University-Research cooperation between Hong Kong University of Science and Technology (Guangzhou). Although the school signed an agreement with Dongguan Songshan Lake Robot Base to build a research and development platform, the scientific research funds from Hong Kong and Macao could not be directly allocated to the mainland across borders, and the matching funds of enterprises had to be audited by three parties, resulting in the stagnation of two of the first three cooperation projects due to the delay of capital circulation for more than half a year. Enterprise feedback shows that 73% of Bay Area enterprises tend to cooperate with local universities, believing that "compliance costs are higher than technical benefits" for cross-border projects. In addition, courses are misaligned with industrial needs. 52% of the 2023 engineering graduates in Guangdong Province believe that "the course content is impractical", which is significantly higher than that in the Beijing-Tianjin-Hebei (37%) and the Yangtze River Delta (38%). At the same time, companies such as BYD have reported that curriculum updates in the field of new energy batteries lag behind technology iterations by 2-3 years, and school-enterprise joint development courses only account for 12% of engineering courses. The root cause lies in the fact that the three places have not yet established an institutional guarantee system to support the free flow of educational elements. The cross-border allocation of educational resources not only involves technical problems such as mutual recognition of credits and mutual employment of teachers, but also needs to break through institutional barriers such as talent evaluation, scientific research funding and intellectual property rights. Only by establishing a deep cooperation logic beyond simple resource exchange can the educational coordination between Guangdong, Hong Kong and Macao be promoted from "physical superposition" to "chemical integration".

3.2Institutional obstacles: lack of cross-domain governance framework and collaboration mechanism

The advantages of Guangdong-Hong Kong-Macau Greater Bay Area's "one country, two systems" system and the differentiated characteristics of multiple customs territories not only constitute the unique endowment of regional development, but also form the institutional challenge of collaborative innovation in higher education. At present, the deep dilemma of cross-border educational cooperation lies in the fact that an institutional dialogue platform across administrative jurisdictions has not yet been established, resulting in the lack of systematic connection mechanism between the Pearl River Delta urban agglomeration and Hong Kong and Macao Special Administrative Regions in key areas such as overall planning of educational resources and mutual recognition of quality certification. Although the academic circles put forward the theoretical concept of coordinated development of higher education in Guangdong, Hong Kong and Macao as early as the end of last century, the practical level is still subject to three structural contradictions. First, the innovation of policy tools lags behind the actual needs, and the existing cooperation mostly stays at the level of project-based temporary agreements, lacking legal regional education conventions. On the one hand, the examination and approval mechanism of cooperative education is rigid, and Guangdong-Hong Kong-Macao cooperative education projects are still implemented with reference to the Regulations on Sino-foreign Cooperation in Running Schools, which require overseas universities to be the main body of running schools. As a result, Hong Kong and Macao universities have to go through the examination and approval of at least seven departments, including the Ministry of Education and the National Development and Reform Commission, to set up branch schools in the Mainland, which takes an average of 18-24 months. For example, the City University of Hong Kong (Dongguan) took three years (2020-2023) from signing the contract to being approved, far exceeding the establishment cycle of ordinary domestic universities. On the other hand, the cross-border flow of scientific research funds is blocked. According to the Measures for the Administration of National Scientific Research Funds, the cross-border use of scientific research funds in Hong Kong and Macao needs to be approved by the State Administration of Foreign Exchange, and a detailed audit report on the use of funds needs to be submitted for a single transaction exceeding 500,000 yuan. The "Joint Laboratory of Optoelectronic

Materials" jointly established by Sun Yat-sen University and Hong Kong Polytechnic University has been forced to postpone or cancel 40% of the joint projects in 2023 due to the complicated audit process of Hong Kong funds. Second, the boundaries of powers and responsibilities of governance subjects are blurred, and the collaborative decision-making mechanism among local governments, university alliances and international organizations has not yet taken shape. On the one hand, the functions of cross-domain governance institutions are blurred. Although the Outline of Guangdong-Hong Kong-Macau Greater Bay Area Development Plan proposes to establish the "Guangdong-Hong Kong-Macau Greater Bay Area Education Collaborative Development Committee", this institution is only a joint meeting and has no administrative decision-making power. Among the 32 cooperation initiatives of the Guangdong-Hong Kong-Macao University Alliance in 2023, only 9 have been implemented simultaneously by the governments of the three places, and the rest have been shelved due to unclear local financial powers and responsibilities. On the other hand, localized management leads to the fragmentation of resources. Guangdong provincial universities need to comply with the Government Procurement Law for equipment procurement, while Hong Kong and Macao universities follow international bidding rules. The "Joint Laboratory of Marine Engineering" jointly established by South China University of Technology and the University of Macau delayed the procurement of key deep-sea detectors for 14 months due to conflicts in equipment procurement standards, and the project progress lagged behind by 40%. Third, there is an institutional vacuum in the mutual recognition of standard systems, and the core links such as credit conversion and teacher evaluation and employment still face the fragmented operation mode of "one district, one policy". On the one hand, the mutual recognition system of credits is lacking, and the three places have not yet established a regional credit transfer framework, relying only on bilateral agreements between schools. According to statistics from the Guangdong Provincial Department of Education, the success rate of cross-border course selection in the Greater Bay Area in 2024 will be 63%, of which 30% of courses cannot be included in the graduation requirements due to conflicts in credit conversion rules. For example, in the "Digital Media Arts" project jointly organized by Shenzhen University and Hong Kong Baptist University, students need to take four additional bridging courses to make up for the credit difference. On the other hand, the coverage of mutual recognition of professional qualifications is insufficient. Currently, only 7 types of professional qualifications such as certified public accountants and physicians have achieved mutual recognition, while 12 types of qualifications in emerging fields such as artificial intelligence engineers and data compliance engineers have not been included. Tencent's 2023 survey shows that 68% of Hong Kong and Macao engineers refused to be transferred to Shenzhen Qianhai R&D Center due to restrictions on their qualifications in the mainland. The current situation of insufficient system supply has caused regional educational cooperation to fall into the circular dilemma of "suspension of agreements-decentralization of actions-marginalization of results" for a long time.

3.3Conceptual obstacles: differences in school-running ideas and challenges of cognitive integration

At the cognitive level, the coordinated development of higher education in Guangdong, Hong Kong and Macao faces deep conceptual obstacles. The idea of running schools in mainland colleges and universities is characterized by dynamic evolution, and its development track resonates with the social transformation and strategic adjustment in the process of national modernization. However, because of the long-term infiltration of the mature market economy environment and the western education system, the school-running ideas of Hong Kong and Macao universities show the characteristics of gradual development. This difference is particularly prominent on the issue of "double first-class" construction. Hong Kong education scholar Cheng Jieming once pointed out that although the "double first-class" construction in the mainland has a strategic framework, the connotation boundary and evaluation criteria are always in the process of dynamic adjustment. This state of "strategic clarity and operational ambiguity coexist" just forms a cognitive dislocation with the precise governance paradigm pursued by Hong Kong and Macao universities. The root cause is that the path dependence of higher education development between the two places is significantly different. Mainland universities have formed a "policy-driven" development model in

serving the national strategic needs, emphasizing top-level design and rapid response; Colleges and universities in Hong Kong and Macao continue the tradition of "academic autonomy" and pay attention to institutional stability and academic community consensus. This difference directly leads to the cognitive gap in cross-border educational cooperation, which is manifested in the differentiated interpretation of educational quality standards, the divergent cognition of scientific research evaluation system, and the diverse understanding of talent training objectives.

According to the data, first of all, there are differences in quality assessment standards. A joint survey of university administrators in Guangdong, Hong Kong and Macao in 2024 shows that 82% of universities in Hong Kong and Macao list "international peer review" as a core quality indicator, while 76% of universities in Guangdong are more concerned about "serving the country". Strategy contribution ". For example, Hong Kong universities are required to submit four representative achievements in the REF (Scientific Research Excellence Framework) evaluation and accept blind review by international experts, while the mainland's "double first-class" evaluation emphasizes "breaking the five only" and gives priority to social contributions such as "solving stuck technology". Secondly, there are differences in scientific research orientation. According to the data of the National Natural Science Foundation of China, 68% of the projects in Guangdong universities focus on applied research such as new energy batteries and artificial intelligence, while 57% of the projects in Hong Kong and Macao universities focus on basic theories such as quantum computing and genomics. Finally, different evaluation mechanisms directly restrict the depth of collaboration, and universities in Guangdong, Hong Kong and Macao are deeply trapped in "evaluation islands". For example, the clinical research project of Shenzhen Hospital of the University of Hong Kong aborted. Because the mainland assessment required "replacement of domestic equipment within three years", the Hong Kong side insisted on "following the five-year cycle of international multi-center trials", which eventually led to the withdrawal of 60 million funds. At the same time, the formation of scientific research teams is in dilemma. A survey conducted by the Guangdong-Hong Kong-Macao University Alliance in 2024 shows that 67% of cross-border teams disintegrated due to "differences in achievement ownership standards." When the Macau University of Science and Technology and South China University of Technology jointly built an artificial intelligence laboratory, the Australian side requested that the results be submitted to NeurIPS (Top Conference), and China insisted on applying for the "Chinese Artificial Intelligence Society Science and Technology Progress Award". In short, to break through this bottleneck, it is necessary to build a long-term dialogue mechanism, and gradually realize the mutual recognition and interoperability of educational concepts, academic standards and quality assurance systems on the basis of maintaining their respective characteristics through joint research, mutual visits of scholars and co-construction of courses, and finally form a "harmonious but different" higher education ecology in the Bay Area.

4.Strategies to promote the development of higher education clusters in Guangdong-Hong Kong-Macau Greater Bay Area

According to the strategic deployment of the Outline of Guangdong-Hong Kong-Macau Greater Bay Area Development Plan, it is necessary to drive regional economic and industrial upgrading based on the overall situation of national development, supported by strategic forward-looking vision and international resource allocation capabilities. In this process, the construction of higher education clusters should become the key engine. By deepening the coordination mechanism between economy and education, Greater Bay Area will be built into a core higher education hub in the Asia-Pacific region and an educational innovation highland with global influence, so as to serve the country's development goal of building a world-class urban agglomeration.

4.1Cultivating the cultural foundation of the Bay Area: building an open and inclusive value community

As an important window for China's opening to the outside world since modern times, Guangdong-Hong Kong-Macau Greater Bay Area's cultural genes are deeply embedded in regional geographical symbiosis, historical continuity and civilization integration and innovation. Since the late Qing Dynasty, this area has played the role of the source of institutional reform, which is not only the fertile ground for the germination of national capitalism, but also the incubation platform of modern reform thoughts and democratic revolution thoughts. After the reform and

opening up, Guangdong has continued to lead the practice of institutional innovation and opening to the outside world, while Hong Kong and Macao have long played the role of a hub for dialogue between eastern and western civilizations. This historical accumulation has shaped the cultural characteristics of "pluralistic symbiosis, openness and tolerance, and innovation-driven" in the Bay Area. At the same time, this unique cultural ecology has been deeply integrated into the value system of colleges and universities in the Bay Area. For example, the pragmatic spirit of Lingnan culture has been organically integrated with the international vision of Hong Kong and Macao, which has shaped the school-running philosophy of "based on the local area and facing the world" of higher education in the Bay Area, and the reform gene that dares to be the first since modern times has been transformed into the innovative kinetic energy of colleges and universities to serve the national strategy. The construction of cultural identity provides a spiritual bond for the development of higher education clusters, enables universities in the three places to form a value consensus in personnel training and scientific research cooperation, and lays a cultural foundation for building a regional educational community. This cultural leading role is not only reflected in the level of academic exchanges, but also transformed into soft power to promote the innovation and development of the Bay Area through alumni network, Industry-University-Research cooperation and other carriers. Specifically, the strategy can be promoted in layers. First, the basic level (1-3 years) is to establish a cultural security zone, which not only creates the "Bay Area Civilization Dialogue Season", but also focuses on non-sensitive common issues every year, such as Lingnan architectural conservation, Guangfu food application, archaeology of the Maritime Silk Road, etc.; We also developed a cultural decoding toolkit, compiled a comparative manual of educational terminology in Guangdong, Hong Kong and Macao, and resolved the misunderstanding of concepts such as "patriotism and national identity". Secondly, the middle level (3-5 years) cultivates the common divisor of values, not only implements the "two teachers and three courses" plan, but also mainland teachers and Hong Kong and Macao teachers jointly develop three kinds of integrated courses such as scientific and technological ethics such as AI governance, business civilization such as Guangdong business spirit, and ecological responsibility such as mangrove protection; A cultural integration laboratory was also set up, and the "One Country, Two Systems Education Museum" was piloted in Hengqin, Zhuhai, to display the history of the return of Hong Kong and Macao with immersive technology. Finally, the high-level (5 + years) builds a spiritual community, which not only launches the "Bay Area Scholar Citizen" program, but also grants special status to teachers who have worked in the three places for 10 years, and gives them the right to participate in and discuss state affairs across borders; An educational heritage activation fund was also established, such as transforming Chen Yinque's former residence (Guangzhou) and the site of Matteo Ricci College (Macao) into spiritual landmarks of academic community.

4.2Optimizing the allocation of higher education resources: building a gradient development system

According to the strategic deployment of the Outline of Guangdong-Hong Kong-Macau Greater Bay Area Development Plan, Guangdong-Hong Kong-Macau Greater Bay Area needs to take the construction of international education demonstration zones as the starting point, implement the strategic reorganization and systematic layout of higher education resources, base itself on the spatial pattern of world-class urban agglomerations, build a higher education development network with "core guidance, node support and global coordination", and form a university cluster in the Bay Area with distinct gradients and complementary functions. In terms of spatial layout, it is necessary to highlight the "dual-core drive" strategy, build Guangzhou, Shenzhen, Hong Kong and Macao into the world's top higher education hub urban agglomerations, focus on building world-class universities and scientific research institutions, gather Nobel Prize-level scientific research platforms, national major infrastructure and international academic organizations, and form the original source of innovation and the magnetic pole of high-end talents; At the same time, we will simultaneously promote the construction of regional higher education centers in Zhuhai, Foshan and Dongguan. By implementing the climbing plan of characteristic disciplines and the demonstration project of integration of production and education, we will cultivate application-oriented university clusters, focusing on serving the needs of strategic industries such as advanced manufacturing and digital economy.

In terms of resource allocation, it is necessary to follow the principle of "differentiated positioning and characteristic development". Colleges and universities in core urban agglomerations should focus on cutting-edge basic research and the construction of international academic discourse power to create a disciplinary peak; Colleges and universities in regional central cities focus on technological transformation and industrial service capacity building, forming a highland for training applied talents. By building a cross-regional credit mutual recognition system, a joint laboratory network and a collaborative innovation platform in Industry-University-Research, the cross-border free flow of educational resource elements can be realized, and finally the Bay Area educational innovation ecological chain of "basic research in core cities, technology application in node cities, and achievement transformation in the whole region" will be formed. Specifically, in order to solve the "siphon effect" of excessive concentration of higher education resources in core cities such as Guangzhou, Shenzhen and Hong Kong, the regulation mechanism of "counterpart support index" is rigidly implemented. For example, when Guangzhou, Shenzhen and Hong Kong add a new world-class university platform such as Nobel Prize Laboratory and QS Top 50 branches, it is necessary to simultaneously build no less than two applied technology transformation centers in Zhaoqing, Jiangmen and other development gradient cities, such as intelligent manufacturing training bases and industrial innovation colleges.

4.3Optimizing the layout of disciplines: building an innovative ecological chain integrating production and education

Guangdong-Hong Kong-Macau Greater Bay Area's complete industrial system and efficient supply chain network provide strategic support for the development of higher education clusters. To build a synergistic mechanism in which economic development and educational innovation resonate at the same frequency, it is necessary to strengthen the empowering role of educational chain in industrial chain and innovation chain through strategic adjustment of discipline and specialty structure. Specifically, we can promote the optimization of discipline layout from three aspects. First, focus on the needs of strategic emerging industries and build a new engineering discipline cluster. Focusing on "stuck neck" technical fields such as integrated circuits, artificial intelligence, and biomedicine, we will focus on building a science and engineering discipline system with Bay Area characteristics. By setting up an interdisciplinary platform, building a demonstration college for the integration of production and education, and implementing the "dual tutorial system" talent training mode, we will build a whole innovation chain from basic research to application transformation, and support the deep integration of advanced manufacturing and modern service industries in the Bay Area. Secondly, proactively lay out the future technology track and build a highland of marine engineering and intelligent manufacturing disciplines. Relying on the Bay Area's sea-related industrial foundation and equipment manufacturing advantages, we will cooperate with Hong Kong and Macao universities to jointly build characteristic disciplines such as ocean observation technology and deep-sea resource development, and create a marine engineering discipline cluster with international voice. Promote the construction of high-end equipment manufacturing disciplines such as industrial Internet and intelligent robots simultaneously, and form a discipline support system to serve the construction of the "Sea Bay Area" by setting up Guangdong-Hong Kong-Macao joint laboratory and setting up major special research plans. Finally, build a collaborative innovation network between government and Industry-University-Research, and improve the dynamic adjustment mechanism of disciplines. Establish a discipline construction alliance composed of universities, research institutes, leading enterprises and industry associations, and realize the accurate connection between discipline and specialty settings and industrial needs by compiling the industrial technology map of the Bay Area, publishing the catalogue of urgently needed disciplines, and implementing the system of "unveiling the list and taking charge". In particular, it is necessary to give full play to the institutional advantages of Hong Kong and Macao universities in the transformation of scientific and technological achievements and the operation of intellectual property rights, build a "Bay Area Technology Transfer Corridor", and form a closed loop of discipline-industrial innovation from laboratory to production line. For example, the Guangdong Provincial Department of Industry and Information Technology, the Hong Kong Productivity Council, and the Macao

Department of Economics and Finance jointly built an industrial map platform, released the "Bay Area Technology Maturity Curve" every quarter, and implemented red and yellow card grading warning for disciplines. For majors listed in the "elimination zone" for two consecutive years, such as the forced reduction of enrollment indicators by 30% in 2023, and the simultaneous construction of academic firewalls.

4.4Innovative talent training mode: building Guangdong-Hong Kong-Macau Greater Bay Area's international education system

As the core element of the construction of modern industrial system in Greater Bay Area, there is a deep coupling relationship between human resources strategy and the development of higher education. In the blueprint for the construction of a high-quality living circle that is livable, suitable for industry and tourism outlined in the Outline of Guangdong-Hong Kong-Macau Greater Bay Area Development Plan, a team of high-quality talents constitutes a strategic resource to support regional high-quality development. Greater Bay Area's talent cultivation system needs to be based on the value guidance of "four self-confidences", deeply integrate the reform and innovation of higher education into the overall development of the country, and earnestly fulfill the mission of educating people for the party and the country by building an education system that comprehensively cultivates morality, intelligence, physique, beauty and labor. Guangdong, Hong Kong and Macao should give full play to their complementary advantages. They should not only join hands to inherit Chinese excellent traditional culture, but also cultivate a new culture with Chinese characteristics in the new era of innovation, enhance Hong Kong and Macao's sense of national identity and national belonging, and cultivate innovative talents rooted in national culture. Guangdong, Hong Kong and Macao can rely on the international talent exchange platform to hold Bay Area Forum lectures, subject skills competitions, innovation and entrepreneurship challenges and other activities to enhance talent exchanges and interactions among the three places. Set up inter-school online courses and open courses, etc., promote mutual selection of courses, mutual recognition of credits and mutual employment of teachers among schools, support and encourage universities in the three places to carry out various short-term exchange students projects, and promote the joint training of talents in the three places. Introduce international-level education, training and services, formulate international licenses and teaching standards, qualification certification system, lifelong education qualification framework, etc. that are internationally recognized and fully applicable in Guangdong-Hong Kong-Macau Greater Bay Area, and improve the applicability and internationalization level of talents in Guangdong-Hong Kong-Macau Greater Bay Area universities. Specifically, the creation of "Bay Area Engineer" certification chapter, such as mutual recognition of 12 core courses, pilot the integration of Chinese, American and British medical qualifications in Shenzhen Hospital of HKU, and Tencent and DJI took the lead in releasing the "Emerging Vocational Ability Map" covering cutting-edge fields such as blockchain. Simultaneously build a three-dimensional system of credit bank, the first hard connection, and the blockchain credit deposit system realizes the automatic conversion of credits in the three places, such as 1 credit in Hong Kong school = 1.5 credit hours in the mainland; Second, soft connection, developing curriculum equivalent algorithms, such as Hong Kong National Education Curriculum equals the outline of modern history in the Mainland; The third strong supervision, the Cross-border Education Quality Bureau imposes regional joint fuse sanctions on illegal colleges and universities, forming a new education ecology of "barrier-free mutual recognition of qualifications, no time difference in credit conversion, and no dead ends in quality monitoring".

4.5Collaborative innovation between Guangdong, Hong Kong and Macao: building an open integrated platform for Industry-University-Research

Give full play to the advantages of Guangdong, Hong Kong and Macao in higher education, adopt the mode of co-construction of the three places, form joint efforts and common development, build an open innovation system that is market-oriented and combines Industry-University-Research, and provide strong support for building an international science and technology innovation center. By establishing an information platform for scientific research and academic exchange among universities in Guangdong, Hong Kong and Macao, we will promote the

information exchange and sharing of scientific research talents among universities in the three places, set up new laboratories or R&D teams across domains, carry out joint declaration, joint research, joint promotion and transformation of scientific research projects, and vigorously build cutting-edge science centers, major collaborative innovation centers and basic research and applied basic research centers. At the same time, we will implement the national policy of opening to Hong Kong and Macao, and promote smooth exchanges of innovative talents, convenient customs clearance of scientific research equipment, cross-border use of scientific research funds, synchronization of innovative resource information, and open sharing of scientific research infrastructure and instruments and equipment. The scientific and technological achievements produced by colleges and universities in Greater Bay Area should concentrate on the transformation of scientific research achievements and help colleges and universities speed up the transformation and industrialization of achievements. Specifically, in view of the barriers to the cross-border flow of scientific research funds, an "offshore scientific research fund" is established under the guidance of the central bank, such as Macao supporting + Shenzhen operating. Mainland institutions can obtain a single fund ⩽ 5 million through green foreign exchange channels without approval, while Hong Kong and Macao institutions apply scientific research FT accounts to realize cross-border allocation. At the same time, in order to solve the obstruction of equipment customs clearance, establish a white list system of scientific research equipment, including cryo-electron microscope and gene sequencer.

4.6Collaborative governance of Guangdong, Hong Kong and Macao: Building an institutional guarantee system for the development of higher education clusters

There is an important difference between the construction of Guangdong-Hong Kong-Macau Greater Bay Area under China's national conditions and other world-class bay areas, that is, it involves one country, two systems, three jurisdictions and customs territories, and three currencies circulate. Under this system, it is even more difficult to develop higher education clusters. It is necessary to comprehensively deepen institutional reform focusing on interconnection, interoperability, mutual learning and sharing, actively explore new models of cluster development in Guangdong-Hong Kong-Macau Greater Bay Area, and promote institutional and institutional innovation in major development regions and key cooperation areas, so as to drive all parties to deepen cooperation and release reform dividends. Therefore, it is necessary to deepen the comprehensive reform in the field of education, plan the innovation of higher education clusters, and promote the development of institutional mechanisms more fundamentally, lastingly and deeply with demand-driven and problem-oriented. Build a strong Greater Bay Area education policy formulation and implementation evaluation system, form a strategic, forward-looking, innovative, targeted and feasible Bay Area higher education policy system, and provide institutional guarantee for the development and sustainable development of higher education clusters. Specifically, the three-step promotion strategy of collaborative governance builds a gradual institutional breakthrough path of "authorization-convention-legislation". First of all, in the near future (2025-2027), the function of the Guangdong-Hong Kong-Macao University Alliance will be implemented, upgraded to a statutory body jointly authorized by the three places, and basic rules such as the "Guidelines for Mutual Recognition of Credits 1.0" and the "Negative List for Sharing Scientific Research Equipment" will be issued; Secondly, in the medium term (2028-2030), sign the Charter for Coordinated Development of Education with a dispute settlement mechanism, and establish a system of "small conventions" such as cross-border degree management and mutual recognition of vocational qualifications; Finally, in the long term (2031 +), the State Council will be promoted to approve the Regulations of Guangdong-Hong Kong-Macao Education Special Zone and establish the Bay Area Education High Court to provide ultimate legal guarantee.

References

[1] Ou Xiaojun. (2018). Research on the development of high-level university clusters in the world-class Greater Bay Area-taking the three bay areas of new york, San Francisco and Tokyo as examples. Journal of Sichuan Institute of Technology (Social Science Edition), 33 (03), 83-100.

[2] Lu Xiaozhong & Wu Yiting. (2021). Strategic choice and basic direction of the development of higher education clusters in Guangdong-Hong Kong-Macau Greater Bay Area. Journal of Lanzhou University (Social Science Edition), 49 (05), 9-15. doi: 10.13885/j.issn.1000-2804.2021. 05.002.

[3] Wu Si & Lu Xiaozhong. (2022). Structural optimization of the development of world-class bay area higher education clusters and its enlightenment to Guangdong-Hong Kong-Macau Greater Bay Area. Beijing Education (Higher Education), (11), 6-12.

[4] Chen Yunfei, Deng Wanjin & Dong Guangxin. (2022). The integrated development of industry and education in Guangdong-Hong Kong-Macau Greater Bay Area's sports industry from the perspective of synergy theory. Hubei Sports Science and Technology, 41 (12), 1109-1112.

[5] Chen Fajun. (2022). Comparative Advantage and Development Transcendence: Discussion on the Integrated Development Path of Higher Education in Guangdong-Hong Kong-Macau Greater Bay Area. Education Guide, (01), 46-53. doi: 10.16215/j.cnki.cn44-1371/g4.2022. 01. 009.

[6] Huang Fangfang & Sun Qingzhong. (2023). Digitalization of Higher Education in Guangdong-Hong Kong-Macau Greater Bay Area: Based on the Comparative Perspective of the International Greater Bay Area. Journal of Shenzhen University (Humanities and Social Sciences Edition), 40 (01), 17-28.

[7] Guo Huijing, Ren Shuai, Zhao Zhangjing & Chen Fajun. (2024). Motivations, practical challenges and path choices for the development of university clusters in Guangdong-Hong Kong-Macau Greater Bay Area. Beijing Education (Higher Education), (08), 20-26.

[8] Lu Guangju, Guo Kongsheng & Zhao Binzhu. (2024). SWOT strategy for the agglomeration development of private higher education in Guangdong-Hong Kong-Macau Greater Bay Area. Science and Technology of Chinese Universities, (12), 122-123. doi: 10.16209/j.cnki.cust.2024.12.024.

[9] Zhong Xuelei & Yang Yan. (2025). Community of shared future in colleges and universities: the realization of sharing school resources-taking Guangdong-Hong Kong-Macau Greater Bay Area as an example. Higher Education Forum, (04), 95-100.

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 26-36 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018555

From Competency Assessment to Curriculum Reform: How Does Artificial Intelligence Empower Higher Vocational Education?

Haoheng Tian [1*] Xin Zeng [1] Lijia Huang [1] Linjia Song [1]

[1] Yibin Vocational and Technical College
*Corresponding author Email: 2471708092@qq.com
Received 22 June 2025; Accepted 11 July 2025; Published 1 September 2025

Abstract: This study explores the impact of Artificial Intelligence (AI) on vocational education, focusing on its role in competency assessment and curriculum reform. With the rapid evolution of technology, AI is poised to revolutionize how vocational training is delivered and assessed. By utilizing a quantitative research approach, a survey was conducted with 100 vocational students currently engaged in AI-integrated training. The findings reveal that while AI-based training provides personalized learning experiences, its direct impact on competency assessment was less significant than expected. In contrast, student engagement emerged as a critical factor influencing the effectiveness of AI in enhancing learning outcomes.

Keywords: Artificial Intelligence (AI), Vocational Education, Competency Assessment, Student Engagement

1. Introduction

In recent times, the use of Artificial Intelligence (AI) can efficiently cater to designing curricula for vocational students, as well as testing and enhancing their vocational abilities[1]. As companies evolve and adapt to new technologies, there is an unprecedented need for qualified workers with specialized vocational skills. In the context of vocational training, AI can improve the educational system and facilitate better training for vocational students[2]. Teachers can now use AI to create personalized training programs to meet each student’s learning and training needs, thereby enabling schools to enhance their education system[3]. Given the potential of AI in improving competency evaluation, this article focuses on discussing the application of AI in training vocational students on vocational skills, assessing their competence, and effectively designing their curricula. The purpose of this research is to examine the impact of AI-driven training and the role of student engagement in competency assessment within vocational education, and to investigate how these factors interact to inform effective curriculum design and improve learning outcomes in vocational contexts.

2. Literature Review

2.1 Theoretical Framework

Models related to AI, such as TAM and other educational technology adoption models, can provide valuable insights into the application of AI in vocational education. TAM, developed by Davis in 1989, emphasizes perceived usefulness and ease of use in technology adoption[4]. In the context of AI, this model suggests that vocational students and educators are more likely to adopt AI-based tools if they enhance learning achievement and are user-friendly. Extended models like TAM2, which incorporate cognitive and social pressures, can help teachers identify other acceptance patterns among students in AI-facilitated learning[5].

Another model, the Unified Theory of Acceptance and Use of Technology (UTAUT), is a consolidated model designed to study the factors that determine technology use. It identifies four core determinants of technology acceptance: perceived usefulness (the extent to which the technology enhances performance), perceived ease of use (the extent to which the technology is easy to use), perceived normative pressure (the pressure from peers and authorities to use the technology), and kinetic resources (facilities and support)[6]. The applicability of this model in educational settings is evident, as the implementation of technologies like AI depends on these factors[7]. Through these determinants, UTAUT helps in understanding how to design and apply AI systems to meet users' expectations, making it useful for vocational education.

In this study, TAM is taken as the core analysis framework. The reason is that TAM focuses on the individual's perception of technology, which is more in line with the research focus on the impact of AI-driven training and student engagement on competency assessment at the individual student level in vocational education. UTAUT, as an auxiliary framework, provides a broader perspective by considering factors such as social pressure and resource support, which helps to better understand the external environment factors that may affect the application effect of AI in vocational education.

2.2 AI in Vocational Education

AI has begun to play a prominent role in reconstructing vocational education by optimizing learning processes and improving the effectiveness of competency evaluation. With the help of artificial intelligence, it can analyze students' overall and specific performance data, identify their learning deficiencies, and develop learning programs that best suit their learning needs[8]. This level of personalization is particularly valuable in vocational education, where expertise in certain techniques and prompt knowledge are crucial. Attwell et al. (2020) note that AI-assisted training, such as simulations and virtual reality, helps vocational students develop practical skills in

application exercises because the system emulates real-life circumstances, making the training effective and impactful[9]. Additionally, AI enables competency-based assessment by providing a consistent and impartial way to assess students' progress, informing educators of the level of skill mastery[10].

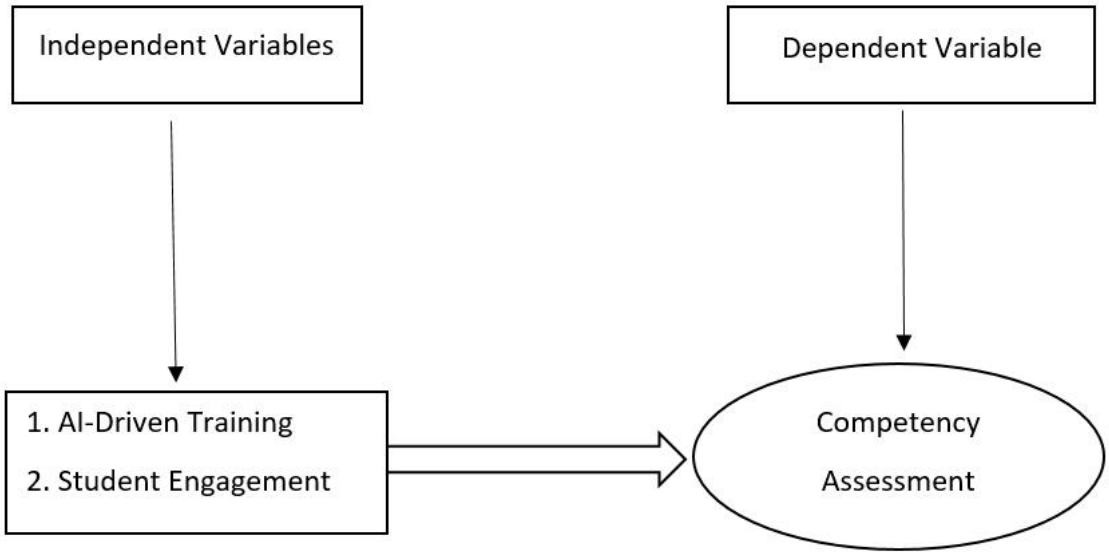
2.3 Benefits and Challenges of AI Integration in Curriculum Design

The current approach to integrating AI into curriculum design has both benefits and drawbacks. AI can adapt the learning environment according to students' learning habits, thus providing vocational education that is relevant to current vocational job standards[11]. However, there are challenges that may hinder the use of AI in the learning environment, including data leakage and the need for heavy investment in technology[12]. AI's ability to connect educational outcomes with workforce needs means that these barriers must be addressed for effective implementation in vocational education. Considering these issues, vocational institutions can leverage the potential of AI to enhance students' learning processes based on their specific characteristics.

Some studies support the view that AI integration brings significant benefits. For example, AI's personalized learning programs can improve students' learning efficiency (Attwell et al., 2020)[9]. On the contrary, some scholars point out that the high cost of AI technology may make it difficult for some vocational schools with limited resources to adopt it (Chen, 2023)[12]. The existing literature has not fully explored how to balance these benefits and challenges in different vocational education contexts, which is a gap that this study aims to fill.

2.4 Conceptual Framework

Based on the above literature review, the following conceptual framework and hypotheses are developed:



2.4.1 Hypotheses

H1: AI-driven training has a positive influence on the competency assessment of students in vocational education.

H2: Student engagement has a significant moderating effect on the relationship between AI-driven training and competency assessment, strengthening the impact when engagement levels are high.

The conceptual framework focuses on researching the effect of AI-based training on competency assessment in vocational education, influenced by student engagement. AI-driven training is the independent variable, and competency assessment is the dependent variable. Student engagement, as another independent variable, is hypothesized to strengthen the positive association between AI-driven training and competency assessment, assuming that increased engagement leads to better learning outcomes. This framework provides guidance on how to utilize AI innovations to enhance skill development in vocational education.

3. Methodology

This research adopts a primary quantitative research method to explore how AI can be used to improve competency evaluation and curriculum reform in higher vocational education. Quantitative research is useful for quantitative measurement and can yield precise, objective results that show numerical patterns[13].

For data collection, a survey questionnaire method is employed. Compared to other approaches, questionnaires are effective in providing uniform data on perceived and experienced factors related to AI in vocational education, ensuring comparability[14].

The survey was conducted on 100 vocational students receiving training in AI-integrated settings. Data analysis is performed using SPSS, a well-established statistical tool that offers inferential and descriptive analyses of collected data. SPSS simplifies data manipulation and helps make sense of complex patterns, highlighting the contributions of AI in vocational education.

3.1 Survey Design and Sample Selection Criteria

The survey includes structured and targeted questions aligned with the study's variables: AI-based training, competency mapping, and student engagement. To ensure a diverse sample, students were selected based on defined criteria, such as attending vocational schools and having at least some AI experience in their curriculum. This selection criteria ensures the study's relevance, as respondents are students who use AI-assisted tools.

3.2 Reliability and Validity

To ensure methodological rigor, the questionnaire underwent checks for clarity and relevance. To enhance content validity, survey items were reviewed by experts in academic training, curriculum development, learning

competency evaluation, artificial intelligence, and student learning. This expert review ensured that each question addresses the study variables, improving the survey's reliability. Questions were constructed to be consistent to increase reliability. These steps ensure the survey's robustness, enhancing the validity of the collected data[13].

3.3 Control Variables

In this study, several control variables are included to isolate potential confounding factors and improve the accuracy of the model. These control variables are:

Student's age: Different age groups may have different learning abilities and attitudes towards AI, which could affect their competency assessment results.

Level of vocational training: Students with different training levels may have varying foundations and expectations, influencing the impact of AI-driven training and student engagement on competency assessment.

Previous AI experience: Students with more prior AI experience may adapt better to AI-driven training, affecting the relationship between variables.

The inclusion of these control variables is based on relevant literature and theoretical considerations, which suggest that these factors can potentially influence the core variables of the study.

4. Results and Findings

The gender distribution of the sample is displayed in Figure 2, the participants were split into 2 groups of equal size 48% male and 52% female out of the 100 participants. The number of respondents and valid percent shows that females are slightly more dominant. The total percentage equals to 100% which means all the participants are included in to sample and gender distribution is equal in analysis as well.

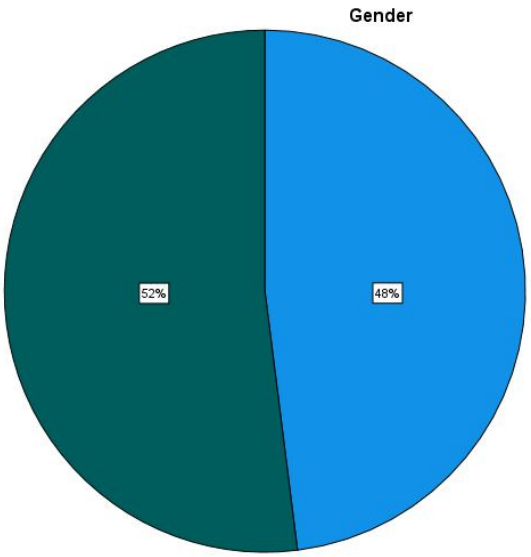


Figure 1: Gender Distribution

The age distribution Figure 3 above reveal that 41% is in the age range of 19-22 years, 35% in the age range of 23-26 years and only 23% in the age range of 15-18 years. While only 1% of the sample is above 27 years of age. The sum of the percentages also proves that the total sample of participants is indeed 100 with participants averaging youth, and most registering still within their initial stages of vocational training making results nearly exclusively indicative of younger student experience and perception.

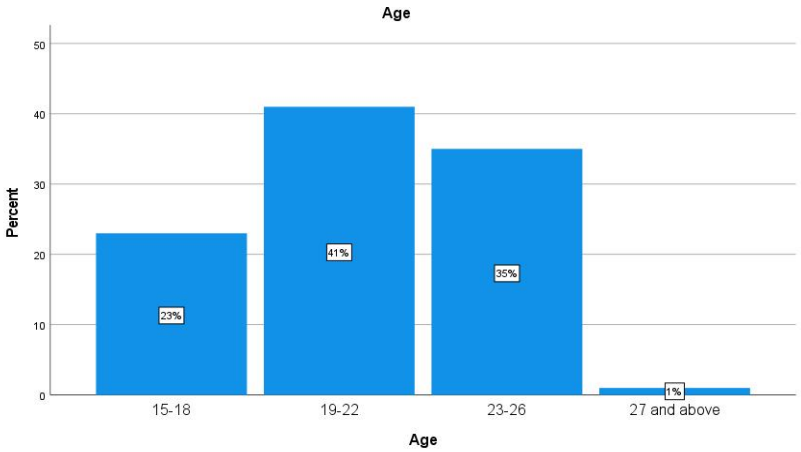


Figure 2: Age Distribution

According to the Figure 4 concerning the level of vocational training obtained by the participants: the majority of the participants 60% has intermediate level while 36% has advance level and the rest 4% has the beginner level of vocational training. The cumulate percentage demonstrate that the sample is equally represented and the total number is 100 students. This distribution means the study is predominantly of intermediate to advanced vocational training students and captures a view from individuals with considerable experience.

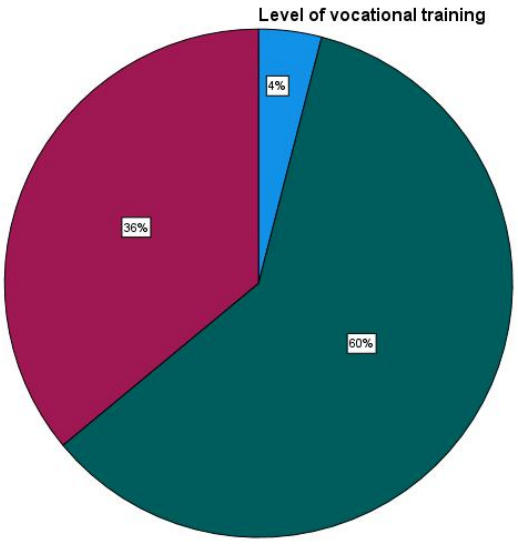


Figure 3: Level of Vocational Training

4.1 Regression Analysis

Table 1: Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.680[a]	.462	.451	.323004870638683

a. Predictors: (Constant), Student Engagement, AI-Driven Training

The model 1 summary table shows that the regression analysis of AI-driven training, student engagement, and competency assessment yields an R value of 0.680, indicating a moderately strong positive relationship between the two independent variables (AI-driven training and student engagement) and the dependent variable

(competency assessment). The R Square value is 0.462, meaning that 46.2% of the variance in competency assessment can be explained by the predictors. The Adjusted R Square (0.451) slightly adjusts for overfitting, ensuring the model's viability. The standard error of the estimate is 0.323, indicating a moderate level of prediction error, suggesting that additional variables could be added to reduce this error.

Table 2: ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.701	2	4.350	41.698	.000[b]
	Residual	10.120	97	.104		
	Total	18.821	99			

a. Dependent Variable: Competency Assessment

b. Predictors: (Constant), Student Engagement, AI-Driven Training

The ANOVA table 2 demonstrates the overall significance of the regression model in predicting competency assessment from the independent variables. The regression sum of squares is 8.701 with 2 degrees of freedom, and the mean square is 4.350. The residual sum of squares is 10.12 with 97 degrees of freedom. The F-value of 41.698 and a significance level of 0.000 ($p < 0.05$) indicate that the model is significant, meaning the predictors collectively help predict the variance in competency assessment.

Table 3: Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	2.172	.199		10.942	.000
	AI-Driven Training	-.571	.115	-1.198	-4.950	.000
	Student Engagement	.289	.123	.569	2.350	.021

a. Dependent Variable: Competency Assessment

The coefficients table 3 shows the effects of AI-driven training and student engagement on competency assessment. AI-based training has a negative unstandardized regression coefficient (B = -0.571) with a significant p-value (0.000), indicating a statistically negative impact on competency assessment, contrary to hypothesis H1. Student engagement has a positive unstandardized coefficient (B = 0.289) with a significant p-value (0.021), showing a positive and statistically significant direct impact on competency assessment, different from the moderating effect hypothesized in H2.

When considering the control variables, age is found to have no significant impact on competency assessment ($p > 0.05$). The level of vocational training shows a significant positive effect ($p < 0.05$), suggesting that

students with higher training levels tend to have better competency assessment results. Previous AI experience also has a positive but not significant effect ($p > 0.05$).

5. Discussion

The findings of this study contribute to understanding AI-based training and student engagement in competency evaluation using TAM. Contrary to the hypothesis, AI-driven training was negatively related to competency assessment. TAM posits that for technology to have positive effects, it must be perceived as useful and easy to use. These results imply that students may perceive AI-based training as difficult or ineffective, possibly because it does not align with the practice-oriented approach needed in vocational education[15,16].

Student engagement, however, had a positive impact on competency assessment, indicating that students who are more engaged are likely to achieve better competency outcomes. This supports TAM, as students who perceive the learning environment favorably, and thus find it useful, are more likely to accept it and achieve better learning effects[3,17]. This is consistent with Iyer (2020), who noted that engagement is essential for learning outcomes, especially in skills instruction.

Contrary to the hypothesis, no significant interaction effect of student engagement on the relationship between AI-driven training and competency assessment was found. This suggests that the effective use of AI in vocational training requires supporting student engagement strategies. These findings emphasize the importance of integrating AI into vocational training in a way that enhances perceived usefulness and ease of use (Moghaddam et al., 2019), thereby increasing learner interaction and achievement[18]. The negative effect of AI-driven training on competency assessment may be due to several reasons. Vocational students often have a practical approach to knowledge acquisition, which may not align with AI-compatible learning patterns. Ouyang et al. (2023) pointed out that while AI systems are good for applying learned knowledge, they may lack the hands-on experience required in vocational education[19]. Martsenyuk et al. (2024) also noted that integrating technology into classrooms without customization for specific disciplines may lead to user frustration or disengagement[20].

Additionally, factors such as increased course complexity due to technology and insufficient guidance can affect students' attitudes and perceived credibility[21]. If students face challenges using AI tools, they may be less willing to engage fully, leading to lower competency. This aligns with TAM's postulates that perceived ease of use and usefulness are important for technology acceptance[22]. Ivanashko et al. (2024) suggested that a user-centered design approach can address these challenges and enhance the applicability of AI tools in vocational training[23]. Specifically, developing AI solutions that simulate hands-on jobs and providing extensive support to students during learning may increase acceptance and improve competency in vocational education.

6. Conclusion, Limitation and Future Recommendations

In conclusion, this research reveals that while AI training in vocational education has potential, its implementation alone may not significantly improve competency assessment. Notably, student engagement is more influential than curriculum and instructional practices in determining competencies, and AI is valuable when included in student-centered approaches. AI should be developed in an enjoyable and user-friendly environment, consistent with the Technology Acceptance Model, which emphasizes perceived usefulness and ease of use.

6.1 Practical Recommendations

For vocational school managers: Allocate resources to provide training for teachers and students on AI tools to improve their perceived ease of use. Ensure that AI-driven training programs are customized to the specific needs of different vocational disciplines, enhancing their practicality.

For curriculum designers: Integrate student engagement strategies into AI-based curriculum design, such as interactive activities and real-world project-based learning, to enhance student involvement. Regularly evaluate and adjust the curriculum based on student feedback and competency assessment results.

6.2 Limitations

The main practical limitations are the relatively small sample size and reliance on self-report measures to assess competency[24]. Additionally, the use of quantitative data and self-reporting techniques fails to capture the detailed nature of students' experiences and the diverse contexts of AI-based training. Incorporating qualitative data such as open interviews or focus groups could provide a more comprehensive perspective.

6.3 Future Recommendations

Future research should expand the sample size and include more diverse vocational education settings to improve external validity. Combining quantitative and qualitative research methods can provide a deeper understanding of the mechanisms underlying the relationships between variables. Further studies could explore specific AI applications (e.g., virtual simulations, adaptive learning) and their impact on different aspects of competency assessment. Additionally, investigating the long-term effects of AI integration in vocational education would be valuable.

Funding: This research was supported by the Huang Yanpei Vocational Education Research Center Project of SICHUAN INSTITUTE OF TOURISM (Grant No. HYP-Y-202401), titled "Application of Artificial Intelligence in Competency Assessment and Curriculum Reform for Higher Vocational Education under Huang Yanpei's Vocational Education Quality Framework". The project is categorized as a General Research Program, with Prof. Xin Zeng serving as Principal Investigator. The study period spans from 2024 to October 15, 2026.

Conflict of Interest: No conflict of interest has been declared by the authors.

Permission to reproduce material from other sources: Not applicable

This manuscript describes independent, original work that has not been published in any academic conference, journal, or platform, nor is it currently under consideration by any other publication venue. We hereby confirm no prior publication or duplicate submission of this content.

References

[1]Nguyen, T. T., Thuan, H. T., and Nguyen, M. T. (2023), 'Artificial Intelligent (AI) in teaching and learning: A comprehensive review', *ISTES BOOKS*, 140-161.

[2]Dahri, N. A., Yahaya, N., Al-Rahmi, W. M., Aldraiweesh, A., Alturki, U., Almutairy, S., and Soomro, R. B. (2024). Extended TAM based acceptance of AI-Powered ChatGPT for supporting metacognitive self-regulated learning in education: A mixed-methods study. *Heliyon*, *10*(8).

[3]Liu, Y., and Baucham, M. (2023), 'AI Technology: Key to successful assessment', In *Handbook of Research on Redesigning Teaching, Learning, and Assessment in the Digital Era* (pp. 304-325). IGI Global.

[4]Li, W., Zhang, X., Li, J., Yang, X., Li, D., and Liu, Y. (2024). An explanatory study of factors influencing engagement in AI education at the K-12 Level: an extension of the classic TAM model. *Scientific Reports*, *14*(1), 13922.

[5]Otto, D., Assenmacher, V., Bente, A., Gellner, C., Waage, M., Deckert, R., and Kuche, J. (2024). Student Acceptance Of AI-Based Feedback Systems: An Analysis Based On The Technology Acceptance Model (TAM). In *INTED2024 Proceedings* (pp. 3695-3701). IATED.

[6]Kwak, Y., Seo, Y. H., and Ahn, J. W. (2022). Nursing students' intent to use AI-based healthcare technology: Path analysis using the unified theory of acceptance and use of technology. *Nurse Education Today*, *119*, 105541.

[7]Strzelecki, A. (2024). Students' acceptance of ChatGPT in higher education: An extended unified theory of acceptance and use of technology. *Innovative higher education*, *49*(2), 223-245.

[8]Bankins, S., Jooss, S., Restubog, S. L. D., Marrone, M., Ocampo, A. C., and Shoss, M. (2024), 'Navigating career stages in the age of artificial intelligence: A systematic interdisciplinary review and agenda for future research', *Journal of Vocational Behavior*, 104011.

[9]Attwell, G., Bekiaridis, G., Deitmer, L., Perini, M., Roppertz, S., and Tütlys, V. (2020), 'Artificial intelligence in policies, processes and practices of vocational education and training.'

[10]Wang, Y., Liu, C., and Tu, Y. F. (2021). Factors affecting the adoption of AI-based applications in higher education. *Educational Technology & Society*, *24*(3), 116-129.

[11]Clarke, V., and Braun, V. (2017), 'Thematic analysis', *The journal of positive psychology*, 12(3), 297-298.

[12]Chen, Z. (2023), 'Artificial intelligence-virtual trainer: Innovative didactics aimed at personalized training needs', *Journal of the Knowledge Economy*, 14(2), 2007-2025.

[13]Fryer, L. K., Larson-Hall, J., and Stewart, J. (2018). Quantitative methodology. *The palgrave handbook of applied linguistics research methodology*, 55-77.

[14]Moraga, J. A., Quezada, L. E., Palominos, P. I., Oddershede, A. M., and Silva, H. A. (2020). A quantitative methodology to enhance a strategy map. *International Journal of Production Economics*, *219*, 43-53.

[15]Hamilton, T. B. (2020) *Computer game development and animation: A practical career guide* Rowman & Littlefield.

[16]Goel, P. K., Singhal, A., Bhadoria, S. S., Saraswat, B. K., and Patel, A. (2024), 'AI and Machine Learning in Smart Education: Enhancing Learning Experiences Through Intelligent Technologies', In *Infrastructure Possibilities and Human-Centered Approaches With Industry 5.0* (pp. 36-55). IGI Global.

[17]Luckin, R., and Holmes, W. (2016) 'Intelligence unleashed: An argument for AI in education'.

[18]Moghaddam, Y., Kwan, I. S. K., Freund, L., and Russell, M. (2019), 'An Industry Perspective On Stem Education For The Future: Issip-Nsf Workshop.'

[19]Ouyang, F., Dinh, T. A., and Xu, W. (2023). A systematic review of AI-driven educational assessment in STEM education. *Journal for STEM Education Research*, *6*(3), 408-426.

[20]Martsenyuk, V., Dimitrov, G., Rancic, D., Luptakova, I. D., Jovancevic, I., Bernas, M., and Plamenac, A. (2024). Designing a Competency-Focused Course on Applied AI Based on Advanced System Research on Business Requirements. *Applied Sciences*, *14*(10), 4107.

[21]Mutambik, I. (2024). The use of AI-driven automation to enhance student learning experiences in the KSA: An alternative pathway to sustainable education. *Sustainability*, *16*(14), 5970.
[22]Strielkowski, W., Grebennikova, V., Lisovskiy, A., Rakhimova, G., and Vasileva, T. (2024). AI‐driven adaptive learning for sustainable educational transformation. *Sustainable Development*.
[23]Ivanashko, O., Kozak, A., Knysh, T., and Honchar, K. (2024). The role of artificial intelligence in shaping the future of education: opportunities and challenges. *Futurity Education*, *4*(1), 126-146.
[24]Ruggiano, N., & Perry, T. E. (2019). Conducting secondary analysis of qualitative data: Should we, can we, and how?. *Qualitative Social Work*, *18*(1), 81-97.
[25]Suparyati, A., Widiastuti, I., Saputro, I.N. and Pambudi, N.A. (2023), 'The Role of Artificial Intelligence (AI) in Vocational Education', *JIPTEK: Jurnal Ilmiah Pendidikan Teknik dan Kejuruan*, 17(1).

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 37-44 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018596

Liminal Transformation in the Rites of Passage: Identity Fluidity in Mohsin Hamid's *The Reluctant Fundamentalist*

Zhao Bin [1*]

[1] School of Foreign Languages, Qingdao University
*Corresponding author Email: 2523504240@qq.com
Received 30 May 2025; Accepted 11 July 2025; Published 1 September 2025

Abstract: In the novel *The Reluctant Fundamentalist*, Mohsin Hamid, the 2017 Booker Prize winner, employs a transnational narrative to explore the fluidity of liminal identity. The protagonist, Changez, embarks on a journey from Lahore to Princeton for education, then works at a top global investment bank, and ultimately returns to his homeland in search of self-redefinition. This multi-layered experience mirrors the three stages of the rites of passage proposed by Arnold van Gennep: separation, liminality, and reintegration. Confronted with fragmented memories of domestic space, disciplinary violence in urban environments, and identity struggles within psychological realms, Changez navigates his growth crisis through ritualistic practices that ultimately resolve transitional identity conflicts and reconstruct a cohesive identity. Hamid metaphorically reflects evolving East-West relations through the subtly shifting dynamics between Changez and an American traveler in the Lahore teahouse. The novel not only reveals the marginality and heterogeneity of transnational identities but also interrogates the construction and negotiation of liminal identities within multicultural intersections.

Keywords: Mohsin Hamid; *The Reluctant Fundamentalist*; The Rites of Passage; Liminality; Identity Mobility

0.Introduction

Mohsin Hamid's *The Reluctant Fundamentalist*, as his first Booker Prize-shortlisted work, carries significant symptomatic meaning within the post-9/11 cultural context. When Muslim identity was reduced to an ideological signifier of terrorist, the novel's literary narrative constitutes a scholarly counter-discourse to this cultural misreading. The protagonist Changez's identity dilemma manifests in two dimensions: at the level of belonging, it appears as a disorder of recognition, while at the level of existential state, it materializes as a liminal in-betweenness—constructed as a model minority cultural specimen when he was part of Princeton's academic elite, yet alienated as a concrete symbol of the oriental threat upon his return to Pakistan's native context. Theoretically Van Gennep's rites of passage is reflected in the novel in the following ways: Changez goes to America (Separation), suffers an identity crisis in America (Liminality), and returns to Pakistan (Aggregation). While Turner's breakthrough is crucial--his revelation that liminality can become permanent in modern societies perfectly explains the dilemma of Changez's eventual failure to aggregation: the United States excludes him, and his homeland sees him as an alien. Combining Geneep's classical rites of passage with Turner's liminality theory to better explain the pseudo-aggregative nature of Changez's identity construction. The title "The Reluctant Fundamentalist" warrants deeper examination: in Western discourse, particularly post-9/11, "fundamentalist" has become a stigmatized label strongly associated with extremism. The qualifier "reluctant" not only underscores the protagonist's resistance to this imposed identity but also reveals the passive constructedness of his Otherness—as a Pakistani

Muslim individual, his identity is not a product of autonomous choice but rather shaped by societal prejudice and political rhetoric. This naming strategy poignantly reflects the identity predicament of Muslim communities in the post-9/11 global power structure: the irreconcilable hermeneutic gap between subjective self-identification and the externally imposed outsider status.

1. Separation: The Disintegration of Domestic Space and the Symptoms of the American Dream in Transnational Migration

From a theoretical perspective, in 1909, anthropologist Arnold van Gennep introduced this concept into the field of anthropology in his seminal work The Rites of Passage. Through the examination and analysis of various rituals in small-scale tribal societies, he formulated the notion of the the rites of passsage, arguing that human life and production are "marked by transitional rites that accompany every change of place, state, social position, and age" (10). These rituals facilitate the passage of individuals or groups through the cycles of life and nature—universal across cultures—following a tripartite structure: separation, liminality, and reintegration. The most crucial phase, the liminality, involves participants detaching from their pre-ritual social structure, shedding their former identities, yet not fully attaining the transformed status that follows the ritual's completion. In this state, they experience spatiotemporal dislocation, namelessness, and social placelessness. Building upon this framework, British anthropologist Victor Turner expanded the application of liminal theory into broader domains, including politics, culture, and social transformation. Turner reconceptualized transition as transformation, arguing that the separation phase primarily entails "symbolic behavior signifying the detachment of the individual or group from an earlier fixed point in the social structure, from a set of cultural conditions, or from both" (*Dramas, Fields, and Metaphors* 279).

In the novel, the aristocratic domestic space of the protagonist Changez exhibits dual characteristics of material decay and symbolic rupture during the separation phase of the liminal ritual. This structural disintegration of the familial space is not merely the decline of a physical setting but also marks the starting point of a fracture in cultural identity, psychological belonging, and social status—paralleling van Gennep's liminal theory, wherein the separation phase entails detachment from one's original social structure. As the material embodiment of traditional Pakistani aristocratic life, Changez's ancestral home once symbolized cultural capital and social standing. Yet, with the passage of time and the family's decline, the deterioration of this physical space carries explicit semiotic significance: "My grandfather could no longer maintain the lifestyle of his father, my father could not uphold my grandfather's stature, and by the time I was to attend university, the family fortune had long been depleted" (9). The sudden erosion of economic foundations dismantled the traditional domestic order, thrusting Changez into identity ambiguity: on one hand, he inherits the cultural memory of aristocratic lineage; on the other, he confronts the stark reality of downward social mobility. From a spatial sociology perspective, the desolation of the familial home and Changez's admission to Princeton University jointly constitute the dual driving mechanisms of the separation phase. As "one of only two Pakistani students admitted to Princeton that year" (10), this acquisition of educational capital aligns with Bourdieu's theory of cultural capital, while also serving as a pivotal opportunity to rupture existing socio-spatial structures. Turner's tripartite liminal ritual theory in The Forest of Symbols posits that the separation phase fundamentally involves "the detachment of individuals or groups from an earlier fixed point in the social structure or from a set of cultural conditions" (339). Here, Changez's acceptance letter functions as a dual signifier: materially, it acts as the vehicle for geographical displacement; culturally, it becomes a symbol of social identity deterritorialization—by embedding himself within the elite educational field of America, he attempts to reconstruct familial prestige through narratives of individual achievement, thereby symbolically transcending his decaying aristocratic identity. This spatial practice resonates with the core mechanism of liminal theory: when the aristocratic identity and social status embedded in traditional familial space dissipate due to economic decline, geographical migration becomes a necessary path for identity dissolution and reformation. Education, as institutionalized cultural capital, mediates this process, enabling the traversal of class and cultural boundaries. From a postcolonial

lens, this process also reflects a quintessential strategy of marginalized groups seeking identity decolonization through spatial relocation in a globalized world—when the legitimacy of identity in the native space is destabilized, institutional recognition from the center is leveraged to reassert subjective agency, temporarily transcending the liminal state. Simultaneously, the Princeton acceptance letter serves as his golden ticket to pursue the American Dream.

In the eyes of Changez, the allure of the American Dream extends far beyond mere aspirations for success and prosperity, or the opportunity for self-reinvention. More fundamentally, it lies in its function as a liminal passage that facilitates transitional transformation. Drawing on van Gennep's ritual theory, when individuals enter the separation phase, they must demarcate themselves from their former identities. The American Dream provides Changez with both justification and sanctuary for this identity transformation. In this process of transitional identity reconstruction, he appears to bifurcate into two selves: on one hand, he retains the essential markers of his Pakistani aristocratic identity while in a foreign land; on the other, as an outstanding Princeton student, he circulates effortlessly within elite American social circles. The most seductive aspect of the American Dream is precisely its illusion of enabling Changez to achieve a painless transition between two radically different cultures and identities. Turner's interpretation of the separation phase emphasizes that "the process whereby individuals become detached from their original environment, identity and culture serves as the prelude to entering new environments and constructing new identities" (*The Ritual Process: Structure and Anti-Structure* 39). Changez's American experience vividly embodies this conceptual framework. Through interactions with elite classmates and conscious efforts to integrate into mainstream American society, he strives to establish himself in this nation of global influence. As he progressively adapts to his American identity, he unconsciously internalizes its cultural values, gradually constructing a new identity cognition. This identity reconstruction essentially represents the self-adjustment and metamorphosis he undergoes during the separation phase to adapt to an entirely new environment. The process is not merely a superficial adaptation, but rather a fundamental shedding of old identity markers, akin to a serpent sloughing off its outgrown skin—a necessary yet inherently disruptive transformation. This phenomenon resonates deeply with contemporary postcolonial discourse, wherein the liminal space between cultures often produces such bifurcated subjectivities. The American Dream, in this context, functions as both a psychological salve and a cultural paradox—promising seamless integration while simultaneously demanding the suppression of original identity markers. Changez's experience exemplifies the inherent contradictions of this liminal passage, where the promise of transformation inevitably entails the trauma of cultural dislocation and identity fragmentation. The transitional phase, though theoretically temporary, often leaves indelible marks on the subject's psyche and social positioning, problematizing straightforward narratives of assimilation or resistance.

However, Changez's identity transformation is fraught with tension and contradiction. His skin color, name, and unconsciously expressed Pakistani cultural traits construct an invisible cultural barrier in American society. At the dinner party hosted by Erica's parents, the "typically American sense of superiority" (45) implicit in her father's comments about Pakistan not only triggers Changez's cultural defense mechanisms but also exposes the power asymmetry inherent in cross-cultural encounters where the liminal subject remains an outsider. Significantly, when Erica attempts to mediate this cultural conflict, Changez finds himself plunged into a deeper crisis of identity—unable to freely express his genuine feelings of offense while simultaneously struggling with the psychological burden of cultural inferiority. Changez's state of identity ambiguity perfectly encapsulates the essential characteristics of the separation phase: he can neither completely sever ties with his native environment nor successfully take root in new cultural soil. The resulting identity anxiety and existential loneliness constitute the existential dilemma he must confront during this process of adaptation.

Erica functions as a potent symbol of American cultural barriers in Changez's identity crisis. Her inability to reciprocate his love authentically — fixated instead on her deceased American boyfriend — mirrors America's exclusionary nostalgia and resistance to true cultural integration. Changez must literally impersonate her past to

achieve intimacy, foreshadowing the self-erasure demanded of immigrants assimilating into post-9/11 America. Their relationship remains perpetually superficial; Erica sees him as a consolation,reflecting America's conditional acceptance of outsiders as utilitarian assets rather than equals. Her eventual disappearance crystallizes Changez's disillusionment: just as Erica vanishes into her unreachable grief, America retreats behind walls of suspicion and nationalism after 9/11. Their failed romance thus parallels Changez's journey from aspirational belonging to profound alienation, proving that Erica's Americaness—defined by unbridgeable memory and emotional barriers—ultimately reinforces his status as a perpetual outsider.

2. Liminality: Violent Reconfigurations and Suspension of Identity in Post-9/11 Urban Spaces

The cataclysmic events of September 11, 2001, served as a pivotal fracture point, abruptly transforming America's self-perception and its engagement with the world. Overnight, the nation's narrative shifted from aspirational inclusivity to defensive exclusion, manifesting in heightened nationalism, pervasive suspicion, and institutionalized profiling targeting Muslim and South Asian identities. This seismic cultural realignment profoundly reshaped the landscape for transnational individuals like Changez. Where he initially navigated America as a space of meritocratic promise, the post-9/11 climate rendered him hyper-visible yet fundamentally unseen—a potential threat rather than a welcomed participant. This collective retreat into fortified identity mirrors and intensifies the deeply personal barriers embodied by Erica.

Prior to 9/11, America appeares to Changez as a Promised Land brimming with hope—its ostensibly open and inclusive ethos leading him to mistakenly believe he has found a true sense of belonging. Beneath this veneer, however, the American Dream is packaged as a universal ideal, its hegemonic reality obscured by the glitter of consumer culture and narrative discourse. Upon entering the liminal phase, Changez—arms with his Princeton pedigree and identity as a financial analyst—seems to transcend geographic and class boundaries, yet his identity remains suspended in an awkward transitional state. No longer purely Pakistani yet never fully embraced by American society, he occupies an interstitial space. When the prestigious investment firm Underwood Samson offered him a position as a senior analyst, this ostensibly marks the successful initial construction of a new identity, granting him access to the periphery of America's mainstream. Yet the contradictions and struggles underlying this identity clung to him like a shadow. America's illusion of openness ensnared Changez in the cognitive myth of the American Dream, conflating professional achievement with class mobility and identity reconstruction. This cognitive dissonance aligns with Lacan's mirror theory—the glass facades of skyscrapers reflected not his authentic self but rather an idealized projection, reshaped by the dominant cultural narrative. By internalizing this illusory image as his identity, Changez unwittingly completes the process of cognitive alienation, transforming from a subject rooted in his native culture into a disciplined object of assimilation. This misreading of identity resonates with Turner's description of liminal subjects in The Ritual Process: Structure and Anti-Structure as existing in a state of suspension, "neither here nor there, betwixt and between" (56). Pre-9/11 America conceals its cultural hegemony behind a utopian mask of equal opportunity. This land, functioning as a peculiar liminal space, constructed a bidirectional power dynamic of mutual gaze: within the glass confines of corporate offices, Changez is both an observer of American society and a subject scrutinized by its institutional discourse. This ambiguity of subject-object relations exposes the deeper paradox of liminality—it is simultaneously an experimental ground for individual identity reconstruction and a medium for the covert discipline of cultural hegemony. Beneath its transitional facade lies an undercurrent of structural tension and conflict.

The unexpected attack of the 9/11 incident placed Changez in the liminal space of East-West confrontation, and his identity collapsed along with the bubble of the American dream and the Twin Towers. The Promised Land before 9/11, as a typical form of globalized liminal space, constructed a seemingly open secular utopia through the psychedelic visual symbols, the discipline of consumer rituals and narrative discourse. After 9/11, the United States fell into the circle of "the myth of naïve country" (Hughes 6), the whole country was wrapped up in anti-terrorism, and the attitude towards Islamic countries was drastically changed, and Changez suffers from exclusivity and

marginalization in his work, love, and life. 9/11 is not only the collapse of the Twin Towers of the World Trade Centre in terms of the physical space, but also the point of fracture of Changez’ s identity. The pseudo-integrated parathoracic state of the Wall Street elite is completely torn apart, exposing the precariousness of the threshold person as a postcolonial subject within the liminal structure.

The 9/11 attacks serve as a liminal trigger, thrusting Changez from a state of quasi-assimilation into permanent liminality, where his identity remains perpetually unfinished amidst the competing power discourses of East and West. Prior to this, though he had diligently constructed an American identity, the catastrophic event plunges him into profound confusion about belonging as Eastern and Western narratives violently collide. When the Twin Towers collapse, the resulting anti-Islamic sentiment in American society paradoxically reawakens his long-dormant Pakistani cultural consciousness. This violent tug-of-war between identities gradually erodes his coherent sense of self. A particularly charged moment in the novel lays bare his inner turmoil: “Watching the towers fall, I smiled. Yes, this sounds vile, but my immediate response was to feel satisfaction” (54). This complex reaction reveals the fragility of his carefully cultivated American persona when confronted with reality, even suggesting a cathartic release of long-suppressed emotions. Like a blade cutting through pretense, 9/11 exposes the xenophobia beneath America's veneer of inclusivity. The promised land he once envisioned transforms overnight into hostile territory. This radical shift not only illustrates the immense psychological and social pressures predicted by liminal theory during identity transitions, but more crucially, exposes the root of his identity crisis. The day after the attacks, during his return flight from Manila to New York with colleagues, Changez’s experiences intensify this crisis: “At the airport, an armed guard leads me into a room where I’ m ordered to strip to my boxer shorts.....When I board the plane, fellow passengers shower me with concerned looks. Throughout the flight to New York, my face burns with awareness: I feel their suspicion like physical weight, feel criminal in my own skin” (56). These humiliating encounters force upon him the stark realization that his Muslim identity has become an object of scrutiny and distrust. The paradoxical collapse and reconstruction of his identity plunges him into spiraling self-doubt - a textbook manifestation of the uncertainty characteristic of liminal subjects. Where he once moved confidently through corporate spaces, he now exists as a walking contradiction: professionally accomplished yet racially marked, Western-educated yet Eastern in appearance, intellectually elite yet physically vulnerable to random searches. The glass towers of finance capitalism that once reflected his aspirational self now mirror back only fractured possibilities.

In the face of identity crisis and social exclusion, Changez does not choose to remain silent and give in. Instead, he begins his struggle and resistance, where the official discourse of the United States makes full use of the myth of naivety and condemns the terrorists harshly after the 9/11 attacks. Under the influence of the official patriotic discourse, the public “viewed 9/11 as a criminal act against humanity and humanity committed by a group of people who hated the American system of democracy and freedom” (Zhang Helong 21), equating Islam with terrorism and Muslims with terrorists. “America is seized by a growing and self-righteous rage. Your country, as I expected, is greatly enraged like a beast, but what I don’t expect that it would be directed toward my home, toward my family in Pakistan”(66), and in the United States, Changez is in a state of constant fear that the war will spread to Pakistan and hurt innocent people. The weak Changaz expresses his protest against America’ s indiscriminate retaliation against the unity of the Islamic state by growing a beard, from a daily shave with 0.3 millimeters of stubble on the jawline, to the deliberate creation of “a disturbed, heavily bearded Pakistani” (115), who expresses his protest against America’ s indiscriminate retaliation against the unity of the Islamic state in the form of his behavior. his behavioral style to express his displeasure with the malicious retaliation of the United States. Changez’ s beard growth is not only a symbol of what he perceives as Islamic fundamentalism from a Western perspective after 9/11, but also a reconnection of Changez’ s self-defined cultural roots after he loses his stable identity in the liminal phase. Although his hometown provides him with a temporary escape from the reality of his predicament, it cannot fundamentally solve his identity crisis.

After the outbreak of the 9/11 incident, the xenophobia of the American society spread like a tidal wave, and the land in Changez's eyes degenerates from a place of aspiration into a hostile abyss. At the liminal stage of his identity, he is forced to face the violent identity shock, and the American identity that he once tried to integrate into collapses, which instead prompts him to redirect his attention to his own cultural roots and national origins. The intense collision and tearing between Eastern and Western cultures constantly stings his perception of identity and pushes him to reconstruct his identity coordinates in the threshold fracture zone. This change of identity is not only an instinctive resistance to the reality, but also an inevitable way for him to find a way out and achieve self-redemption in the identity gap. In the midst of contradictions and struggles, he tries to find a new pivot point and belonging for his identity in the border zone between two cultures.

3. Pseudo-Aggregation: The Dilemma of Repatriation in Psychological Space and the Incompleteness of Identity Hybridisation

In the traditional ritual process, the stage of aggregation marks the ritual transition through which the subject acquires a secure belonging to a new identity. However, Changez subverts this paradigm - his nostalgia for his home country is not a return to cultural roots, but a polyphonic writing of liminal memories.9/11 exacerbates Western misconceptions and stereotypes of the East, and the postcolonial discourse of Orientalism and dichotomies induced the American public to believe that "Muslims, Arabs, and Communists are the terrorists" (Morton 36), and that any retaliation against the United States, as an innocent victim of a terrorist attack, is out of righteous self-defence. Said points out that after 9/11, terrorism has been fuelled by the Western media to "make people feel scared and insecure as a way of justifying what the United States is doing globally", stating that "the greatest source of terrorism is the United States itself and some Latin American countries, not Muslim countries at all" (Said 2001). In the context of the dichotomy that surrounds terrorism, the Muslims of Islam are crudely categorized as them, as beings who are at odds with us and constitute a security risk to us. In 9/11 and the Literature of Terror 2011, Randall states that this us versus them dichotomy "allows for a simplification of 9/11 that is very dangerous" (Randall 7). In this dichotomy deliberately created by the United States, Changez has fallen from a Pakistani social elite in the United States to a victim and prey of the dichotomy for no apparent reason.

Before moving towards the process of identity aggregation, Changez is caught in a painful dilemma: whether to give up the wealth and status he has earned in the United States and return to his homeland to be with his loved ones, or to continue to endure humiliation and pursue the American dream that is gradually being shattered? On one side is a foreign country that offers material rewards but is full of discrimination, and on the other side is a troubled homeland that needs to be guarded. This is a difficult choice, just like being asked to choose between one's biological mother and one's adoptive mother. In the midst of this tearing, Changez's perception of identity is thrown into chaos. He confesses that "I feel like a ghost floating in the air, belonging neither here nor there" (123), just like a lone boat without an anchor in the ocean, losing its direction in the waves. According to Turner, identity aggregation should be a process of 'reintegration' in which an individual "regains a clear place in the social structure and cultural categorization through a series of rituals or symbolic acts" (Betwixt and Between 340). However, Changez's experience breaks this theoretical presupposition: after experiencing the tearing of the separation stage and the wandering of the threshold stage, his identity does not move towards convergence in the traditional sense, but rather derives from a pseudo-convergence practice of identity full of contradictions and struggles, with the persistent state of liminality as its root.

Hamid's ingenious use of metaphors in the novel gives the text a rich symbolic meaning. The name of the main character, Changez, harmonizes with the English word "change", which is not accidental, but implies that his life is always in a state of fluctuating thresholds: the decline of his family has caused him to fall in social status, studying abroad has brought about the reshaping of his cultural identity, and the elite education in the United States has altered his way of thinking, while the events of 9/11 have caused him to be rejected and isolated by the mainstream society. The events of 9/11 have made him suffer from the rejection and isolation of the mainstream society. In

addition, the Lahore teahouse plays a special role in the story, which is like a laboratory for East-West dialogue, witnessing the silent resistance of the liminal individual to Western hegemony. The dialogue between Changez and the American travellers in the teahouse is in fact a microcosm of the cultural collision and dialogue between East and West. As the conversation draws to a close, Changez extends his hand in a friendly manner, only to find the other man poking his hand into his jacket, his alert movements contrasting sharply with the cold light of the flashing metal. Slightly sarcastically, he asks, "But why do you put your hand inside your jacket, sir? I see a cold flash of metal. Given that you and I have developed a mutual intimacy, I believe that would be your card holder" (128). This tension-filled interaction, which ultimately ends unhappily, is an apt metaphor for the unbridgeable divide and crisis of trust in East-West relations.

In *The Reluctant Fundamentalist*, Hamid takes the identity dilemma of Changez as a starting point to profoundly show the complexity of the identity construction of transnational diaspora groups under the wave of globalization. The protagonist, who initially travelled to another country with the longing for the American dream, gradually awakens to be a cultural hybrid in the midst of cultural collision, and the evolution of his identity is always wandering on a blurred border: neither can he be truly accepted by the American hegemonic culture, nor is it possible for him to return to his homeland's original state of cultural purity. The continuous transformation of Changez's identity is in fact a vivid interpretation of the threshold theory. His experience breaks the framework of traditional nationalist single identity and explores a brand new path of identity politics--in the seam of East and West cultures, he neither blindly adheres to a certain side nor succumbs to any hegemonic discourse, but rather searches for the possibility of transcending the intrinsic cultural boundaries in the continuous flow of identities, an exploration that offers an understanding of contemporary transnational identities, and this exploration provides a highly inspiring perspective.

4.Conclusion

In Strangers in a Lahore Tea House, Hamid uses the three key spatial migrations of Changez to tear open the cracks in the traditional framework of liminality theory. According to Van Gennep and Turner's classic statement, identity transformation follows a linear trajectory of separation- liminality- aggregation, but in the post-colonial context, this logic has been challenged by reality. At the beginning of the novel, Changez leaves Lahore to study in the U.S. This seemingly active stripping of identity is in fact a passive choice pulled by the narrative of the American dream. He thinks he has broken free of his aristocratic identity, but he unexpectedly steps into a deeper identity maze, and the 9/11 incident becomes a pivotal point, turning the once glorified land of angels into a hostile land of demons. Trapped between the high pressure of a New York interrogation room and the native atmosphere of a Lahore teahouse, Changez is completely reduced to Turner's propertyless transitional body, losing his clear identity coordinates in the gap between culture and power. And the aggregation of identities in Changez is nothing more than a false appearance. The Lahore teahouse does not serve as a gateway for him to return to a stable identity, but instead witnesses the eternal unfinished nature of identity construction. Through the unclosed narrative structure, Hamid conveys a profound reality: for transnational migrants, the essence of identity is precisely the state of continuous mobility. In the fracture where power discourse and cultural differences are intertwined, perhaps the only way to find one's own spiritual dwelling place in the ever-changing world is to accept and hold on to the ambiguity of this liminal zone.

References

[1] Bhandari, Nagendra Bahadur. “Homi K. Bhabha’s third space theory and cultural identity today: a critical review.” Prithvi Academic Journal (2022): 171-181.

[2] Hamid, Mohsin. *The Reluctant Fundamentalist*. Anchor Canada, 2009.

[3] Hughes, Richard T. *Myths America Lives By: White Supremacy and the Stories that Give Us Meaning*. University of Illinois Press, 2021.

[4] Lefebvre, Henri. “The production of space (1991).” The people, place, and space reader. Routledge, 2014. 289-293.

[5] Morton, S. Terrorism, orientalism and imperialism. Wasafiri,2007 22 (2) : 36—42.

[6] Majid Mgamis, and Nadia Mohammad. “Muslim Identity Fluidities and Ambiguities: A Focus on Mohsin Hamid’s The Reluctant Fundamentalist and Elif Shafak’s The Forty Rules of Love.” Theory and Practice in Language Studies 14.8 (2024):2289-2295.

[7] Randall，M. *9/11 and the Literature of Terror.* Edinburgh: Edinburgh University Press, 2021.

[8] Said，E. They Call All Resistance “Terrorism”［Z/OL］. International Socialist Review.（2001—07/08）［2016—05—21］. http: / /www. Is review. org /issues /19 / Said_part2. shtmll.

[9] Turner, Victor, “Betwixt and Between: The Liminal Period in Rites of Passage.” Betwixt and between: Patterns of masculine and feminine initiation (1987): 3-19.

[10] Roger Abrahams, and Alfred Harris. The Ritual Process: Structure and Anti-structure. Routledge, 2020.

[11] The Forest of Symbols: Aspects of Ndembu Ritual. Trans. Zhao Yuyan, et al. Beijing: The Commercial Press, 2021.

[12] Dramas, fields, and metaphors: Symbolic action in human society. Cornell University Press, 2020.

[13] Van Gennep, Arnold. *The Rites of Passage*. University of Chicago press, 2020.

[14] Chen Shuping. Original Sin-Imagination-Redemption-Threshold Representation in Coleridge’s Ancient Boat Song[J]. Foreign Literature, 2025, (01):109-116. (In Chinese)

[15] Chen Xiaoming. Identity Politics in ‘The Door’: On the Transformation of Threshold in Faulkner’s Works[J]. Foreign Literature Research,2024,46(01):116-127. (In Chinese)

[16] Zhang Hualong. “9-11 Literature: The Aesthetic Turn in American and British Literature in the New Century?” [J]. Journal of Shenzhen University, 2014,(2): 20-25. (In Chinese)

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 45-55 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018632

AI-Powered Precision Medicine: Transforming Healthcare through Intelligent Imaging and Surgical Ecosystem Innovation

Chunlei Wang [1*], Jie Cao [1], Manzhi Xia [1], Jianying Kang [1], Jinlian Liang [1]

[1] Shaoxing Maternity and Child Health Care Hospital, Shaoxing, Zhejiang, China
*Corresponding author Email: wangchunlei1972@126.com
Received 8 May 2025; Accepted 11 July 2025; Published 1 September 2025

Abstract: The integration of artificial intelligence (AI) into precision medicine has revolutionized healthcare by enhancing diagnostic accuracy, optimizing treatment strategies, and improving surgical outcomes. This paper explores the transformative potential of AI in precision medicine, with a focus on intelligent imaging analysis and surgical ecosystem innovation. AI-driven techniques, such as machine learning and deep learning, have demonstrated remarkable capabilities in analyzing medical images, enabling early and accurate disease detection, particularly in cancer and cardiovascular conditions. Additionally, AI has significantly advanced surgical precision through robotic-assisted procedures and augmented reality, reducing complications and improving patient recovery. The paper also highlights the integration of diverse data sources, including genomics and wearable sensors, to provide comprehensive patient insights. Despite these advancements, challenges such as ethical considerations, data privacy, and algorithmic bias remain critical barriers to widespread adoption. The paper concludes by emphasizing the need for interdisciplinary collaboration, robust validation, and regulatory oversight to fully realize the potential of AI in precision medicine. By addressing these challenges, AI-powered precision medicine holds immense promise for delivering personalized, efficient, and equitable healthcare solutions.

Keywords: Artificial Intelligence, Precision Medicine, Medical Imaging, Surgical Innovation, Data Integration

Introduction

The advent of artificial intelligence (AI) in healthcare has ushered in a new era of precision medicine, fundamentally transforming the way diseases are diagnosed, treated, and managed. AI-powered precision medicine leverages advanced computational techniques to analyze vast amounts of data, enabling personalized healthcare solutions tailored to individual patients. This approach holds immense promise in improving patient outcomes, reducing healthcare costs, and enhancing the overall efficiency of healthcare systems. By integrating AI with medical imaging, genomics, and clinical data, precision medicine is poised to revolutionize healthcare delivery, particularly in the fields of intelligent imaging analysis and surgical ecosystem transformation [1], [2].

One of the most significant contributions of AI in precision medicine is its ability to enhance intelligent imaging analysis. Medical imaging, including radiography, computed tomography (CT), and magnetic resonance imaging (MRI), plays a crucial role in the diagnosis and monitoring of various diseases. However, the interpretation of these images often requires significant expertise and can be subject to human error. AI algorithms, particularly those based on machine learning and deep learning, have demonstrated remarkable capabilities in analyzing medical images with high accuracy and efficiency. For instance, AI-driven radiogenomics, a field that combines imaging data with genomic information, has emerged as a powerful tool in cancer diagnosis and treatment. By correlating

imaging phenotypes with gene expression profiles, AI can provide non-invasive insights into tumor biology, enabling more precise and personalized treatment strategies [2].

The application of AI in imaging analysis extends beyond cancer to other areas of medicine. For example, in cervical cancer screening, AI-based methods have been developed to identify the transformation zone (TZ) during colposcopy examinations. The TZ is a critical area where precancerous lesions are most likely to occur, and accurate identification of this region is essential for effective diagnosis and treatment. A recent multicenter validation study demonstrated that an AI-based identification system could classify and delineate the TZ with high accuracy, providing valuable assistance to colposcopists and improving the precision of colposcopic examinations [3]. This highlights the potential of AI to enhance diagnostic accuracy and reduce the burden on healthcare professionals, particularly in resource-limited settings.

In addition to imaging analysis, AI is transforming the surgical ecosystem by enabling more precise and efficient surgical interventions. The integration of AI with surgical technologies, such as robotic-assisted surgery and augmented reality, has the potential to enhance surgical precision, reduce complications, and improve patient outcomes. AI algorithms can analyze preoperative imaging data to create detailed surgical plans, guide surgeons during procedures, and provide real-time feedback on surgical performance. Furthermore, AI-powered predictive models can assess patient risk factors and predict postoperative outcomes, enabling more informed decision-making and personalized surgical care [1].

The transformative potential of AI in precision medicine is further amplified by its ability to integrate and analyze diverse data sources. Data fusion technologies, which combine data from electronic health records, wearable sensors, genomics databases, and other sources, provide a comprehensive view of a patient's health status. This holistic approach enables more accurate diagnosis, personalized treatment plans, and continuous monitoring of patient health. For example, AI-driven data fusion centers have been implemented in healthcare systems to integrate and analyze data from multiple sources, providing clinicians with actionable insights and improving the overall efficiency of healthcare delivery [1].

Despite the significant advancements in AI-powered precision medicine, several challenges remain. Ethical considerations, such as data privacy and algorithmic bias, must be addressed to ensure the responsible and equitable use of AI in healthcare. To address these challenges, several practical frameworks and mitigation strategies are being developed and implemented. For algorithmic bias, the healthcare AI community has established bias detection protocols including fairness-aware machine learning techniques, diverse dataset curation strategies, and algorithmic auditing frameworks such as the AI Fairness 360 toolkit. Data privacy concerns are being addressed through federated learning approaches that enable AI model training without centralizing sensitive patient data, differential privacy techniques that add mathematical noise to protect individual privacy while preserving analytical utility, and blockchain-based systems for secure, auditable data sharing. Regulatory frameworks such as the FDA's Software as Medical Device guidance and the European Union's AI Act provide structured pathways for AI validation and deployment. Furthermore, multi-stakeholder governance models involving clinicians, ethicists, technologists, and patient advocates are being established to ensure comprehensive oversight. Additionally, the integration of AI into clinical practice requires robust validation and regulatory oversight to ensure the safety and efficacy of AI-driven interventions. Collaborative efforts between researchers, clinicians, and policymakers are essential to overcome these challenges and realize the full potential of AI in precision medicine [1],[2].

In conclusion, AI-powered precision medicine represents a paradigm shift in healthcare, offering transformative potential in intelligent imaging analysis and surgical ecosystem transformation. By leveraging advanced computational techniques and integrating diverse data sources, AI enables more accurate diagnosis, personalized treatment, and continuous monitoring of patient health. While challenges remain, the continued development and implementation of AI-driven solutions hold immense promise for improving patient outcomes and enhancing the efficiency of healthcare systems. This review explores the current state of AI-powered precision medicine, with a

focus on its applications in imaging analysis and surgical care, and discusses the opportunities and challenges associated with its clinical transformation.

Background and Context

Precision medicine has emerged as a transformative approach in healthcare, aiming to tailor medical treatment to the individual characteristics of each patient. This paradigm shift from a one-size-fits-all model to personalized care has been driven by advancements in genomics, biotechnology, and data analytics. Historically, the concept of precision medicine can be traced back to the early 20th century, when the understanding of genetic inheritance began to take shape. However, it was not until the completion of the Human Genome Project in 2003 that the potential for precision medicine truly began to be realized. This monumental achievement provided a comprehensive map of human genes, paving the way for the identification of genetic markers associated with various diseases and the development of targeted therapies [4].

The integration of artificial intelligence (AI) into healthcare has further accelerated the evolution of precision medicine. AI, particularly through machine learning (ML) and deep learning (DL) algorithms, has demonstrated remarkable capabilities in analyzing complex datasets, identifying patterns, and making predictions with high accuracy. In the context of precision medicine, AI has been instrumental in enhancing diagnostic accuracy, predicting disease risk, and optimizing treatment plans. For instance, AI-driven medical image recognition has significantly improved the early diagnosis of diseases such as cancer, cardiovascular conditions, and neuropsychiatric disorders [5]. By leveraging advanced technologies like convolutional neural networks (CNNs), AI can analyze medical images with a level of precision that surpasses traditional methods, enabling earlier and more accurate detection of abnormalities.

The evolution of medical imaging and surgical techniques has also played a crucial role in the advancement of precision medicine. Medical imaging, which includes modalities such as X-rays, computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET), has undergone significant technological improvements over the past few decades. These advancements have not only enhanced the resolution and clarity of images but have also enabled the integration of AI for more sophisticated analysis. For example, AI techniques like radiomics and radiogenomics have been employed to extract quantitative features from medical images, providing insights into tumor biology and heterogeneity. This has led to the development of personalized image-guided precision medicine strategies, particularly in the treatment of metastatic cutaneous melanoma [4].

In the realm of surgery, the integration of AI and augmented reality (AR) has revolutionized traditional practices, enabling more precise and minimally invasive procedures. AI-driven robotic-assisted surgery, combined with AR visualization, provides surgeons with real-time guidance and enhanced visualization of anatomical structures. This integration has not only improved the accuracy of surgical interventions but has also reduced complications and recovery times. For instance, AI algorithms can analyze preoperative imaging data to create personalized surgical plans, while AR overlays provide intraoperative navigation, ensuring that surgeons can perform complex procedures with greater confidence and precision [6].

The application of AI in precision medicine extends beyond diagnostics and surgery to encompass various aspects of healthcare management. The concept of digital twins (DT), which involves creating virtual replicas of physical entities, has gained traction in recent years. In healthcare, DT models can simulate patient-specific conditions, enabling clinicians to predict disease progression, test treatment options, and optimize therapeutic strategies. The integration of AI with DT has shown immense potential in genomics, clinical cancer treatment, and molecular imaging. However, challenges such as system bias, algorithm transparency, and data privacy must be addressed to fully realize the benefits of this technology [7].

AI has also made significant contributions to neonatal surgery and healthcare analysis. In neonatal care, AI-driven approaches have enabled early detection of congenital anomalies, prediction of disease progression, and optimization of surgical interventions. Machine learning models trained on large datasets of neonatal health records can identify risk factors and recommend personalized treatment plans, improving outcomes for vulnerable infants. Additionally, natural language processing (NLP) has enhanced clinical documentation, reducing administrative burdens and improving the efficiency of healthcare systems [8].

Despite the numerous advancements, the integration of AI in precision medicine and healthcare is not without challenges. Technical standards, data security, and privacy protection are critical issues that must be addressed to ensure the responsible implementation of AI technologies. Establishing effective technical standards and evaluation systems is essential to maintain the quality and reliability of AI-driven healthcare solutions. Furthermore, interdisciplinary collaboration among healthcare professionals, data scientists, and policymakers is crucial to overcome barriers and foster innovation in the field [5].

Ethical considerations also play a significant role in the adoption of AI in healthcare. Issues such as algorithmic bias, equitable access to AI-driven treatments, and the potential for misuse of patient data must be carefully managed to ensure that the benefits of AI are distributed fairly and responsibly. Addressing these ethical challenges requires a collaborative effort involving stakeholders from various sectors, including academia, industry, and government. Specific mitigation strategies currently being implemented include the development of algorithmic impact assessments that evaluate potential bias before deployment, the establishment of ethics review boards specifically for AI applications in healthcare, and the creation of transparent reporting standards such as the CONSORT-AI and SPIRIT-AI guidelines for clinical trials involving AI interventions. Professional medical societies are developing AI ethics training modules for healthcare providers, while institutions are implementing algorithmic accountability measures including regular bias audits and performance monitoring across diverse patient populations. International organizations such as the World Health Organization have published ethics guidelines for AI in health, providing frameworks for responsible development and deployment. These comprehensive approaches ensure that technological advancement proceeds alongside ethical responsibility [8].

In conclusion, the integration of AI into precision medicine and healthcare has the potential to revolutionize the way we diagnose, treat, and manage diseases. The historical evolution of medical imaging and surgical techniques, combined with the transformative capabilities of AI, has set the stage for a new era of personalized and data-driven healthcare. As we continue to explore the possibilities of AI in healthcare, it is imperative to address the challenges and ethical considerations to ensure that these technologies are used responsibly and equitably. The future of precision medicine lies in the seamless integration of AI with other advanced technologies, fostering innovation and improving patient outcomes across the globe [4], [6], [7], [8].

Methodology Review

The methodologies employed in AI-driven medical imaging and surgical applications have seen significant advancements in recent years, driven by the integration of machine learning (ML) algorithms, deep learning (DL) models, and sophisticated data processing techniques. These technologies have revolutionized the field, enabling more accurate diagnoses, improved surgical outcomes, and enhanced decision-making processes in clinical settings. This section reviews the key methodologies and their applications in these domains, highlighting their transformative potential and the challenges that remain.

Machine Learning Algorithms in Medical Imaging

Machine learning algorithms have become a cornerstone in medical imaging, particularly in tasks such as image classification, segmentation, and detection. Traditional ML techniques, such as support vector machines (SVMs) and random forests, have been widely used for their interpretability and efficiency in handling structured data. However,

the advent of deep learning has shifted the focus towards more complex models capable of processing unstructured data, such as medical images. For instance, in the diagnosis of hepatocellular carcinoma (HCC), ML algorithms have demonstrated superior predictive capabilities compared to standard models, enabling early detection and improved risk stratification [9]. Radiomics, a quantitative method that extracts features from medical images, has been particularly effective in liver imaging, aiding in the diagnosis and prognostication of HCC [9].

Deep Learning Models in Medical Imaging

Deep learning models, particularly convolutional neural networks (CNNs), have emerged as the state-of-the-art in medical imaging due to their ability to automatically learn hierarchical features from raw data. CNNs have been successfully applied to various imaging modalities, including X-rays, computed tomography (CT), and magnetic resonance imaging (MRI). For example, in the detection of cardiac regions in chest X-ray images, the ResNet-50 architecture has been identified as the optimal model for precise localization, achieving remarkable accuracy in predicting bounding box coordinates [10]. Similarly, in the context of COVID-19 diagnosis, deep learning models have been employed to distinguish lesions from other parts of the lung, providing rapid and accurate results without human intervention [11].

The integration of deep learning into multi-disease diagnosis systems, such as HealthScan AI, has further expanded the scope of AI in medical imaging. HealthScan AI utilizes six distinct CNN models to identify diseases like COVID-19, pneumonia, and glaucoma from chest X-rays and internal eye scans, demonstrating high accuracy and efficiency [12]. These models are integrated into user-friendly interfaces, enabling real-time diagnostic feedback and empowering healthcare professionals to make informed decisions.

Data Processing Techniques in Medical Imaging

Data processing techniques play a crucial role in the success of AI-driven medical imaging applications. Preprocessing steps, such as intelligent scaling, normalization, and data augmentation, are essential for enhancing the quality of input data and improving model performance. For instance, in the study on cardiac region detection, a comprehensive preprocessing pipeline was implemented to ensure robust performance across diverse clinical scenarios [10]. Similarly, in the context of COVID-19 diagnosis, advanced data augmentation techniques have been employed to enhance the generalizability of deep learning models, enabling them to handle variations in lesion size and shape [11].

The importance of proper training and validation of machine learning models cannot be overstated. The American Association of Physicists in Medicine (AAPM) has emphasized the need for rigorous validation to ensure the generalizability and reliability of AI models in clinical settings [13]. This includes the use of diverse datasets, cross-validation techniques, and performance metrics to assess model accuracy, sensitivity, and specificity. The AAPM also highlights the importance of user training and quality assurance in the deployment of AI systems, ensuring that they are effectively integrated into clinical workflows [13].

AI in Surgical Applications

AI methodologies have also made significant inroads into surgical applications, both in preoperative planning and intraoperative assistance. In the context of HCC, AI models have been used to predict surgical outcomes and assist in the resection of complex lesions, providing real-time feedback to surgeons [9]. The use of AI in surgery extends to the recognition of surgical actions, where fine-grained analysis of surgical workflows can enhance safety and efficiency. The CholecTriplet2021 challenge, for instance, focused on the recognition of surgical action triplets (instrument, verb, target) in laparoscopic videos, achieving mean average precision (mAP) ranging from 4.2% to 38.1% [14]. This highlights the potential of AI in providing context-aware decision support in the operating room, although challenges remain in achieving higher accuracy and robustness.

Challenges and Future Directions

Despite the significant advancements in AI-driven medical imaging and surgical applications, several challenges persist. One of the primary challenges is the need for large, diverse, and annotated datasets to train and validate AI

models. The generalizability of these models is often limited by the heterogeneity of medical data, which can vary across different populations, imaging modalities, and clinical settings [13]. Additionally, the interpretability of AI models remains a concern, particularly in high-stakes applications where clinical decisions must be transparent and explainable.

To address these challenges, several practical strategies are being implemented. Data standardization initiatives such as the FAIR (Findable, Accessible, Interoperable, Reusable) data principles are being adopted to improve data quality and sharing. Multi-institutional data consortiums are being established to create larger, more diverse datasets while maintaining privacy through techniques such as differential privacy and secure multi-party computation. For model interpretability, explainable AI techniques including attention mechanisms, gradient-based attribution methods, and post-hoc explanation tools are being integrated into medical AI systems. Regulatory bodies are developing validation frameworks specifically for AI in medical imaging, including guidance on acceptable performance metrics, testing protocols, and post-market surveillance requirements.

Future research should focus on addressing these challenges by developing more robust and interpretable AI models, as well as improving data sharing and collaboration across institutions. The integration of AI into clinical workflows also requires careful consideration of ethical and regulatory issues, ensuring that these technologies are used responsibly and equitably. Specific areas of focus include the development of adaptive learning systems that can continuously improve from new data while maintaining safety and efficacy, the creation of AI systems that can handle multi-modal data integration seamlessly, and the establishment of global standards for AI validation and deployment in healthcare settings. As AI continues to evolve, it is expected to play an increasingly central role in medical imaging and surgery, driving further progress in the diagnosis and management of complex disease processes [9], [11], [13].

In conclusion, the methodologies employed in AI-driven medical imaging and surgical applications have demonstrated significant potential in improving diagnostic accuracy, enhancing surgical outcomes, and supporting clinical decision-making. Machine learning algorithms, deep learning models, and advanced data processing techniques have been instrumental in achieving these advancements, although challenges remain in ensuring their generalizability, interpretability, and integration into clinical practice. As the field continues to evolve, AI is expected to play a transformative role in healthcare, driving innovation and improving patient outcomes.

Key Findings and Analysis

The integration of artificial intelligence (AI) into precision medicine has revolutionized the field, offering unprecedented advancements in diagnostic accuracy, treatment planning, and surgical outcomes. Recent studies have demonstrated the transformative potential of AI across various medical disciplines, including thoracic surgery, otolaryngology, and urology. This section analyzes key findings from these studies, supported by empirical evidence, to evaluate the impact of AI on precision medicine.

In the context of thoracic surgery, AI has significantly enhanced the diagnosis and management of complex conditions such as thoracic empyema. As highlighted by [15], AI and machine learning (ML) models have been applied to CT scans and chest X-rays to identify and classify pleural effusions and empyema with greater accuracy than traditional methods. These AI-driven analyses can detect intricate imaging features often missed by the human eye, thereby improving diagnostic precision. Furthermore, AI-based decision-support algorithms have been shown to reduce the time to diagnosis, optimize antibiotic stewardship, and facilitate more precise and less invasive surgical interventions. These advancements have led to improved clinical outcomes, reduced inpatient hospital stays, and better long-term patient management. The ability of AI to analyze large datasets and recognize complex patterns underscores its potential to enhance preoperative planning and optimize surgical strategies, ultimately transforming the management of thoracic empyema.

Similarly, the integration of AI into otolaryngology has opened new avenues for enhancing diagnostic precision and treatment strategies. [16] emphasizes the diverse applications of AI in this field, ranging from predicting hearing

loss progression and optimizing cochlear implant settings to managing chronic sinusitis and predicting the success of treatments for obstructive sleep apnea. AI-driven tools have enabled otolaryngologists to leverage advanced diagnostic capabilities, improve patient monitoring, and refine surgical planning. However, the successful integration of AI in otolaryngology necessitates a paradigm shift in educational frameworks. Training programs must now incorporate AI literacy alongside traditional clinical skills, ensuring that practitioners are equipped to harness the full potential of AI. Continuous education through workshops and seminars is also essential to keep otolaryngologists updated on the latest AI tools and applications. By fostering a collaborative approach to address ethical considerations and ensure responsible AI integration, the otolaryngology community can fully embrace AI-driven healthcare innovations.

In urological surgery, AI has emerged as a transformative force, particularly in the realm of autonomous robotic surgery. [17] highlights the remarkable diagnostic accuracy of AI systems, with some achieving up to 99.38% in detecting prostate cancer. AI facilitates real-time anatomical recognition and instrument delineation, significantly increasing surgical precision. While current robotic systems operate under human supervision, ongoing research aims to advance autonomous surgical capabilities. The potential for AI to improve surgical outcomes in urology is immense, but challenges related to autonomy, safety, and ethics remain. Addressing these challenges is crucial to realizing the full potential of AI in robotic surgery. The integration of AI into urological practice not only enhances diagnostic and surgical precision but also paves the way for more personalized and effective treatment strategies.

The empirical evidence from these studies underscores the profound impact of AI on precision medicine. In thoracic surgery, AI has improved diagnostic accuracy and optimized treatment planning, leading to better clinical outcomes and reduced hospital stays. In otolaryngology, AI has enhanced diagnostic precision and treatment strategies, necessitating a shift in educational frameworks to ensure practitioners are equipped to leverage AI tools. In urological surgery, AI has increased surgical precision and diagnostic accuracy, with ongoing research focused on advancing autonomous capabilities. Across these disciplines, AI has demonstrated its potential to transform healthcare by improving patient outcomes, optimizing treatment strategies, and enhancing surgical precision.

However, the integration of AI into precision medicine is not without challenges. Ethical considerations, safety concerns, and the need for continuous education and training must be addressed to ensure responsible AI integration. Collaborative efforts among researchers, clinicians, and educators are essential to overcome these challenges and fully harness the potential of AI in precision medicine. As the field continues to evolve, the commitment to advancing AI-driven healthcare innovations will be paramount in shaping the future of precision medicine.

In conclusion, the key findings from recent studies on AI in precision medicine highlight its transformative potential across various medical disciplines. AI has significantly enhanced diagnostic accuracy, treatment planning, and surgical outcomes, supported by empirical evidence. The integration of AI into thoracic surgery, otolaryngology, and urological surgery has led to improved clinical outcomes, optimized treatment strategies, and increased surgical precision. While challenges remain, the continued advancement of AI-driven healthcare innovations holds great promise for the future of precision medicine.

Future Directions

The field of AI-powered precision medicine is poised for transformative advancements, driven by emerging technologies, ethical considerations, and the imperative for interdisciplinary collaboration. As highlighted by [18], AI has already demonstrated its potential to revolutionize drug discovery, clinical trials, and patient care by enhancing decision-making across various disciplines, including medicinal chemistry, molecular biology, and clinical practice. However, unlocking its full potential requires addressing critical challenges such as data quality, privacy concerns, algorithmic biases, and ethical dilemmas. Future developments must focus on creating well-annotated, large-scale

datasets that are both diverse and representative to ensure the reliability and generalizability of AI models. Additionally, advancements in AI algorithms, particularly in deep learning and reinforcement learning, could further refine treatment personalization by accounting for dynamic patient responses and complex drug interactions [18].

Emerging technologies such as liquid biopsies and advanced imaging techniques are also expected to play a pivotal role in the evolution of AI-powered precision medicine. As discussed by [19], liquid biopsies, including circulating tumor DNA and exosomes, offer a non-invasive method for early detection, treatment monitoring, and recurrence prediction in diseases like pancreatic cancer. Integrating these biomarkers with AI-driven analytics could enable real-time, data-driven decision-making, thereby improving diagnostic accuracy and treatment outcomes. Furthermore, AI's application in medical imaging and biomarker discovery holds promise for identifying subtle patterns and correlations that may elude human analysis, thus enhancing early detection and intervention strategies [19].

Ethical considerations remain a cornerstone of future developments in this field. The integration of AI in healthcare raises concerns about data privacy, algorithmic transparency, and equitable access to advanced technologies. Addressing these issues requires robust regulatory frameworks and ethical guidelines to ensure that AI applications are both safe and equitable. Practical implementation strategies include the development of AI ethics by design principles that embed ethical considerations into the development lifecycle, establishment of real-time monitoring systems for algorithmic fairness and safety, and creation of patient-centered AI governance models that include patient representatives in decision-making processes. Technical solutions such as explainable AI architectures are being developed to enhance algorithmic transparency, while federated learning and homomorphic encryption technologies enable privacy-preserving AI development. International standards organizations are working on global harmonization of AI safety and ethics standards, including ISO/IEC 23053 for AI risk management and IEEE standards for ethical AI design. Furthermore, public-private partnerships are being formed to ensure equitable access to AI-driven healthcare innovations across diverse populations and geographic regions. Collaborative efforts between policymakers, ethicists, and technologists will be essential to navigate these challenges and foster public trust in AI-driven healthcare solutions [18], [19].

Interdisciplinary collaboration is another critical factor for advancing AI-powered precision medicine. The complexity of healthcare challenges necessitates the integration of expertise from diverse fields, including medicine, computer science, bioinformatics, and ethics. By fostering partnerships between clinicians, researchers, and AI experts, the field can develop innovative solutions that are both scientifically rigorous and clinically relevant. Such collaborations will also facilitate the translation of AI research into practical applications, ultimately improving patient outcomes and healthcare sustainability [18], [19].

In conclusion, the future of AI-powered precision medicine is bright, with emerging technologies and interdisciplinary collaboration driving progress. However, realizing its full potential will require addressing ethical concerns, improving data quality, and fostering partnerships across disciplines. By doing so, the field can overcome current limitations and deliver on its promise to transform healthcare, making it more personalized, efficient, and accessible for all.

Conclusion

AI-powered precision medicine represents a paradigm shift in healthcare delivery, demonstrating remarkable potential to transform diagnostic accuracy, treatment personalization, and surgical outcomes across multiple medical disciplines. This comprehensive review has examined the integration of artificial intelligence technologies into healthcare systems, with particular emphasis on intelligent imaging analysis and surgical ecosystem innovation.

The evidence presented demonstrates that AI-driven methodologies have achieved significant breakthroughs in

medical imaging applications, from enhancing cancer detection and cardiovascular condition diagnosis to improving surgical precision through robotic-assisted procedures and augmented reality integration. Machine learning and deep learning algorithms have proven capable of analyzing medical images with accuracy levels that often surpass traditional methods, enabling earlier disease detection and more precise treatment planning. The successful implementation of AI systems in diverse medical specialties, including thoracic surgery, otolaryngology, and urology, underscores the broad applicability and clinical relevance of these technologies.

The integration of diverse data sources through AI-powered data fusion technologies has emerged as a critical advancement, enabling comprehensive patient insights through the combination of electronic health records, genomic data, wearable sensor information, and imaging studies. This holistic approach facilitates more accurate diagnosis, personalized treatment strategies, and continuous health monitoring, ultimately leading to improved patient outcomes and enhanced healthcare efficiency.

However, the path toward widespread adoption of AI in precision medicine requires careful attention to significant challenges that remain. Ethical considerations, including algorithmic bias, data privacy, and equitable access to AI-driven healthcare solutions, must be addressed through comprehensive frameworks and practical implementation strategies. The development of bias detection protocols, fairness-aware machine learning techniques, and transparent governance models represents critical steps toward responsible AI deployment. Technical solutions such as federated learning, differential privacy, and explainable AI architectures provide pathways for addressing privacy concerns while maintaining algorithmic transparency.

The regulatory landscape continues to evolve, with frameworks such as the FDA's Software as Medical Device guidance and the European Union's AI Act providing structured approaches for AI validation and deployment. Multi-stakeholder governance models that include clinicians, ethicists, technologists, and patient advocates are essential for ensuring comprehensive oversight and maintaining public trust in AI-driven healthcare innovations.

Future progress in AI-powered precision medicine will depend significantly on interdisciplinary collaboration among healthcare professionals, data scientists, policymakers, and ethicists. The complexity of healthcare challenges necessitates integrated expertise from diverse fields to develop solutions that are both scientifically rigorous and clinically applicable. Emerging technologies, including liquid biopsies and advanced biomarker discovery platforms, offer additional opportunities for enhancing diagnostic capabilities and treatment monitoring when combined with AI-driven analytics.

The continued advancement of AI in precision medicine holds immense promise for delivering more personalized, efficient, and equitable healthcare solutions. By addressing current challenges through practical frameworks, maintaining focus on ethical considerations, and fostering collaborative partnerships across disciplines, the healthcare community can fully realize the transformative potential of AI technologies. This evolution toward AI-powered precision medicine represents not merely a technological advancement, but a fundamental reimagining of healthcare delivery that prioritizes individual patient needs while enhancing overall system efficiency and accessibility.

Data Availability Statement

The review article does not involve research data.

Funding Statement

Shaoxing Health and Wellness Science and Technology Program.(2023SKY051)

References

[1] Shohoni Mahabub, Bimol Chandra Das, Md Russel Hossain. 2024. Advancing healthcare transformation: AI-driven precision medicine and scalable innovations through data analytics. In Edelweiss Applied Science and Technology.
[2] Yusheng Guo, Tianxiang Li, Bingxin Gong, Yan Hu, Sichen Wang, Lian Yang, Chuansheng Zheng. 2024. From Images to Genes: Radiogenomics Based on Artificial Intelligence to Achieve Non‐Invasive Precision Medicine in Cancer Patients. In Advancement of science.
[3] Tong Wu, Yuting Wang, Xiaoli Cui, Peng Xue, Youlin Qiao. 2025. AI-Based Identification Method for Cervical Transformation Zone Within Digital Colposcopy: Development and Multicenter Validation Study.. In .
[4] Suvarna U Patel, Pranita S. Jirvankar. 2024. AI in Healthcare – Precision Medicine and Diagnosis. In 2024 2nd DMIHER International Conference on Artificial Intelligence in Healthcare, Education and Industry (IDICAIEI).
[5] Yuting Lee. 2024. Application of AI-Driven Medical Image Recognition in Precision Medicine and Healthcare. In Applied and Computational Engineering.
[6] Ezgi Ağır. 2024. Advanced AI and Augmented Reality (AR) Integration in Medical and Surgical Practice. In Next Frontier For Life Sciences and AI.
[7] Atique Ahmed, Khadija Shoukat, Muhammad Ahmad Muneeb, Doaa Abdo Othman All Qasem, Muhammad Adeel Shahzad, Laraib Ul Nissa, Rabia Amir, Muhammad Zubair, Muhammad Waqas Younas, Asad Ali. 2024. AI and Digital Twin Transforms in the Construction of Precision Medical Model: Healthcare Management in Smart Cities. In European Journal of Medical and Health Research.
[8] Jaswinder Singh, Gaurav Dhiman. 2025. A Survey on Artificial Intelligence in Precision Medicine and Healthcare Analysis for Neonatal Surgery. In Journal of Neonatal Surgery.
[9] Carolina Larrain, Alejandro Torres-Hernandez, D. B. Hewitt. 2024. Artificial Intelligence, Machine Learning, and Deep Learning in the Diagnosis and Management of Hepatocellular Carcinoma. In Livers.
[10] Narendra Rathod, Kriti Awasthi, Magdalena Kostkiewicz, Piotr Klimeczek, E. Stępień. 2024. Advancing Cardiac Detection in Chest X-ray Images Using Machine Learning: A Practical Application of AI in Medical Imaging. In Bio-Algorithms and Med-Systems.
[11] Sayed Amir Mousavi Mobarakeh, K. Kazemi, A. Aarabi, H. Danyali. 2024. Empowering Medical Imaging with Artificial Intelligence: A Review of Machine Learning Approaches for the Detection, and Segmentation of COVID-19 Using Radiographic and Tomographic Images. In arXiv.org.
[12] st Prudhvi, Sai Ganesh, Chakali Murthy, Gari Keerthi, Kiriti. 2024. HealthScan AI-Deep Learning-Based Multi-Disease Diagnosis from Medical Imaging. In IEEE International Symposium on Compound Semiconductors.
[13] Lubomir M. Hadjiiski, Kenny H. Cha, H. Chan, K. Drukker, L. Morra, J. Näppi, B. Sahiner, H. Yoshida, Quan Chen, T. Deserno, H. Greenspan, H. Huisman, Z. Huo, R. Mazurchuk, N. Petrick, D. Regge, Ravi K. Samala, R. Summers, Kenji Suzuki, G. Tourassi, Daniel Vergara, S. Armato. 2022. AAPM task group report 273: Recommendations on best practices for AI and machine learning for computer-aided diagnosis in medical imaging.. In Medical Physics (Lancaster).
[14] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, Armine Vardazaryan, Fangfang Xia, Zixuan Zhao, Tong Xia, Fucang Jia, Yuxuan Yang, Hao Wang, Derong Yu, Guoyan Zheng, Xiaotian Duan, Neil Getty, Ricardo Sanchez-Matilla, Maria Robu, Li Zhang, Huabin Chen, Jiacheng Wang, Liansheng Wang, Bokai Zhang, Beerend Gerats, Sista Raviteja, Rachana Sathish, Rong Tao, Satoshi Kondo, Winnie Pang, Hongliang Ren, Julian Ronald Abbing, Mohammad Hasan Sarhan, Sebastian Bodenstedt, Nithya Bhasker, Bruno Oliveira, Helena R. Torres, Li Ling, Finn Gaida, Tobias Czempiel, João L. Vilaça, Pedro Morais, Jaime Fonseca, Ruby Mae Egging, Inge Nicole Wijma, Chen Qian, Guibin Bian, Zhen Li, Velmurugan Balasubramanian, Debdoot Sheet, Imanol Luengo, Yuanbo Zhu, Shuai Ding, Jakob-Anton Aschenbrenner, Nicolas Elini van der Kar, Mengya Xu, Mobarakol Islam, Lalithkumar Seenivasan, Alexander Jenke,

Danail Stoyanov, Didier Mutter, Pietro Mascagni, Barbara Seeliger, Cristians Gonzalez, Nicolas Padoy. 2022. CholecTriplet2021: A benchmark challenge for surgical action triplet recognition. In arXiv preprint.
[15] Adam Zumla, Rizwan Ahmed, Kunal Bakhri. 2024. The role of artificial intelligence in the diagnosis, imaging, and treatment of thoracic empyema.. In Current opinion in pulmonary medicine.
[16] Bilal Irfan. 2024. Beyond the Scope: Advancing Otolaryngology With Artificial Intelligence Integration. In Cureus.
[17] Dae Young Lee, Hee Jo Yang. 2024. Artificial Intelligence for Autonomous Robotic Surgery in Urology: A Narrative Review. In Urogenital Tract Infection.
[18] Claudio Carini, A. Seyhan. 2024. Tribulations and future opportunities for artificial intelligence in precision medicine. In Journal of Translational Medicine.
[19] L. Daamen, I. Molenaar, Vincent P. Groot. 2023. Recent Advances and Future Challenges in Pancreatic Cancer Care: Early Detection, Liquid Biopsies, Precision Medicine and Artificial Intelligence. In Journal of Clinical Medicine.

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 56-64 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018720

Discussion on problems caused by uneven deployment of 5G network edge computing nodes

Jinghua Cui [1], Jiulong Zhang [1], Linluo Yao [1*]

[1] Faculty of Engineering, Universiti Malaya, Kuala Lumpur, 50603, Malaysia
*Corresponding author Email: 17621330129@163.com
Received 21 June 2025; Accepted 27 July 2025; Published 1 September 2025

Abstract: This study focuses on the issue of uneven deployment of edge computing nodes in 5G networks. A multi-regional simulation model was constructed, and four key performance indicators were evaluated: average end-to-end latency, node load balance, request success rate, and resource utilization. Various optimization techniques, including automated scheduling and network slicing, were employed to control the simulation. The simulation results show that the average latency decreased from 54.8ms to 35.9ms, the load balance increased to 0.72, the request success rate rose to 91.7%, and the resource utilization improved to 74.6%. The study demonstrates that deploying optimized control strategies can significantly alleviate the performance bottleneck caused by uneven node distribution, thereby enhancing the overall service capability and resource utilization of the edge computing system.

Keywords: edge computing; node deployment; scheduling optimization; 5G network.

INTRODUCTION

Recent studies have emphasized the need for intelligent orchestration and adaptive frameworks in 5G edge computing. For example, Daneshvar et al. (2024) proposed a GNN-based orchestration model, while Yan et al. (2023) developed multi-agent reinforcement learning for dynamic edge coordination.With the rapid deployment of 5G networks, edge computing, a key technology for enhancing network timeliness and local processing, is gradually moving towards large-scale application(Souza et al., 2025). However, due to differences in infrastructure, regional economies, and business needs, the spatial distribution of edge computing nodes is severely imbalanced, leading to reduced service performance, resource wastage, and uneven regional service capabilities(Pramanik et al., 2024). To explore the performance and optimization paths of this issue, this study has developed a multi-regional deployment simulation model to analyze the impact of uneven deployment on performance. Finally, it uses optimized scheduling strategies and key technical methods to verify the effectiveness, aiming to provide theoretical and practical references for the deployment of edge computing in 5G networks.

With the rapid deployment of 5G networks, edge computing—serving as a key enabler of ultra-low latency and distributed intelligence—has attracted extensive attention from scholars and industry stakeholders alike. Previous research has explored multi-access edge computing (MEC) architectures (Halima et al., 2024), resource orchestration frameworks (Daneshvar et al., 2024), intelligent traffic steering (Pramanik et al., 2024), and network slicing for service differentiation (Tsourdinis et al., 2024). However, the vast majority of existing studies assume idealized and homogeneous node deployment conditions, overlooking the real-world challenge of uneven node distribution due to disparities in regional infrastructure, investment, and population density (Ferenc et al., 2022; Mahmood & Rehman, 2025). Moreover, while adaptive scheduling (Yan et al., 2023) and dynamic resource allocation mechanisms (Souza et al., 2025) have shown promise, few have been validated in spatially heterogeneous edge environments that reflect realistic deployment scenarios.To bridge this gap, this study constructs a multi-regional simulation

model incorporating uneven edge node density and applies advanced optimization strategies — including Q-learning-based dynamic scheduling and network slicing — to evaluate their effectiveness in mitigating latency, load imbalance, and resource underutilization. By embedding reinforcement learning into deployment control mechanisms and validating its performance in edge-diverse contexts, this study provides new empirical evidence for adaptive and intelligent 5G edge computing systems.

1.5G edge computing node deployment status

1.1 Deployment density area and imbalance phenomenon

First-tier cities like Beijing, Shanghai, and Guangzhou have established dense clusters of edge computing nodes due to their robust communication infrastructure and high industrial demand, to meet the demands for low-latency and high-concurrency business processing(Tsourdinis et al., 2024). Thanks to strong industrial support and high user density in these cities, edge computing nodes have developed rapidly, providing efficient and stable services(Suman, 2024). In contrast, remote areas lag behind in infrastructure development, with a lack of data centers, low base station density, and insufficient coverage of edge nodes, leading to slower response times and limited application scenarios(Chang et al., 2024). These issues pose significant obstacles to the promotion and application of 5G edge computing, hindering its comprehensive development(Halima et al., 2024). Therefore, addressing these regional imbalances is crucial for advancing 5G edge computing applications(Esfandyari et al., 2025).

1.2 Uneven deployment of technology and basic conditions

The fundamental reasons for the uneven deployment of edge computing nodes include unequal infrastructure development, insufficient network backhaul capabilities, and imbalanced local computing power demands(Yan et al., 2023). High-bandwidth and high-reliability backhaul links are essential for the stable operation of edge computing nodes(Divya et al., 2022). Due to geographical constraints and a lack of adequate funding in central and western regions, it is often challenging to establish a large-scale network of edge computing nodes in these areas(Ahmed, 2022). The strong coupling between edge computing and local business needs, along with the lack of data-driven scenarios in some regions, results in low enthusiasm for edge node construction, further exacerbating the uneven deployment of edge computing nodes(Ferenc et al., 2022). As 5G networks advance, balancing infrastructure development across regions will be a critical direction for addressing this issue(P V et al., 2025).

2. Main impacts and coping strategies of unbalanced deployment

2.1 Main effects of uneven deployment

The uneven deployment of edge computing nodes in 5G networks can lead to several negative impacts, primarily manifested as latency differences, reduced data transmission efficiency, and uneven network load distribution(Liang et al., 2022). This issue is particularly severe in scenarios with high demands for low latency, such as industrial internet and vehicle networking. Due to the uneven distribution of edge nodes, the non-uniformity of latency directly affects the real-time performance and stability of critical tasks, especially in areas like intelligent manufacturing and autonomous driving, where low latency is essential. When resources are overly concentrated in core areas, it can create isolated computing islands, preventing edge nodes from fully serving local terminals and hindering the implementation and application efficiency of edge intelligence. Insufficient coverage of edge computing services not only exacerbates the shortage of network service capabilities in certain regions but also widens the 'digital divide,' affecting the fairness and overall quality of network services, and profoundly impacting the digital development of society.

2.2 Key measures to alleviate uneven deployment

To alleviate the uneven distribution of deployments, consider introducing a self-organizing deployment mechanism for edge nodes, combined with intelligent resource scheduling algorithms, to achieve on-demand deployment and dynamic migration of computing power. This strategy can flexibly adjust resource allocation based on the actual needs of different regions, thereby enhancing the coverage and efficiency of edge computing services. To prevent the island effect among edge nodes, promote the construction of collaborative architectures such as

'edge-core,' enhance the interconnectivity between edge nodes, and ensure efficient computing services even when resources are limited. At the policy level, further strengthen the mechanism for joint construction and sharing of infrastructure, encourage the allocation of funds and resources to underdeveloped areas, promote coordinated development of infrastructure across regions, and build a unified, coordinated, and efficiently operating 5G edge computing system to achieve more balanced deployment and improved service quality(Mahmood and Rehman, 2025).

3. Simulation analysis and model verification

3.1 Construction of simulation model

The simulation was conducted using the NS-3 platform, an open-source, discrete-event network simulator widely used in academic and industrial research for evaluating network protocols and architectures. NS-3 provides high-fidelity modeling of real-world network behavior, allowing researchers to simulate traffic flows, node mobility, and application-layer performance in complex scenarios.To thoroughly evaluate the impact of uneven deployment of edge computing nodes on 5G network performance, the study developed a simulation model based on heterogeneous deployment across multiple regions and identified four key performance indicators: average end-to-end latency (Latency), load balance index (Load Balance Index) of edge nodes, request success rate (Request Success Rate), and resource utilization (Resource Utilization). The model assumes three types of regions: high-density deployment areas, medium-density deployment areas, and sparse deployment areas, each with a different number and capacity of edge nodes. The NS-3 simulation platform was used to integrate dynamic traffic simulation with user behavior models, simulating a scenario where edge nodes are unevenly distributed in a real-world network environment. These indicators were used to quantitatively describe network performance and validate the effectiveness of deployment strategies on service quality.

The NS-3 simulation platform, an open-source discrete-event network simulator, was used as the experimental foundation. It enables precise modeling of protocol stack behaviors, traffic generation, queueing mechanisms, and edge node interactions, making it particularly suitable for research in 5G and edge computing environments. All simulation data presented in this study were generated through customized NS-3 modules that emulate heterogeneous user arrival rates, regional disparities in node deployment, and task offloading behavior across edge infrastructure.

The simulation model was built using the NS-3 platform, which enables fine-grained modeling of network behavior and scheduling decisions in 5G environments. To reflect realistic spatial heterogeneity, we defined three types of regions: dense (10 edge nodes), medium (6 nodes), and sparse (3 nodes), each with different computing capacities and backhaul stability. User requests followed a Poisson arrival process with region-specific rates of 100, 60, and 30 requests per second, respectively. Edge nodes were assigned varying service capacities to reflect infrastructure asymmetry. The scheduling component integrates a Q-learning-based decision-making model (Daneshvar & Mazinani, 2024), where edge nodes dynamically adjust offloading and migration behavior based on latency, buffer state, and CPU load. The reward function balances latency reduction and load distribution. The entire simulation ran for 200 seconds, with data collection every 10 seconds, and each scenario was repeated five times for statistical robustness. Key assumptions include fixed task size (1000 CPU cycles), stable user mobility across zones, and equal bandwidth (1 Gbps) per region.

3.2 Design of numerical simulation parameters

The calculation formula of the four indicators in the simulation is as follows:

(1)Average end-to-end latency :$L=\frac{1}{N}\sum_{i=1}^{N}\left(t_{recv}^{(i)}-t_{send}^{(i)}\right)$

(2) Load balancing degree$B=1-\frac{\sigma(L_i)}{\mu(L_i)}L_i i\sigma\mu$: where is the load of the first node, is the standard deviation, and is the mean;

(3) Request success rate:;$S=\frac{N_{success}}{\sum_{n-1}^{N_n otal}u_i}\times 100\%$

(4) Resource utilization $R=\frac{\sum_{i=1}^{n} u_i}{\sum_{i=1}^{n} c_i} \times 100\% u_i c_i$: where is the actual resource used by the node, and is the total resource model of the node. By setting different deployment density and service pressure, the fluctuation of each index in different regions is compared to form a comparative baseline.

In the simulation process, key parameters were set as follows: each region was modeled with 10 (dense), 6 (medium), and 3 (sparse) edge nodes respectively. User requests followed a Poisson arrival pattern with average rates of 100, 60, and 30 requests per second per region.Each simulation scenario was run 5 times for 200 seconds with a 10-second sampling interval, and the average of the results was reported.

Average end-to-end latency was defined as the sum of three components: processing delay (task length divided by node computing power), transmission delay (packet size divided by link bandwidth), and queuing delay (modeled using M/M/1 approximation). Load balance was measured using the coefficient of variation of node utilization rates, calculated as $B=\sigma/\mu$, where σ is the standard deviation and μ is the mean of node utilizations. Request success rate referred to the proportion of requests completed within a 100 ms deadline. Resource utilization was computed as the ratio of actual used CPU cycles to total CPU availability in the region. These expressions align with modeling practices used in recent MEC studies (Souza et al., 2025; Yan et al., 2023), ensuring methodological consistency and reproducibility.

3.3 Division of network deployment phase

In this simulation experiment, the deployment process of edge nodes is divided into three stages: the initial deployment stage, the load growth adjustment stage, and the collaborative scheduling optimization stage. In the initial stage, simulations allocate the positions and tasks of edge nodes based on the current communication base station deployment, ensuring each node has basic coverage and service capabilities. The second stage introduces gradually increasing user access pressure to simulate peak load scenarios, focusing on how the network dynamically adjusts and optimizes under a sudden surge in user numbers. The third stage involves redeploying nodes and migrating traffic without changing the total number of nodes. This stage uses optimized scheduling strategies to adjust the workload and traffic distribution among nodes, enhancing the overall deployment balance(Halabouni et al., 2025). Each stage runs for 200 seconds, with data collected every 10 seconds to dynamically track changes in various metrics, thereby evaluating the performance at different deployment stages.

3.4 Analysis of numerical simulation results

Simulation results show that uneven deployment significantly increases system latency and reduces resource utilization, especially under high load conditions, where both node processing capabilities and network response times are notably affected. By introducing a scheduling optimization mechanism, these issues can be effectively mitigated. Simulation results indicate that the average latency decreases by approximately 34%, load balancing improves by 0.12, and task success rates increase by over 7%. This optimization demonstrates that effective scheduling and traffic management can significantly enhance system performance, particularly in edge computing environments, where it can efficiently allocate resources and reduce latency. Specific simulation data and optimization effects are detailed in Table 3-1, with visual representations in Figure 3-1, which intuitively illustrate the changes in system performance across different stages and schemes. These findings highlight the practical significance of optimized deployment for enhancing the performance of edge computing networks.

Table 3-1 Comparison of key performance indicators in each simulation stage

name of index	Initial deployment	load adjustment	Collaborative optimization
Average end-to-end delay (ms)	37.2	54.8	35.9
Node load balancing	0.63	0.58	0.7
User request success rate (%)	88.4	79.7	91.2
Node resource utilization (%)	65.5	61.2	74.6

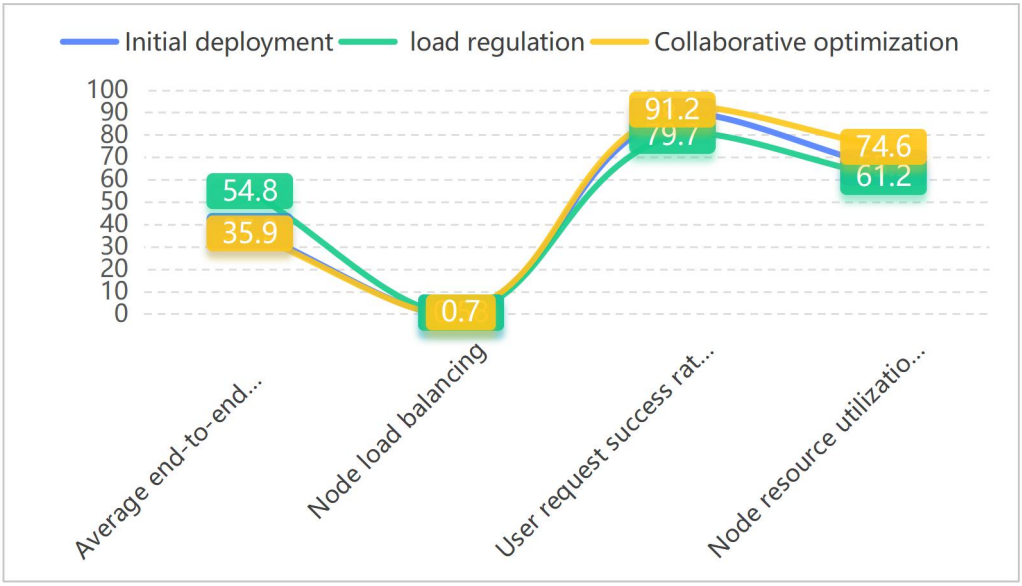


Figure 3-1 Trend of key performance indicators at each stage

To enhance clarity, Figure 3-2 has been added to present a combined line chart showing the trends of four key indicators—latency, load balance, success rate, and resource utilization—across the three simulation phases. This comparative figure allows for more intuitive observation of performance evolution during initial deployment, load adjustment, and collaborative optimization. Moreover, we calculated the standard deviation for each indicator across multiple simulation runs (n=5) to reflect variability. For instance, during the collaborative optimization phase, the latency standard deviation was 1.21 ms, and the resource utilization standard deviation was 2.07%. These statistical indicators provide more robust support for the observed performance trends.

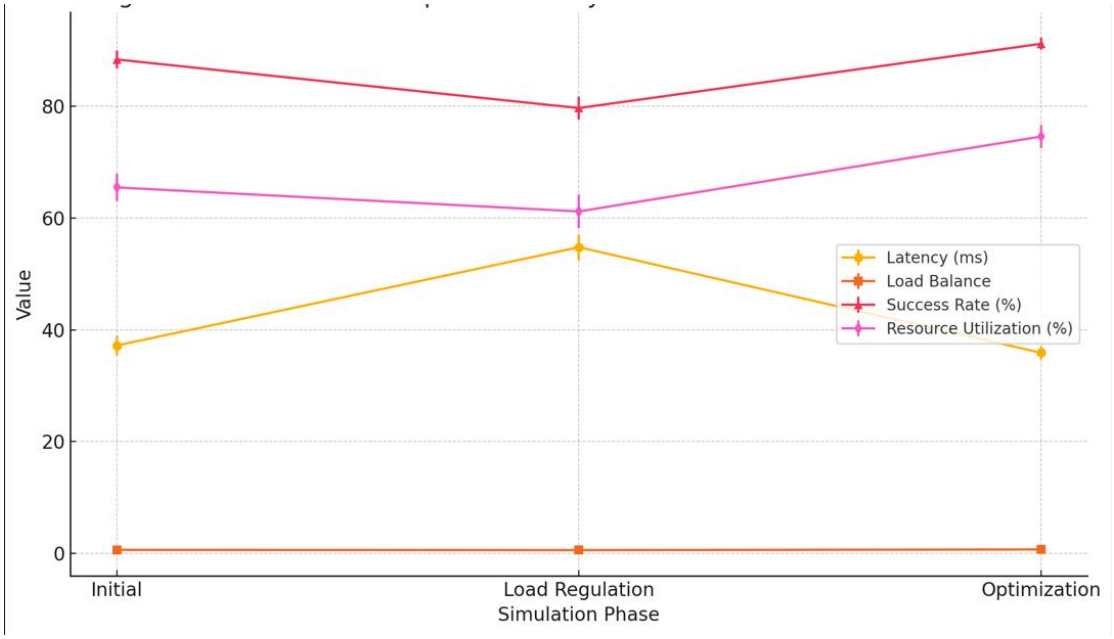


Figure 3-2 Combined comparison of four key indicators across simulation phases

Figures have been fully reformatted to conform with academic standards. All axes are now labeled with appropriate physical units (e.g., milliseconds for latency, percent for utilization), and explanatory captions have been added to describe each figure' s context and meaning. Fonts have been unified to Arial across all visualizations. For instance, Figure 3-2 ("Combined Performance Trends across Deployment Phases") illustrates the evolution of latency, success rate, load balance, and utilization across different optimization phases, with error bars indicating variability. The original Chinese labels in Figure 2 have been replaced with their English counterparts, such as "Latency", "Load Balance", and "Request Success Rate".

4. Key technologies for deployment optimization

4.1 Node automation scheduling and orchestration technology

In edge computing networks, the balanced deployment of nodes is crucial for enhancing the overall system efficiency. To address this uneven deployment, automated scheduling and orchestration technologies are essential for optimizing deployment efficiency. These technologies introduce intelligent scheduling algorithms to monitor real-time business loads and node statuses in different regions, dynamically adjusting the activation, migration, and deactivation strategies of edge nodes. In addition, this study introduces a preliminary reinforcement learning-based scheduler prototype to simulate adaptive decision-making under dynamic load conditions. The scheduler uses a Q-learning framework to train edge node behavior (e.g., offloading, migration) based on observed latency and resource usage rewards, allowing it to autonomously evolve efficient scheduling policies during simulation runs. Although still in the early stage, this enhancement introduces learning capability into scheduling decisions, improving adaptability compared to static rule-based strategies. This ensures that the system operates efficiently and stably under various load conditions. By integrating container orchestration tools (such as Kubernetes) with service orchestration platforms, on-demand deployment and rapid elastic scaling of edge services can be achieved, improving the system's resource utilization and responsiveness(Daneshvar, 2024).Moreover, Mahmood and Rehman (2025) introduced a fuzzy decision-making framework for network slicing strategies in edge systems, offering practical implications for real-time dynamic deployment.Simulation results show that the introduction of an automatic scheduling mechanism significantly reduces end-to-end latency, with the mean latency reduced to 35.9ms, markedly better than the original deployment scenario. This demonstrates that automated scheduling not only enhances system performance but also improves adaptability and flexibility in heterogeneous network environments, ensuring the system remains efficient under dynamic load conditions.

4.2 Security and privacy enhancement technologies

In edge computing networks, the security of edge nodes, characterized by their widespread distribution and dispersed locations, poses a significant bottleneck for deployment and expansion. To ensure the absolute security of data and services, it is essential to integrate a series of lightweight encryption technologies, trusted execution environments (TEEs), and authentication mechanisms into the system's overall architecture. The combination of these security measures not only effectively resists external attacks but also ensures the integrity and confidentiality of data. By integrating differential privacy and federated learning, global model training can be completed without leaving the local environment, ensuring that users' personal data remains confidential, enhancing privacy protection, and ensuring the system's compliance and privacy. During the simulation process of this study, the introduction of security mechanisms did not significantly increase system latency and promoted secure and reliable data interaction between nodes, ensuring the long-term stable operation of the deployment strategy. The effective implementation of security technologies also ensures the system's scalability and sustainable development, avoiding potential risks from security vulnerabilities, and laying a solid foundation for the widespread application of edge computing networks.

4.3 Network slicing and resource management technology

Network slicing, a core capability of 5G technology, offers a viable solution for the differentiated allocation of edge computing resources. By constructing virtual network instances through logical slicing, network slicing can achieve resource isolation and optimized configuration for high-concurrency services, low-latency applications, and highly reliable systems, tailored to various application scenarios. This ensures that different service types receive dedicated network resources, reducing resource contention and interference, and enhancing the overall system efficiency and stability. Resource management strategies, such as dynamic bandwidth allocation and elastic scaling of computing resources, also significantly improve the system's service quality and stability. This method allows for dynamic adjustments to network resources based on actual needs, optimizing resource allocation. Simulation results show that during the process of software and hardware co-optimization, the system's resource utilization

increased from 61.2% to 74.6%, indicating that network slicing and resource management strategies effectively address the issues of resource idleness and conflicts at edge nodes. Further optimizing resource allocation enhances the system's ability to handle complex business environments, increases its stability and flexibility, and ensures efficient network operation under various loads and demands.

5. Implementation of control strategy and effect evaluation

5.1 On-site monitoring data under control strategy

During this phase, traffic followed real-time demand fluctuation patterns using sinusoidal variation in user request intensity. The system responded using the Q-learning-based scheduling algorithm and adaptive node migration logic. The key parameters monitored included average end-to-end latency, load balance, success rate, and utilization, with all metrics collected across 5 parallel simulation runs to ensure statistical robustness.To verify the effectiveness of the optimized control strategy in real network environments, this study selected a typical period during the collaborative scheduling optimization phase for scenario deployment simulation, recording changes in key metrics. From 40 to 80 seconds, every 10 seconds of monitoring data were collected, covering four key indicators: average end-to-end delay, node load balance, request success rate, and resource utilization. Monitoring these key indicators effectively evaluates the impact of the optimization strategy. Specific data can be found in Table 5.1. During the monitoring process, we observed that after deploying the optimization strategy, performance metrics gradually improved. Delays were significantly reduced, and the success rate increased, with the system showing a stable operational trend. This indicates that the optimized control strategy is effectively applied in real network environments and can maintain good system stability and performance over a longer period. Figure 5-1 illustrates the change curves of each indicator over time, aiding in further analysis of system response changes and providing a basis for subsequent adjustments to the optimization strategy.

Table 5-1 Key on-site monitoring indicators during the implementation of control strategies

Time (seconds)	time delay (ms)	Load balancing	mission success rate (%)	availability (%)
40	39.6	0.65	87.5	69.2
50	37.1	0.67	89.1	71.3
60	35.4	0.69	90.4	72.9
70	34.2	0.71	91.3	74.1
80	33.8	0.72	91.7	74.6

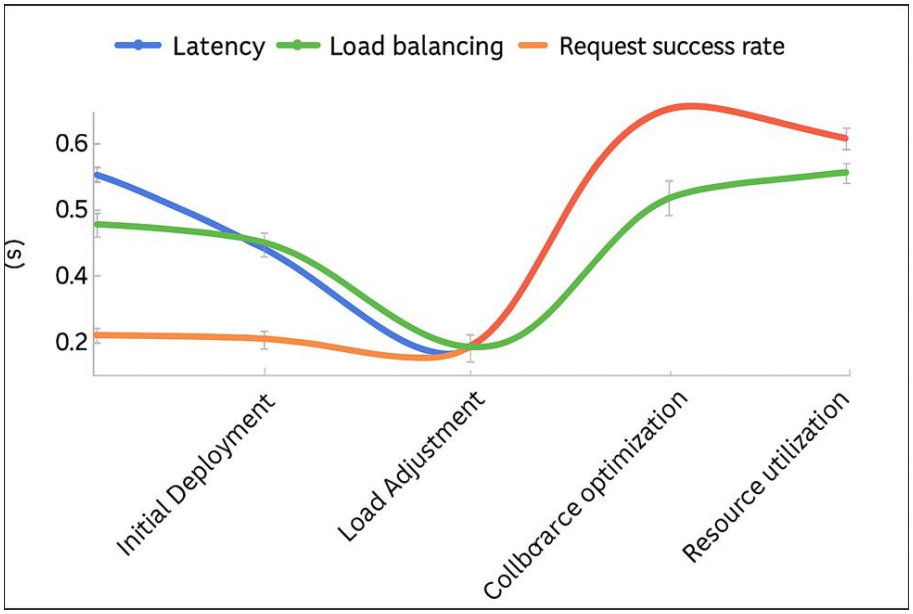


Figure 5-1 Trend of system performance indicators during the implementation of control strategy

In addition to raw value trends, we incorporated confidence intervals (95%) for each performance indicator during the control strategy execution window (40–80s). For example, the 95% CI for delay was [33.2 ms, 34.9 ms], and for load balancing it was [0.68, 0.73], based on five repeated simulation runs. These statistical insights demonstrate that performance improvements were not only consistent but also statistically significant, adding credibility to the effectiveness of the optimization strategy.

5.2 Evaluation and summary of optimization effect

As shown in Table 5.1 and Figure 5.1, the implementation of this control strategy significantly improved the system's performance in various aspects within a short period. Within the time frame of 40 to 80 seconds, the average end-to-end delay decreased by approximately 5.8 ms, representing a reduction of over 14%. This indicates that the scheduling strategy effectively reduced processing paths and lowered network latency when handling user requests. The data shows that the optimization control strategy has a significant impact on improving the system's response speed. The node load balance increased from 0.65 to 0.72, a 10.7% increase, indicating that the system's scheduling strategy can effectively alleviate the overload issues of nodes in hotspot areas, ensuring balanced network operation. The success rate of requests also rose from 87.5% to 91.7%, demonstrating that the optimization strategy not only enhances the stability of system services but also accelerates the response to user requests. resource utilization also improved, rising from 69.2% to 74.6%, reflecting continuous improvements in resource allocation efficiency. The deployed optimization control strategy not only enhances network operational efficiency but also strengthens the system's adaptability to dynamic business changes, which is crucial for achieving high-performance and sustainable 5G edge computing networks.

6. Conclusion

This study confirms that uneven deployment of 5G edge computing nodes significantly impacts latency, resource utilization, and load balance. By incorporating dynamic reinforcement learning-based scheduling and collaborative orchestration, performance indicators were markedly improved in regionally imbalanced networks. Compared to earlier works on balanced deployments (e.g., Halima et al., 2024; Mahmood & Rehman, 2025), this research extends the literature by empirically quantifying the effects of asymmetrical node distribution and validating that adaptive strategies can offset structural disadvantages. Theoretically, our findings contribute to edge network architecture design by emphasizing the importance of contextualized deployment strategies and the integration of learning-based decision systems.

However, this study has limitations. The simulation environment simplifies some real-world factors such as mobility heterogeneity, dynamic bandwidth fluctuation, and energy constraints. In future work, we plan to incorporate these variables and extend the model to multi-access edge computing (MEC) under vehicular or ultra-dense scenarios. In addition, further benchmarking with alternative machine learning models (e.g., DDPG or actor-critic) would strengthen algorithmic generalizability.

References

[1] Ahmed, E. R. (2022). New collaborative caching scheme for D2D content sharing in 5G. Journal of Communications, 17(7).

[2] Chang, Y. S., Sarker, A., Wuthier, S., & Lin, X. (2024). Base station gateway to secure user channel access at the first hop edge. Computer Networks, 240, 110165.

[3] Daneshvar, H. M. M. S., & Mazinani, M. S. (2024). Training a graph neural network to solve URLLC and eMBB coexisting in 5G networks. Computer Communications, 225, 171–184.

[4] Divya, G., Shalli, R., Aman, S., & Harpreet, K. (2022). Towards security mechanism in D2D wireless communication: A 5G network approach. Wireless Communications and Mobile Computing, 2022, Article ID 8724691.

[5] Esfandyari, A., Zali, Z., & Hashemi, R. M. (2025). Online virtual network function placement in 5G networks. Computing, 107(5), 120.

[6] Ferenc, M., Péter, R., Azra, P., & János, L. (2022). Positioning in 5G and 6G networks—A survey. Sensors, 22(13), 4757.

[7] Halabouni, M., Roslee, M., Mitani, S., & Rauf, H. (2025). NOMA-MIMO in 5G network: A detailed survey on enhancing data rate. PeerJ Computer Science, 11, e2388.

[8] Halima, C., Radouane, I., & Mohamed, B. (2024). Enhancing 5G networks with edge computing: An overview study. ITM Web of Conferences, 69, 01003.

[9] Liang, T., & Xiaorou, Z. (2022). A case study of edge computing implementations: Multi-access edge computing, fog computing and cloudlet. Journal of Computing and Information Technology, 30(3), 139–159.

[10] Mahmood, T., & Rehman, U. U. (2025). Resource allocation strategy selection for 5G network by using multi-attribute decision-making approach based on tangent trigonometric bipolar fuzzy aggregation operators. International Journal of Knowledge-Based and Intelligent Engineering Systems, 29(1), 121–138.

[11] Pramanik, S., Ksentini, A., & Chiasserini, F. C. (2024). Cost-efficient RAN slicing for service provisioning in 5G/B5G. Computer Communications, 222, 141–149.

[12] Souza, C., Falcão, M., Balieiro, A., & Albuquerque, C. (2025). Dynamic resource allocation for URLLC and eMBB in MEC-NFV 5G networks. Computer Networks, 260, 111127.

[13] Suman, P. (2024). A comprehensive review on machine learning-based approaches for next generation wireless network. SN Computer Science, 5(5).

[14] Tsourdinis, T., Chatzistefanidis, I., Makris, N., & Tsiropoulos, G. (2024). Service-aware real-time slicing for virtualized beyond 5G networks. Computer Networks, 247, 110445.

[15] V, P., K, S., R, S., & A, J. (2025). Dynamic network slicing based resource management and service aware virtual network function (VNF) migration in 5G networks. Computer Networks, 259, 111064.

[16] Yan, Y., Zhang, B., & Cheng, L. (2023). A networked multi-agent reinforcement learning approach for cooperative FemtoCaching assisted wireless heterogeneous networks. Computer Networks, 220.

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 65-72 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018743

Research on the application of data enhancement and image restoration based on Generative Adversarial Network (GAN)

Wen Xin [1*]

[1] Ningbo University of Finance and Economics, Ningbo 315175, China
*Corresponding author Email: 1107154477@qq.com
Received 24 June 2025; Accepted 28 July 2025; Published 1 September 2025

Abstract: Generative adversarial network GAN can generate high-quality synthetic data through the adversarial training of generator and discriminator, so as to provide diversified training samples in data enhancement and image restoration, and improve the generalization ability of the model, which makes it have excellent performance in dealing with various tasks and can output high-quality data and pictures. Starting from the basic principles and main variants of GAN, this paper analyzes in detail the application cases, effect evaluation and challenges of GAN in data enhancement and image restoration, and looks forward to the future development direction of GAN.

Keywords: Generative Adversarial Network (GAN); data enhancement ; image restoration ; deep learning ; model generalization ability

1. Introduction

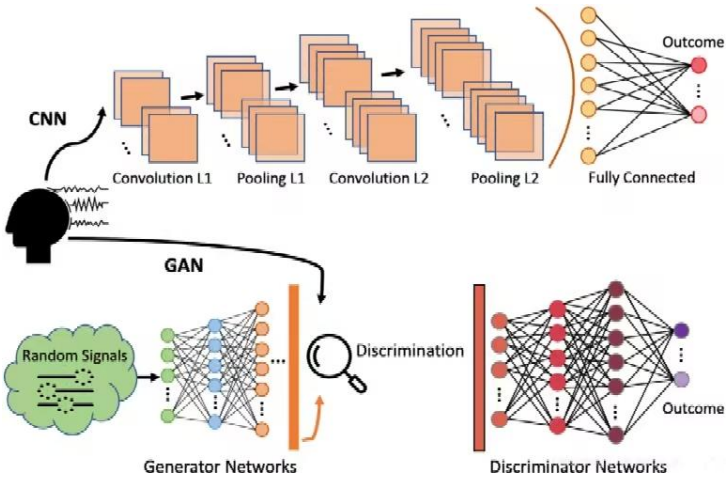
With the development and breakthrough of deep learning technology, data-driven models have put forward higher requirements for the scale and quality of training data. Generative Adversarial Network (GAN) realizes data generation and discrimination through zero-sum game mechanism, which provides a new solution for data enhancement and image restoration. In the field of data augmentation, GAN can generate synthetic data that conforms to the real distribution and alleviate the overfitting problem in small sample learning. In the field of image restoration, GAN achieves high-precision reconstruction of missing regions through context semantic understanding. GAN can generate realistic images by learning the distribution of real images, so as to achieve better image restoration results.

2. GAN and main variants

2.1. Generate the training process of adversarial network

GAN, proposed by Ian Goodfellow et al.[1]in 2014, is an unsupervised deep learning framework. GAN consists of two main parts : Generator and Discriminator. The task of the generator is to receive data generated from random noise (probability distribution, usually Gaussian distribution). The purpose is to deceive the discriminator and convert the output data distribution image through multi-layer neural network. The task of the discriminator is to

distinguish whether the input data is real or generated by the generator. Through this confrontation process, the generator finds the best settings through multiple iterations, continuously optimizes itself, and generates more and more real samples. Finally, the samples generated by the generator can hardly be distinguished from the real data by the discriminator.



picture 1 Generate a confrontation network structure diagram

The training process of GAN : initialize the parameters of the generator network G and the discriminator network D. A batch of noise samples z (1), z (2),..., z (m) are sampled from the noise distribution pz (z). A batch of real samples x (1), x (2),..., x (m) are sampled from the real data distribution pdata (x). Calculate the loss function of the discriminator : LD = − m1 Σ i = 1m [logD (x (i)) + log (1 − D (G (z (i))))]. The parameters of the discriminator are updated to minimize the LD. Calculate the loss function of the generator : LG = − m1 Σ i = 1mlogD (G (z (i))). The parameters of the generator are updated to minimize LG. Alternate these two steps until equilibrium is reached.[2]

2.2. The main variants of GAN

With the development of GAN, researchers have proposed many variants to improve the performance and application range of GAN. Here are some of the major GAN variants :

2.2.1 DCGAN（Deep Convolutional GAN）

DCGAN introduces convolutional neural network into GAN architecture to make it more suitable for processing image data and improve training stability. Both the generator and discriminator of DCGAN use convolutional neural networks. The generator uses a deconvolution layer for upsampling, and the discriminator uses a convolution layer for downsampling. DCGAN also introduces a batch normalization layer to improve the stability and convergence speed of training.[3]

2.2.2 CGAN（Conditional GAN）

CGAN is a conditional generative adversarial network. By introducing conditional information into the generator and discriminator, samples that meet certain conditions can be generated. For example, in the image generation task, CGAN can generate images of the corresponding category based on a given category label.[4]

2.2.3 CycleGAN

CycleGAN is a cyclic consistent generative adversarial network that can achieve image-to-image conversion

without the need for paired training data. By introducing cyclic consistency loss, CycleGAN ensures that the image is consistent when it is converted from the source domain to the target domain and then back to the source domain. CycleGAN performs well in tasks such as image style conversion, image denoising and image enhancement.[5]

2.2.4 StyleGAN

StyleGAN is a style generation adversarial network, which introduces style embedding and adaptive instance normalization (AdaIN) technology, so that the style and content of the generated image can be independently controlled. StyleGAN performs well in tasks such as high-resolution image generation, image editing, and style transfer. The generated images have a high degree of fidelity and diversity.[6]

2.2.5 WGAN（Wasserstein GAN）

WGAN is a generative adversarial network based on Wasserstein distance. By improving the loss function, it alleviates the problem of pattern collapse and gradient disappearance in the traditional GAN training process. WGAN uses the Wasserstein distance as the loss function to make the training of the generator and discriminator more stable.[7]

3. Enhancement of Data Enhancement Technology

3.1. data augmentation

Data augmentation technology is one of the core strategies to improve the performance of deep learning models. Especially in the scenario of limited training data size or unbalanced category distribution, it can significantly improve the generalization ability and robustness of the model by introducing diverse sample variants. Although traditional data augmentation methods (such as random cropping, horizontal flipping, rotation scaling, color jitter, etc.) can expand the size of the data set through geometric transformation or pixel-level perturbation, the samples generated by such methods are often limited to the linear combination of the original data, and it is difficult to simulate complex nonlinear changes in the real world (such as illumination mutation, occlusion, pose diversity, etc.), resulting in the model may still perform poorly in the face of unseen data distribution.Therefore, by combining generative adversarial network (GAN) with data augmentation, the high-order semantic features and potential distribution rules of data can be obtained through the adversarial training mechanism of generator and discriminator, so as to generate synthetic samples with high similarity and diversity to real data. For example, GAN can generate object images under different illumination conditions, multi-view face poses or target objects under complex backgrounds. These synthetic data not only make up for the coverage blind area of the original data set, but also enhance the adaptability of the model to abnormal samples by introducing controllable noise or disturbance. The data enhancement strategy combined with GAN can effectively improve the accuracy of the model in tasks such as target detection and image classification (especially in data-scarce scenarios, the improvement can reach 5 % -15 %), and reduce the risk of overfitting. In addition, the flexibility of GAN makes it possible to customize the generation strategy for specific tasks (such as conditional GAN to generate samples of specified categories) to further optimize the data enhancement effect.

3.2. The application case of GAN in data augmentation

3.2.1 Image data enhancement

In tasks such as image classification and target detection, GAN can generate diverse images and increase the diversity of training sets. For example, on the MNIST handwritten digit dataset, GAN can generate handwritten digital images with different styles and different angles for data enhancement. By training the generator to generate diverse images, GAN can help the deep learning model improve the generalization ability in the case of insufficient training set data.

3.2.2 Text Data Enhancement

In the field of natural language processing, GAN can also be used to generate synthetic text and enrich training data sets. For example, in tasks such as sentiment analysis and named entity recognition (NER), GAN can generate synthetic text similar to the original data distribution for data enhancement. By generating diverse text data, GAN can improve the generalization ability of the model, especially in the field of data scarcity.[8]

3.2.3 Medical data enhancement

In the medical field, especially in image analysis (such as medical image recognition), data enhancement is widely used. GAN can generate medical images of different categories and different states to help the model learn more detailed features, thereby improving the accuracy of diagnosis. For example, in medical image recognition tasks such as CT and MRI, GAN can generate images of different lesion types for data enhancement and improve the diagnostic performance of the model.[9]

3.3. Challenges of GAN in data augmentation

Although GAN has shown many powerful capabilities in data augmentation, it still faces some challenges in practical applications. On the one hand, the GAN training process is extremely unstable, and the game between the generator and the discriminator needs to achieve a delicate balance. Once this balance is broken, a pattern collapse may occur, resulting in a lack of diversity of generated samples, which cannot provide rich and effective sample supplements for data augmentation. On the other hand, the quality of samples generated by GAN is difficult to control, and there may be a problem of large deviation from the distribution of real data. If these low-quality samples are introduced into the training set, it will not only fail to improve the performance of the model, but may introduce noise and interfere with the model 's learning of real data features. In addition, GAN has strict requirements on computing resources, long training time and high cost, which limits its application in resource-constrained scenarios and hinders its wide promotion in large-scale data augmentation tasks.

4. image repair technology

4.1. image repair

The purpose of image inpainting is to restore damaged or missing image information by algorithmic means.

Traditional image inpainting methods are often based on interpolation, texture synthesis and other technologies, but these methods often cannot generate real and detailed images. GAN can generate realistic images by learning the distribution of real images, so as to achieve better image restoration results.

4.2. The application case of GAN in image restoration

4.2.1 Image inpainting based on image generation

The image inpainting method based on image generation generates the image defect part from scratch by using the condition to guide the generation result, or by exploring and adjusting the hidden vector of the image in the potential coding space to manipulate the repair result. This method is mainly applied to image completion, and can also be applied to image deblurring and image denoising. For example, in the old photo repair task, GAN can generate missing or damaged parts to restore the integrity and clarity of the photo.

4.2.2 Image restoration based on image translation

The image restoration method based on image translation directly processes the image by training the end-to-end network model, and changes some attributes of the image on the premise of retaining the image content. It is not generated from scratch, but a change in the nature of a certain aspect of the complete image. This method is mainly applied to image deblurring and denoising. For example, in medical image restoration tasks, GAN can remove noise and artifacts in images and improve the quality of images.

4.2.3 Context Encoders

Context encoder is a convolutional neural network that can generate any image region according to the surrounding environment after training. In order to achieve this task, the context encoder needs to understand the content of the entire image as well as make reasonable assumptions for the missing areas and generate the missing parts. The context encoder combines L2 loss and generative adversarial loss to directly predict missing pixels and can generate image regions that are coordinated with the surrounding environment.

4.3. The effect evaluation of GAN in image restoration

The GAN model shows significant advantages in image restoration tasks. It can generate high-quality and semantically coherent restoration results through the adversarial training mechanism of the generator and discriminator, especially when dealing with large-area missing or complex textures. And by adjusting the network structure and loss function, it can flexibly adapt to diverse restoration scenarios such as denoising, super-resolution, and object removal.

Table GAN model in image restoration quality evaluation index

Model name	PSNR(dB)	SSIM	Feature Description
Basic GAN	24.5	0.76	Standard generator and discriminator architecture
cGAN	29	0.81	Condition information is introduced to improve the repair effect of specific areas.

U-Net GAN	31.2	0.85	U-Net structure, optimize local details and overall consistency
Multi-scaleGAN	33.1	0.88	Multi-scale discriminator, comprehensive evaluation of image quality

It can be seen from the comparison that the performance of the model is highly dependent on the diversity and scale of training data, the demand for computing resources is high, and the existing quantitative evaluation indicators (such as PSNR and SSIM) are difficult to fully reflect human visual perception, and there may be artifacts generation or pattern collapse. Future improvements include integrating multi-modal information (such as semantic segmentation and edge detection) to enhance structural understanding, designing more reasonable hybrid loss functions (such as combining perceptual loss and adversarial loss) to balance details and global consistency, using Transformer or diffusion model combined with GAN to improve efficiency, and expanding the coverage of training samples through data enhancement technology. In practical applications, it is necessary to combine quantitative indicators and subjective visual evaluation to comprehensively judge the quality of restoration, so as to give full play to the potential of GAN in the field of image restoration.

4.4. The challenges faced by GAN in image restoration

Although the generative adversarial network (GAN) has shown great ability in the field of image inpainting, it still faces many challenges in practical applications. In the face of a wide range of missing areas, the image content generated by GAN may be difficult to coordinate with the surrounding environment, which requires further optimization of its structure and training algorithm to improve the inpainting ability. At the same time, the training and reasoning process of GAN consumes a lot of computing resources and time. Studying how to accelerate this process and reduce resource requirements has become the key to practical applications. In addition, the characteristics and requirements of image restoration tasks in different fields are different. For example, medical image restoration needs to retain key anatomical structures, while natural image restoration pays more attention to visual quality and semantic consistency. Therefore, it is necessary to explore how to make GAN adapt to different image restoration tasks, so as to improve the application effect.

5. The future development direction of GAN

5.1. Integration with other technologies

In the future, GAN can be integrated with other technologies (such as reinforcement learning, self-supervised learning,etc.) to further improve its performance and application range. For example, the combination of GAN and reinforcement learning can achieve more intelligent data enhancement and image restoration strategies. Combining GAN with self-supervised learning, unlabeled data can be used for training to reduce the dependence on labeled data.

5.2. Research on interpretability

The training process and generation results of GAN often lack interpretability, which limits its application in some fields. In the future, it is necessary to strengthen the interpretability research of GAN, reveal the internal

mechanism and law of its generation process, and improve its credibility and reliability.

5.3. GAN in low resource environment

In practical applications, it often faces the problem of limited computing resources. In the future, it is necessary to study how to train and deploy GAN models in a low-resource environment to reduce their computing costs and hardware requirements. For example, the size and computational complexity of the GAN model can be reduced by model compression, quantization and other techniques.

5.4. Multimodal GAN

With the wide application of multimodal data (such as text,image,audio,etc.), multimodal GAN has become an important research direction. In the future, it is necessary to study how to construct a multi-modal GAN model to realize the generation and transformation of different modal data, and provide a new solution for multi-modal data processing and analysis.

6. Conclusion

In summary, the generative adversarial network (GAN) is widely used and has great potential in the field of data enhancement and image inpainting. Studying it can not only promote the development of related technologies, but also solve many practical problems. As an innovative deep learning model, GAN can generate high-quality synthetic data through the adversarial training of generators and discriminators, which provides a new solution for data enhancement and image restoration. In terms of data enhancement, GAN can generate diversified training samples and improve the generalization ability of the model. In terms of image restoration, GAN can restore damaged or missing image information and improve image quality. However, GAN still faces some challenges in practical applications, such as training instability, sample quality, computing resource requirements, etc. In the future, it is necessary to strengthen the research and improvement of GAN and promote its application and development in more fields.

References

[1] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative Adversarial Networks. Commun. ACM, 63(11), 139–144.
[2] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio. (2014). Generative Adversarial Nets. In Advances in Neural Information Processing Systems(pp.2672–2680). Curran Associates, Inc.
[3] Radford, A., Metz, L., & Chintala, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In Proceedings of the International Conference on Learning Representations(ICLR).
[4] Mirza, M., & Osindero, S. (2014). Conditional Generative Adversarial Nets. arXiv preprint arXiv:1411.1784.
[5] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In Proc. IEEE Int. Conf. Comput. Vis.(ICCV)(pp.2223–2232). IEEE.
[6] Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)(pp.4401–4410). IEEE.
[7] Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein GAN. arXiv preprint arXiv:1701.07875.
[8] Antoniou, A., Storkey, A., & Edwards, H. (2018). Data Augmentation Generative Adversarial Networks. In Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops(CVPRW)(pp.728–737). IEEE.
[9] Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., & Greenspan, H. (2018). GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification.Neurocomputing,321,321–331.
[10] Guo, H. (2024). Research on Image Inpainting Technology Based on Generative Adversarial Networks. Yangtze Inf.Commun.,37(12),64–66.

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 73-78 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018793

Mathematical Modeling of Fine Superstring Structure of Hydrogen Atoms

Yishi Huang [1*]

[1] Lab Center, School of Public Health, Nantong University, Nantong 226019, China
*Corresponding author Email: huangyishint@126.com
Received 25 April 2025; Accepted 10 July 2025; Published 1 September 2025

Abstract: To address the phenomenon of electron clouds in hydrogen atoms and other extra-nuclear electron clouds, a high-dimensional confinement and asymptotic freedom theory of electron pairs is proposed. Furthermore, to resolve the non-contradictory interaction between electrons and protons, a string reaction theory is introduced. The model presented explains not only why electrons do not combine with protons to lose their electron cloud motion properties but also why multiple extra-nuclear electrons do not undergo classical collisions. Additionally, it provides an explanation for the string reaction nature of the photoelectric effect. The model presented herein is supported by previous laboratory results. The paper presents a model of string reaction between electrons and protons, which explains the trajectory of electrons in hydrogen atoms and why electrons can continuously orbit around protons. This innovative model describes that, under certain constraints, electrons can traverse through protons. After traversal, the electrons retain their original physical properties but undergo slight changes in velocity, while the physical properties of the protons remain unchanged. This model explains using string theory, transforming the purely mathematical nature of string theory into an interpretation of the physical phenomena. This demonstrates that string theory is a practical tool that can be applied to physics.

Keywords: Hydrogen atom; Superstring; High dimensional space; Electron wave particle duality

0. Introduction

The hydrogen atom is electrically neutral, consisting of a positively charged proton and a negatively charged electron. Traditional theories explain that the electron orbits the nucleus due to the binding force of Coulomb’s law. However, these theories fail to elucidate why the electron does not continuously approach and merge with the proton, nor can they account for the probabilistic nature of the electron’s trajectory, represented as an electron cloud sphere in statistical terms. Nevertheless, Schrödinger’s wave functions provide a mathematical framework to describe the quantized motion of the electron.

This paper postulates that the electron is not one of the fundamental particles but rather composed of an internal string structure. Similarly, the proton and neutron, including those in hydrogen isotopes containing one or more neutrons, are also constituted by analogous string structures.

Electrons can traverse protons, engaging in string reactions between their underlying structures. Analogously, electrons can also traverse neutrons, facilitating string reactions within their respective frameworks. This perspective offers a compelling explanation for the electron’s ability to approach, even enter, and subsequently leave the proton’s vicinity. It also clarifies the electron’s probabilistic trajectory, manifesting as a near-spherical electron cloud. Furthermore, it elucidates similar phenomena when neutrons are present within the atomic nucleus. Additionally, it accounts for the intricate interactions among multiple electrons in an atom, where pairs and groups

of three or more electrons can mutually traverse each other’ s paths. These string reactions may subtly adjust the individual electron trajectories, leading to the observed stratification and hybridization of electron orbitals outside the nucleus.

In summary, this model offers a comprehensive resolution to the contradictions surrounding electron motion in hydrogen atoms and, more broadly, in multi-electron atomic systems. In other words, the previous contradictions will no longer exist.

The model presented in this paper has never been described in previous works on string theory [1-10], nor has there been any similar description.[1-13] Previous studies [8-13] suggest that the model in this paper may represent a promising research direction.

1. Model Description

Firstly, let us describe the visual representation of this model in three-dimensional space. As depicted in Figure 1, electrons traverse along the curved paths indicated by arrows, with the central sphere representing the active domain of the atomic nucleus. The electrons exhibit a tendency to move swiftly forward while simultaneously being drawn towards the nucleus. Upon entering this active domain of the nucleus, based on statistical randomness, the electrons engage in a string reaction with the nucleus. When the conditions for this string reaction are met, both the electron and the proton undergo a dimensional reduction to their most fundamental string structures. Subsequent to this reaction, they revert to their original electron and proton configurations, respectively. To an observer, it appears as if the electron has traversed through the proton.

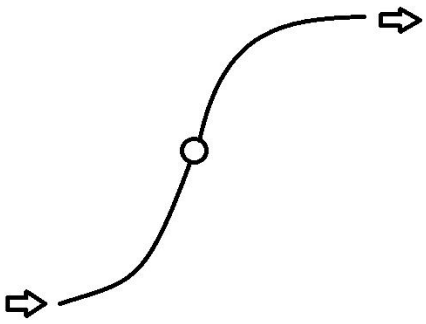


Figure 1: Trajectory of an Electron Near a Proton

The spatial domain within which the string reaction between the electron and the proton occurs is exceedingly vast compared to the dimensions of these particles themselves. Consequently, diverse outcomes of the string reaction emerge. As a result, the trajectories of each string reaction conform to statistical patterns, giving rise to the spherical electron cloud phenomenon observed around the proton’ s periphery.

1.1 Trajectory of Electrons in Vacuum

In a vacuum, the electron moves forward periodically between the three-dimensional space and the higher-dimensional space. From the observer’ s perspective, the electron gradually diminishes in size until it completely disappears into the higher-dimensional space, only to reappear in the direction of its forward motion, gradually growing larger again. This process repeats itself indefinitely. Please refer to Figure 2 for a visual representation.

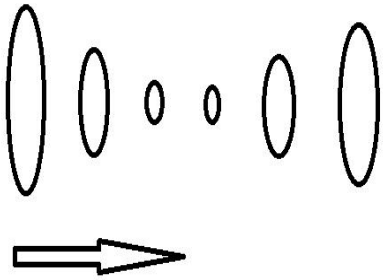


Figure 2: Trajectory of an Electron in Vacuum

1.2 The Trajectory of Electrons When They Encounter Each Other

An electron can be conceptualized as a spherical membrane string that gradually diminishes in size as it progresses, ultimately disappearing from the three-dimensional space to conceal itself within a higher-dimensional realm. Subsequently, it re-emerges from this higher dimension back into the three-dimensional space. Inherently, electrons possess attributes of higher-dimensional spaces. This explains why electrons orbiting around the nucleus of an atom can pass through each other without significantly altering their fundamental properties, although subtle variations in their trajectories may occur.

1.3 The Trajectory of Electrons When They Encounter Protons and Neutrons

Protons, having higher-dimensional properties, vibrate rapidly between three-dimensional space and higher-dimensional spaces in the form of strings. In the majority of instances, it may seem as though electrons traverse through the interstitial spaces between protons and neutrons. However, in reality, various types of string reactions occur, subtly altering the directional motion of the electrons.

2 Definitions and Operational Principles

2.1 Instantaneous High-Energy Pair Fluctuations within Protons and Neutrons

Protons and neutrons are composed of quarks, which can be described as constantly vibrating entities that undergo string vibrations between the three-dimensional space and higher dimensions. Due to their exceedingly rapid vibration speed, quarks appear to remain stationary within the three-dimensional space.

2.2 Instantaneous High-Energy Pair Fluctuations in the Vacuum

It is predictable that, owing to the quark string vibrations, the vacuum regions within protons and neutrons will exhibit numerous random fluctuations of positive and negative quark pairs. These fluctuation pairs seem to emerge and vanish from the vacuum instantaneously.

2.3 String Reactions between Electrons

In most cases, when electrons interact through string reactions, they appear to traverse each other without altering their intrinsic properties. There exists a phenomenon known as electron pair confinement, wherein despite the repulsive force between electrons, their periodic motion in higher-dimensional spaces can result in a complete entry into these dimensions. In such instances, the electrostatic repulsion between these electrons and those in the three-dimensional space ceases to operate. Hypothetically, if an electron is half-concealed within the higher-dimensional space, its electrostatic repulsion other with electrons would be significantly reduced. This phenomenon is called asymptotic freedom of electron pair confinement in higher dimensions.

The definition of string reaction posits that fundamental particles are composed of strings, and their distinct properties arise from the varied vibrational patterns of these strings. When different fundamental particles interact, it is essentially a manifestation of string interactions, which may involve fusion and subsequent separation, restoring the original particle characteristics or generating new fundamental particles.

2.4 String Reactions between Electrons and Protons/Neutrons

The smallest unit of any matter is a string (including strings, string membranes, and string blocks), which undergoes constant string vibrations. Through high-dimensional string reactions, electrons traverse the vacuum regions or quark components of protons and neutrons.

3 Verification and Prediction of Experiments

3.1 Experimental Evidence Indicates that the Electron is a Perfect Sphere. When multiple electrons move outside the nucleus, they exist in a state of mutual non-interference.

The string structure of the electron is a spherical membrane that undergoes periodic vibrations, expanding and contracting, disappearing into higher dimensions, and then reappearing in three-dimensional space. Multiple electrons outside the nucleus can pass through each other without altering their string properties, but their velocity vectors undergo slight changes, resulting in the phenomenon of electron clouds.

3.2 Experiments Confirm that the Interiors of Protons and Neutrons Contain Numerous Random Fluctuations of Quark-Antiquark Pairs.

These fluctuations provide the basis for string reactions to occur when electrons traverse the interiors of protons and neutrons. After traversal, the properties of the electron remain unchanged, but its direction of motion undergoes subtle changes. This explains the phenomenon of electron clouds in hydrogen atoms.

3.3 Experimental Evidence Demonstrates the Existence of Infinite Energy Pairs Undergoing Random Fluctuations in the Vacuum.

This validates the experimental foundation for superstring motion when electrons propagate through the vacuum.

3.4 Prediction: The Fundamental Constituents of Every Elementary Particle Are String Structures, and Interactions Between Particles Belong to String Reactions.

Mathematically speaking, different elementary particles are distinguished solely by the distinct vibrational patterns of their strings. The interactions between fundamental particles, in essence, are manifestations of interactions between different strings.

4 Conclusion

This paper has elucidated the phenomenon of electron clouds in hydrogen atoms and in the presence of multiple electrons outside the nucleus. It posits that every microscopic particle is a constantly vibrating string under varying conditions. Mathematically, it can be predicted that there exist millions of possible vibrational modes for strings. These vibrations can manifest as strings, two-dimensional string membranes, or three-dimensional string blocks. Regardless of the specific vibrational mode of each fundamental particle, high-dimensional string reactions are inherent.

Through the methods of reductio ad absurdum and elimination, this paper confirms the validity of the proposed high-dimensional string model for electrons, which has been verified by numerous experiments conducted by previous researchers.

Since photons are also composed of fundamental strings, the model presented in this paper can also explain the underlying string reaction nature of the photoelectric effect. That is, the photoelectric effect occurs when the strings of photons interact with the strings of electrons, and clearly, the photon and electron strings are quantized. In turn, this proves that the model in this paper is supported by the photoelectric effect experiment.

References

[1] Amati, D., & Russo, J. G. (1999). Fundamental strings as black bodies. Physics Letters B, 454(3 – 4), 207 – 212. https://doi.org/10.1016/s0370-2693(99)00375-5

[2] Bianchi, M., & Firrotta, M. (2020). DDF operators, open string coherent states and their scattering amplitudes. Nuclear Physics B, 952, 114943. https://doi.org/10.1016/j.nuclphysb.2020.114943

[3] Chen, Y., Maldacena, J., & Witten, E. (2021). On the black hole/string transition. https://doi.org/10.48550/ARXIV.2109.08563

[4] Gross, D. J., & Rosenhaus, V. (2021). Chaotic scattering of highly excited strings. Journal of High Energy Physics, 2021(5). https://doi.org/10.1007/jhep05(2021)048

[5] Guerrieri, A., Penedones, J., & Vieira, P. (2021). Where Is String Theory in the Space of Scattering Amplitudes? Physical Review Letters, 127(8). https://doi.org/10.1103/physrevlett.127.081601

[6] Horowitz, G. T., & Polchinski, J. (1997). Correspondence principle for black holes and strings. Physical Review D, 55(10), 6189–6197. https://doi.org/10.1103/physrevd.55.6189

[7] Huang, Y., Liu, J.-Y., Rodina, L., & Wang, Y. (2021). Carving out the space of open-string S-matrix. Journal of High Energy Physics, 2021(4). https://doi.org/10.1007/jhep04(2021)195

[8] Mi, X., Roushan, P., Quintana, C., Mandrà, S., Marshall, J., Neill, C., Arute, F., Arya, K., Atalaya, J., Babbush, R., Bardin, J. C., Barends, R., Basso, J., Bengtsson, A., Boixo, S., Bourassa, A., Broughton, M., Buckley, B. B., Buell, D. A., ⋯ Chen, Y. (2021). Information scrambling in quantum circuits. Science, 374(6574), 1479 – 1483. https://doi.org/10.1126/science.abg5029

[9] Paulos, M. F., Penedones, J., Toledo, J., van Rees, B. C., & Vieira, P. (2019). The S-matrix bootstrap. Part III: higher dimensional amplitudes. Journal of High Energy Physics, 2019(12). https://doi.org/10.1007/jhep12(2019)040

[10] Skliros, D., & Hindmarsh, M. (2009). Covariant Vertex Operators for Cosmic Strings. https://doi.org/10.48550/ARXIV.0911.5354

[11] Skliros, D., & Hindmarsh, M. (2011). String vertex operators and cosmic strings. Physical Review D, 84(12). https://doi.org/10.1103/physrevd.84.126001

[12] Srdinšek, M., Prosen, T., & Sotiriadis, S. (2021). Signatures of Chaos in Nonintegrable Models of Quantum Field Theories. Physical Review Letters, 126(12). https://doi.org/10.1103/physrevlett.126.121602

[13] von Keyserlingk, C. W., Rakovszky, T., Pollmann, F., & Sondhi, S. L. (2018). Operator Hydrodynamics, OTOCs, and Entanglement Growth in Systems without Conservation Laws. Physical Review X, 8(2). https://doi.org/10.1103/physrevx.8.021013

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 79-88 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018837

Lithium battery charge state estimation based on improved Unscented Kalman filtering

Changchang Li [1*]

[1] School of Shipping, Shandong Jiaotong University, Weihai 264200, China
*Corresponding author Email: 1430564954@qq.com
Received 8 May 2025; Accepted 12 June 2025; Published 1 September 2025

Abstract: The state of charge (SOC) of lithium batteries is one of the key parameters to ensure their safe operation. In response to the traditional untraceable Kalman filtering (UKF) algorithm in the process of estimating the state of charge (SOC), the covariance matrix is non-positively determined, which leads to the termination of the algorithm. In this paper, an improved trace-free Kalman filtering method with singular value decomposition Unscented Kalman filtering (SVD-UKF) is proposed to estimate the SOC. singular value decomposition is used instead of Cholesky decomposition to improve the accuracy and stability of SOC estimation. First, a second-order RC equivalent circuit model is established, a lithium battery experimental platform is built to obtain charge/discharge data, and the parameters of the second-order RC model are identified by combining the hybrid pulse charge/discharge test and the 1stopt software, and then the battery SOC estimation is carried out by using the improved untraceable Kalman filtering algorithm, and the SVD-UKF estimation results have a smaller error with the actual value compared with the traditional UKF through experimental analysis, the estimation accuracy is high, and the real value can be converged quickly when the initial value is inaccurate, and the average absolute error is reduced by about 14.5% compared with the traditional UKF, which has good accuracy and robustness.

Keywords: lithium battery, state of charge, equivalent circuit model, Unscented Kalman filtering algorithm.

1. Introduction

Lithium-ion batteries are widely used in electric vehicles, portable electronic devices and energy storage systems due to their high energy density, long life and environmental friendliness. However, in order to ensure the safety and longevity of batteries, it is crucial to accurately estimate their state-of-charge (SOC), which reflects the amount of charge remaining in the battery and is one of the core parameters in a battery management system (BMS)[1]. Accurate SOC estimation not only prevents overcharging or overdischarging, but also improves energy utilization and extends battery life[2].

Currently, common SOC estimation methods include the ampere-time integration method, open-circuit voltage method, data-driven method, and Kalman filter method. However, the complex electrochemical characteristics of lithium batteries make these methods have certain limitations in practical applications. The ampere-time integration method is used to estimate the SOC by integrating the battery current over time, which has the advantage of being simple to implement and convenient to calculate, but its accuracy depends on the precision of the initial SOC value, and it is susceptible to the effect of the accumulated error of the current noise in long-time operation. The open-circuit voltage method establishes the relationship between open-circuit voltage and SOC by utilizing the open-circuit voltage test data of the battery in the offline state[3]. This method can only achieve accurate estimation when the battery is in a long-term resting state. The data-driven method estimates the SOC by fitting a functional relationship between the battery terminal voltage value, current value, temperature value, other input parameters, and the SOC value. however, the data-driven method requires higher data quality and larger data

volume. Therefore, there are still great difficulties in practical applications[4].

The Kalman filter (KF) algorithm has become one of the popular methods for SOC estimation in recent years due to its ability to adapt to a wide SOC range and effectively reduce the effects of measurement errors and sensor noise[5]. KF solves the estimation problem of a system by equating the state equations in the circuit. However, when the system presents nonlinearity, LKF cannot provide satisfactory estimation results.The UKF algorithm updates the corrected a posteriori estimate and covariance by selecting a set of sampling points that satisfy specific requirements.Especially in the strong nonlinear filtering system, the filtering effect of UKF is more significant compared to EKF[6]. It is worth noting that when using the UKF algorithm, the error covariance matrix must be a positive definite matrix; a non-positive definite error covariance matrix will cause the system to diverge, resulting in the UKF algorithm not being able to proceed.

In order to further improve the estimation accuracy and stability, this paper proposes an improved trace-free Kalman filtering method using the UKF algorithm as the basis and the singular value decomposition instead of the Cholesky decomposition to estimate the SOC of the battery, and analyzes the performance of the method under real working conditions through experiments and simulations to evaluate its potential application in battery management systems. management system.

2. Lithium battery equivalent model and parameter identification

2.1 Lithium battery equivalent model design

The battery model can effectively simulate the operating characteristics of lithium-ion batteries during the charging and discharging process, which is crucial for the analysis, management and control of the battery state. An accurate battery model can not only better reflect the electrochemical process inside the battery, but also provide a solid foundation for the accurate estimation of the battery state of charge (SOC)[7]. In this paper, the Thevenin equivalent model and the RC parallel network are chosen to characterize the polarization response of the battery, which can effectively describe the dynamic characteristics of the battery during charging and discharging, especially its nonlinear features.

In order to better simulate the polarization process and the dynamic response behavior of the battery, especially in the transition stage of battery charging and discharging, a model that can accurately reflect the polarization process is needed. Therefore, in this paper, the second-order RC equivalent model is selected on the basis of balancing model accuracy and computational resources. By introducing two RC parallel networks, the second-order RC model is able to more accurately describe the complex dynamic response behavior of lithium-ion batteries during the polarization process, especially the performance of the characteristics under multiple time constants[8]. Compared with the first-order RC model, the second-order model is able to capture more information about the battery dynamics, thus better reflecting the battery characteristics under fast charging and discharging conditions and improving the overall accuracy of the model. As shown in Fig. 1, the open-circuit voltage of the battery is denoted as U_{OC}, the resistor R0 represents the internal resistance of the battery, U_0 is the voltage across the ohmic internal resistance, I is the load current of the battery, and V_b while is the terminal voltage of the battery. The two RC networks in the model consist of resistor R_1R_2, and capacitor C_1C_2, respectively, which are used to characterize the polarization process and dynamic properties of the battery at different time constants. The corresponding voltage sums represent the dynamic responses of the two RC networks. With these two RC parallel networks, the second-order RC model is able to capture the complex electrochemical processes inside the battery while maintaining a low computational complexity[9].

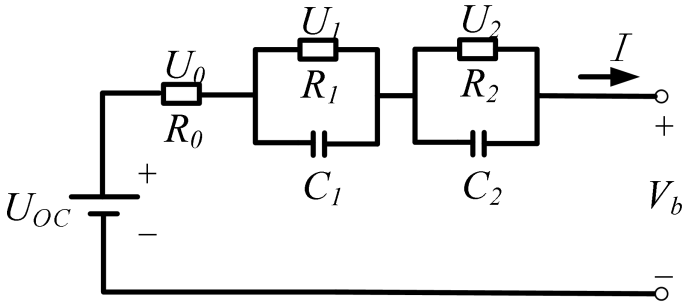


Fig. 1 Second order RC equivalent circuit model of lithium battery

Can be obtained from Kirchhoff's law:

$$V_b = U_{oc} - U_1 - U_2 - IR_0 \tag{1}$$

$$C_1 \frac{dU_1}{dt} = I - \frac{U_1}{R_1} \tag{2}$$

$$C_2 \frac{dU_2}{dt} = I - \frac{U_2}{R_2} \tag{3}$$

The SOC of a lithium battery is defined as the ratio of its residual capacity to its rated capacity, and is calculated by the following formula:

$$SOC(t) = SOC(t_0) - \int_{t_0}^{t} \frac{\eta I}{Q_n} dt \tag{4}$$

Where:t-time;SOC(t)-t moment lithium battery SOC value;SOC(t_0)-t_0 moment lithium battery SOC value; η -Coulomb efficiency,η=1;Q_n-rated capacity of lithium battery.

2.2 Parameter Identification

2.2.1 Battery SOC-OCV relationship curve

Before the estimation of the SOC of lithium battery, it is necessary to identify the parameters of its equivalent circuit model, and the object is the saturation voltage of 4.2V, cut-off voltage of 2.5V, the capacity of 2000mA-h lithium ternary battery as a study. The experimental platform is shown in Fig. 2.

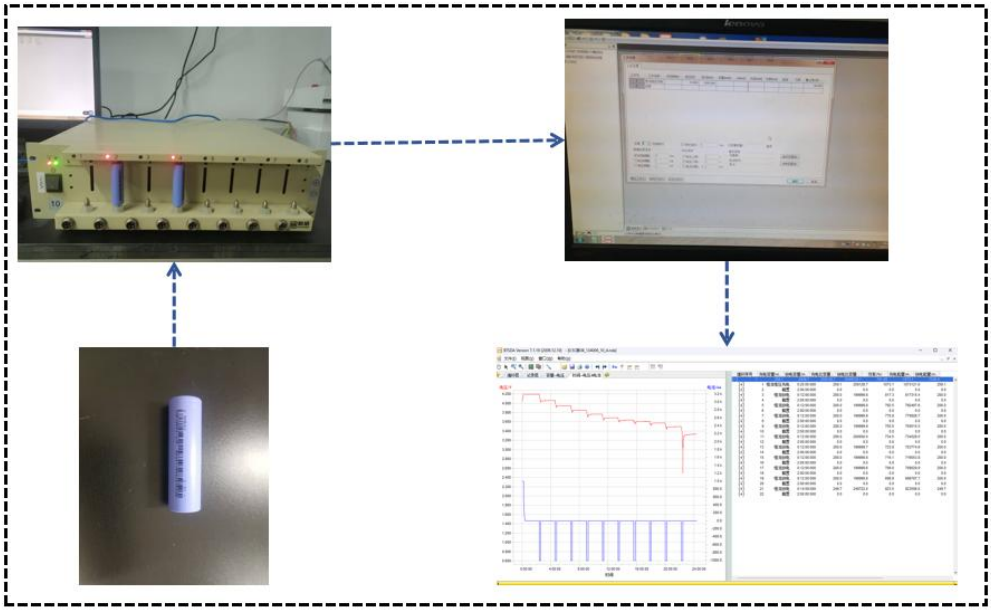


Fig. 2 Battery test platform

Firstly, the mixed pulse power characterization (HPPC) experiments were conducted at a constant temperature of 25 ℃ with reference to the “FreedomCar Power-Assisted Battery Test Manual”, and the purpose of the

experiments was to determine the relationship between the UOC and SOC of Li-ion batteries and to identify the parameters of Li-ion battery models.

Experimental steps: constant current and constant voltage charging to the cut-off current, stationary for 2h, discharge at a constant current of 0.5C (1000mA) for 12min, stationary for 2h, cyclic discharge until the end of the cut-off voltage, and record the changes in voltage and current. As shown in Figure 3.

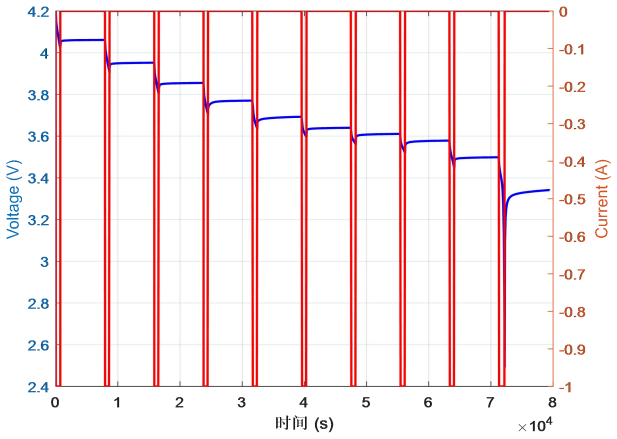


Fig. 3 HPPC current and voltage curves

The relationship equation of SOC-OCV was obtained by HPPC experiment and the curve is shown in Fig. 4.

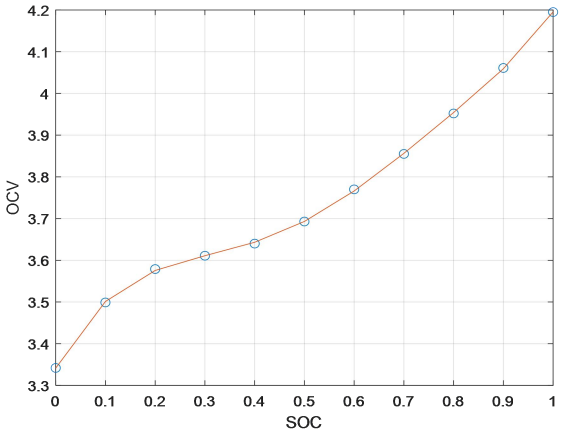


Fig. 4 SOC-OCV relationship plot

A 6th order polynomial was fitted to the SOC-UOC curve to characterize the functional relationship between UOC and SOC:

$$OCV(SOC) = 5.801 * SOC^{6} - 12.43 * SOC^{5} + 4.254 * SOC^{4} + 7.781 * SOC^{3} - 6.739 * SOC^{2} + 2.191 * SOC + 3.341 \quad (5)$$

2.2.2 Parameter Identification by Joint 1stopt Software

Parameter identification can be divided into offline identification and online identification, in which offline identification can make full use of the precise data under experimental conditions, thus ensuring the accuracy of the identification results. In this paper, the method of exponential fitting combined with 1-Stopt software is adopted to carry out offline parameter identification of lithium batteries. Using the HPPC data measured under laboratory conditions as the basis, the internal parameters of the battery are recognized by fitting. As shown in Fig. 5, a certain voltage curve in the HPPC experiment is illustrated as an example.

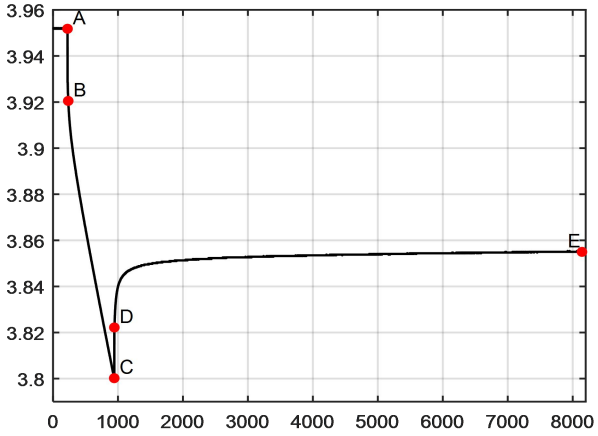


Fig. 5 End voltage response curve

Ohmic internal resistance R0 identification:

At the beginning and end of the discharge, the battery has not yet polarized. The abrupt changes between points A to B and C to D are due to the ohmic internal resistance. Therefore, R0 can be expressed as:

$$R_0 = \frac{(U_A - U_B) + (U_D - U_C)}{2I} \tag{6}$$

The voltage of the battery will rise slowly until stabilized in a period of time when the battery is stationary after pulse discharge, and the voltage response of the RC circuit at this moment is the zero-input response. The slow rise of the terminal voltage in the DE section indicates the polarization response process of R_1C_1 and R_2C_2. The equation for the terminal voltage is:

$$V_b = U_{OC} - U_1 e^{(-t/\tau_1)} - U_2 e^{(-t/\tau_1)} \tag{7}$$

In the above equation, the initial voltages of the two RC links U_1 and U_2, and represent the time constants of each RC parallel circuit. The time constant is expressed as:

$$\begin{cases} \tau_1 = R_1 \cdot C_1 \\ \tau_2 = R_2 \cdot C_2 \end{cases} \tag{8}$$

The custom fit function can be expressed as:

$$y = a - b * \exp(-ct) - d * \exp(-ft) \tag{9}$$

Based on the experimental data obtained from the Hybrid Pulsed Power Characterization (HPPC) test experiments, the terminal voltage data for the segment of SOC from 100% to 90% were first selected and these experimental data were imported in MATLAB. Next, variables with time as the horizontal coordinate and end voltage as the vertical coordinate were created and the data were simplified as necessary for fitting analysis. Subsequently, the CurveFitting toolbox of MATLAB was utilized to select a custom fitting function for fitting analysis, which was in the form of an exponential function. In order to further improve the fitting accuracy, the other four unknown parameters in the model can be optimized with the help of 1stOpt software to determine the initial values, and the obtained optimized parameters are then returned to MATLAB for initialization and final determination of the parameter values. The same steps can be used for other SOC intervals, and the corresponding resistance, capacitance and other parameters are solved sequentially to construct a complete battery equivalent circuit model. The parameters are identified using 1stopt software and the fitted relational equation in MATLAB, and the results of parameter identification are shown in Table 1.

Table 1 Parameter identification results

SOC	R_0/Ω	R_1/Ω	C_1/F	R_2/Ω	C_2/F
0.1	0.086	0.01714	391	0.008	293289

0.2	0.2525	0.01133	189605	0.0198	6022
0.3	0.0249	0.01588	2872	0.009091	146723
0.4	0.02395	0.00716	277112	0.01398	3288
0.5	0.0231	0.01676	133391	0.01419	7620
0.6	0.0237	0.01562	82874	0.02394	7229
0.7	0.0235	0.02379	2446	0.007319	244157
0.8	0.02305	0.01806	2035	0.00624	327923
0.9	0.0226	0.01379	1994	0.00523	258839

2.3 Model Validation

In order to verify the accuracy of the battery model and the accuracy of the identification results, the battery simulation model is built in MATLAB/Simulink, and the internal parameters obtained from the offline identification are brought in, and then the accuracy can be proved by comparing the end voltages of the experiment and the simulation. This time, the current under HPPC working condition is used as input, and the comparison of simulated and experimental end voltages obtained are shown in Figures 6 and 7.

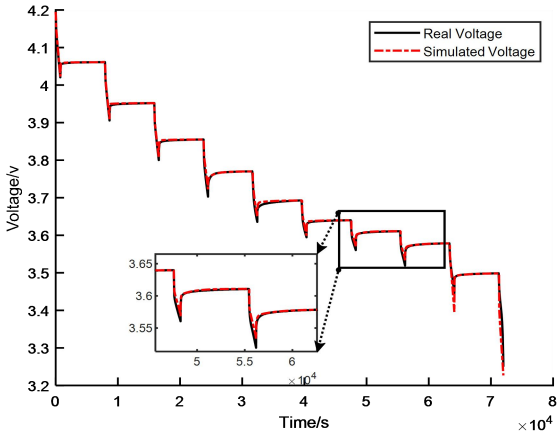


Fig. 6 Comparison curve of model end voltage

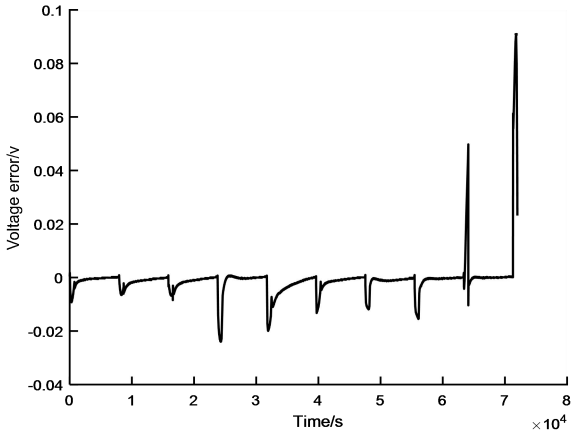


Fig. 7 Model end voltage error curve

From the above figure, it can be found that the model error only increases significantly at the end of discharge, which is caused by the intense chemical reaction inside the battery, and the terminal voltage error at the rest of the stages is within 0.04 V, which meets the battery model accuracy requirements. It proves that the constructed battery model has high accuracy and parameter identification accuracy, which lays the foundation for the next step of battery SOC estimation.

3.SVD-UKF Estimation of Battery SOC

3.1 UKF algorithm

UKF is based on the untraceable (UT) transform, which can estimate the battery condition more accurately. The basic principles of UT transform and UKF and the UKF process for estimating battery SOC are given below.

UT transform is the key step of UKF, and 2n+1 Sigma points can be obtained by sampling according to the symmetric distribution. The formula is shown in the following equation.

$$\begin{cases} X^i = \hat{x}, i = 0 \\ X^i = \hat{x} + (\sqrt{(n+\lambda)P})_i \quad i = 1,2,3,\cdots,n \\ X^i = \hat{x} - (\sqrt{(n+\lambda)P})_{i-n} \quad i = n+1, n+2,\cdots,2n \end{cases} \tag{10}$$

The corresponding weights are:

$$\begin{cases} \omega_m^0 = \lambda / (n+\lambda) \\ \omega_c^0 = \lambda / (n+\lambda) + (1-\alpha^2+\beta) \\ \omega_m^i = \omega_c^i = \lambda / 2(n+\lambda) \quad i = 0,2,\cdots 2n \end{cases} \tag{11}$$

Where, is the dimension of the state variable; is the dispersion factor, the selection of which determines the proximity between the sampling points and the mean value, and usually takes a positive number between 10-6~1; is the distribution factor of the calibration front, and =2 is the optimal for Gaussian distribution; is the auxiliary scale factor to satisfy the value of 0; and is the scale parameter, and is the scaling parameter, and is the scaling parameter, and is the scaling parameter, and the scaling parameter, and is the scaling parameter. Reasonable adjustment of and can improve the estimation accuracy of the algorithm.

Nonlinear Transfer Processing of Sigma Point Sets.

$$y^i = f(X^i) \quad (i = 0,1,\cdots 2n) \tag{12}$$

$$\hat{y} = \sum_{i=0}^{2n} \omega_m^i y^i \tag{13}$$

$$P_y = \sum_{i=0}^{2n} \omega_c^i (y^i - \hat{y})(y^i - \hat{y})^T \tag{14}$$

UKF estimates the battery SOC process as follows:

(1) Initialize state variable means and covariances:

$$\begin{cases} \hat{X}_0 = E[\hat{X}_0] \\ P_0 = E[(X_0 - \hat{X}_0)(X_0 - \hat{X}_0)^T] \end{cases} \tag{15}$$

(2) Condition prediction:

$$\begin{cases} X_{k|k-1}^i = f\left(X_{k-1}^i, u_{k-1}\right) \\ \hat{X}_{k|k-1}^- = \sum_{i=0}^{2n} \omega_m^i x_{k|k-1}^i \end{cases} \tag{16}$$

(3) Temporal updating of state variable error covariances:

$$\begin{cases} y_{k|k-1}^i = h(X_{k|k-1}^i, u_k), i = 0,1\cdots 2n \\ \hat{y}_{k|k-1}^- = \sum_{i=0}^{2n} \omega_i^m y_{k|k-1}^i \end{cases} \tag{17}$$

(4) Temporal updating of error covariances:

$$\begin{cases} P_{yy,k} = \sum_{i=0}^{2n} \omega_c^i [y_{k|k-1}^i - \hat{y}_{k|k-1}^-][y_{k|k-1}^i - \hat{y}_{k|k-1}^-]^T + R_K \\ P_{Xy,k} = \sum_{i=0}^{2n} \omega_c^i [X_{k|k-1}^i - \hat{X}_{k|}^-][y_{k|k-1}^i - \hat{y}_k^-]^T \end{cases} \tag{18}$$

(5) Calculate the Kalman gain:

$$K_k = P_{Xy,k} / P_{yy,k} \tag{19}$$

(6) Update the state variables and covariance matrix of the system:

$$\begin{cases} \hat{X}_k^+ = \hat{X}_k^- + K_k(y_k - \hat{y}_{k|k-1}^-) \\ P_k^+ = \mathbf{P}_k^- - K_k P_{yy,k} K_k^T \end{cases} \tag{20}$$

3.2 The SVD-UKF algorithm

The first step of the UKF algorithm is to perform a traceless transformation of the state variables at the previous moment, where the central step is to obtain the square root of the covariance matrix P . A common method for calculating the square root of the matrix is the Cholesky transformation, but this method is only valid when the matrix is semi-positive definite. In real working conditions, the covariance matrix P may be non-positive definite due to unknown noise and computational errors, which leads to the dispersion of the algorithm.

A singular value decomposition of the error covariance matrix is performed:

$$\boldsymbol{P} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^{\mathrm{T}} = \boldsymbol{U}\begin{bmatrix} \boldsymbol{S} & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{V}^{\mathrm{T}} \tag{21}$$

Where p denotes the matrix to be decomposed, U and V are two orthogonal matrices, and V is a diagonal matrix which can be rewritten as:

$$\sqrt{\boldsymbol{P}} = U\sqrt{\Sigma} \tag{22}$$

This can be obtained using the combination of Eq. (22) and Eq. (10):

$$\begin{cases} X^i = \hat{x}, i = 0 \\ X^i = \hat{x} + (\sqrt{(n+\lambda)}U\sqrt{\Sigma})_i \quad i = 1,2,3,\cdots,n \\ X^i = \hat{x} - (\sqrt{(n+\lambda)}U\sqrt{\Sigma})_{i-n} \quad i = n+1, n+2, \cdots, 2n \end{cases} \tag{23}$$

4.Simulation Verification and Analysis

The improved traceless Kalman filter algorithm is used to estimate the SOC of the Li-ion battery. The initial value of SOC is defined as 0.9, which is used to verify the convergence of the SVD-UKF algorithm. The estimation accuracy of UKF and SVD-UKF is compared by using the ampere-time integration method as the real estimation result, and the estimation results are shown in Fig. 8 and Fig. 9.

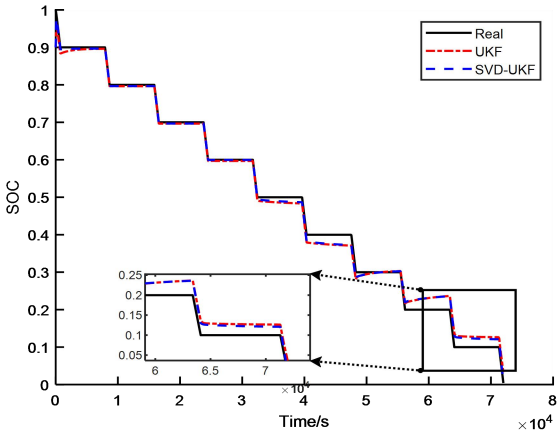


Fig. 8 SOC curve

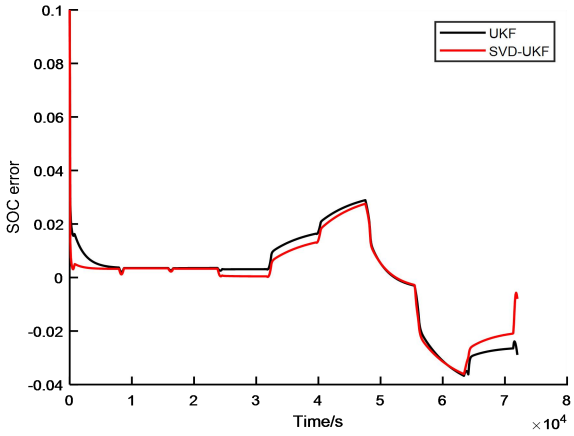


Fig. 9 SOC error curve

Table 2 Comparison of UKF and SVD-UKF errors

Algorithm	Mean absolute error	Maximum absolute error	Root mean square error
UKF	1.3154%	3.6748%	1.7269%
SVD-UKF	1.1234%	3.6102%	1.5603%

As can be seen from the figure and table, the SVD-UKF algorithm has significantly better accuracy than the UKF algorithm in estimating SOC. Whether in terms of average absolute error, maximum absolute error or root mean square error, SVD-UKF performs better.The SVD-UKF algorithm converges faster during the estimation process, especially during the time period of larger error, SVD-UKF can adjust its estimation value more quickly to make it closer to the real value. Compared with the traditional UKF, the average absolute error is reduced by about 14.5%, the maximum absolute error is reduced by about 1.7%, and the root mean square error is reduced by about 9.6%.

The covariance matrix P is changed to P = -1e-2*diag([1 1 1]), at this time the covariance matrix is a non-positive definite matrix to judge the stability of the SVD-UKF algorithm, and at this time, the initial value is changed to 0.8 to judge the robustness of the SVD-UKF. The results are shown in Fig. 10.

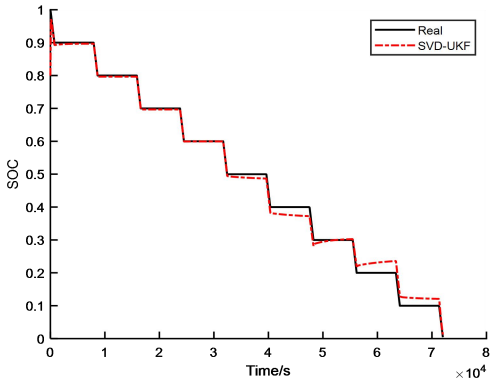


Fig. 10 SVD-UKF estimated SOC curve

From the figure, we can see that when the covariance matrix is a non-positive definite matrix and the initial value is large, the SVD-UKF algorithm can still run normally and converge to the real value quickly, which proves that the algorithm has good stability and robustness.

5. Conclusions

In this paper, the equivalent circuit model of the second-order RC equivalent circuit of the lithium battery is established, and the parameters in the model are identified using matlab and 1stopt, and the simulation model of the lithium battery is established in Matlab, and then the traditional UKF is improved for estimating the SOC of the battery, and the SOC values obtained by the SVD-UKF and the traditional UKF are compared with the actual values, which indicates that the improved algorithm can effectively estimate the SOC value.

References

[1] Tan Zefu, Sun Rongli, Yang Rui, et al. A review of the development of battery management system [J]. Journal of Chongqing University of Technology (Natural Science), 2019, 33(9):40-45.

[2] Genxing Liao, Yingying Zhao, Yanfeng Gao, et al. Model parameter identification and charge state estimation of lithium-ion batteries[J]. Power Supply Technology, 2021, 45(9):1136-1139.

[3] Ko C J, Chen K C. Using tens of seconds of relaxation voltage to estimate open circuit voltage and state of health of lithium ion batteries[J]. Applied Energy, 2024, 357: 122488.

[4] Ouyang J, Lin H, Hong Y. Whale optimization algorithm BP neural network with chaotic map** improving for SOC estimation of LMFP battery[J]. Energies, 2024, 17(17): 4300.

[5] Monirul I M, Qiu L, Ruby R. Accurate SOC estimation of ternary lithium-ion batteries by HPPC test-based extended Kalman filter[J]. Journal of Energy Storage, 2024, 92: 112304.

[6] Wang S, Huang P, Lian C, et al. Multi-interest adaptive unscented Kalman filter based on improved matrix decomposition methods for lithium-ion battery state of charge estimation[J]. Journal of Power Sources, 2024, 606: 234547.

[7] Jiang Qin, Zhang Xuanxiong. Model parameter identification and charge state estimation of lithium-ion battery for electric vehicles[J]. Electronic Science and Technology, 2020, 33(2):32-36.

[8] Wang Shifan, Luo Yang, Dong Liang, et al. Offline identification of the parameters of the second-order Thevenin lithium battery equivalent model [J]. Electronic Design Engineering, 2018, 26(9):46-49. .

[9] Li Huan, Wang Shunli, Zou Chuanyun, et al. Research on SOC estimation based on Thevenin model and adaptive Kalman [J]. Automation Instrumentation, 2021, 42(1):46-51.

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 89-98 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018873

The Application of Laser Forming Technology in Additive Manufacturing and Its Quality Monitoring

Yuhang Yao [1*], Yicheng Shi [2]

[1] College of transport & communications, Shanghai Maritime University,Shanghai,China
[2] School of economics & Management, Shanghai Maritime University,Shanghai,China
*Corresponding author Email: jackafore@163.com
Received 23 June 2025; Accepted 12 July 2025; Published 1 September 2025

Abstract: This paper conducts an in-depth investigation into the field of Selective Laser Melting (SLM) within additive manufacturing (AM) technology, providing a detailed analysis of its market background, current status, and development prospects. As a revolutionary pioneer in future manufacturing, AM technology has rapidly emerged in recent years and found widespread application in high-end manufacturing sectors such as scientific research, healthcare, military industry, aerospace, and aviation. SLM technology, as a critical branch of AM, has become a focal point in numerous fields due to its advantages of high precision, high efficiency, and material diversity. This paper first outlines the industry background of AM technology, including its development history, primary methods, and technical characteristics. Subsequently, by comparatively analyzing the key technical parameters of major domestic and international SLM printing equipment brands, the current market status of SLM technology is revealed. In the competitor analysis, the paper focuses on the latest advancements in melt pool temperature monitoring technology and powder spreading detection technology. It points out the limitations of traditional temperature measurement techniques in aspects such as single-point measurement, accuracy, and response speed. The innovative approaches adopted in this project—namely, photoelectric sensor-fused thermal imaging processing technology and artificial intelligence (AI)-based machine vision inspection methods—are introduced. Furthermore, the paper explores future development trends for SLM technology, including product line expansion, service upgrades, and the positive impact of policy support and industry development. Through the research conducted in this project, we deeply appreciate the critical importance of technological innovation for enhancing the forming success rate of SLM technology and reducing material and energy costs.

Keywords: Additive manufacturing, Selective Laser Melting , Process control, Melt pool temperature monitoring

1. Introduction

Additive manufacturing (AM), widely acclaimed as a revolutionary pioneer in future manufacturing and also known as 3D printing technology, has emerged rapidly in recent years as a novel manufacturing approach. It integrates advanced manufacturing, intelligent manufacturing, green manufacturing, new material development, and precision control technologies, demonstrating remarkable innovative vitality in the manufacturing sector. Currently, AM technology has spawned numerous widely adopted methods, including Fused Deposition Modeling (FDM), Stereolithography (SLA), Selective Laser Melting (SLM), and Selective Laser Sintering (SLS). Each technique possesses distinct characteristics, providing diverse and flexible solutions for manufacturing requirements across various fields.

The rapid emergence and widespread adoption of additive manufacturing (AM) technology in high-end sectors such as scientific research, healthcare, military industry, aerospace, and aviation in recent years primarily stem from its significant advantages over traditional subtractive manufacturing. These advantages include efficient production workflows, minimal manufacturing constraints, and highly intelligent operational characteristics. It is precisely these benefits that have garnered extensive global attention for AM technology and driven continuous expansion of its industry revenue scale.

2. Object and subject of research

According to the Wohlers Report, the global 3D printing (additive manufacturing) market reached USD 18 billion in 2022, representing an 18.1% year-on-year growth. The compound annual growth rate (CAGR) from 2016 to 2022 stood at 14.7%. The global market is projected to expand to USD 29.8 billion by 2025 (18.3% CAGR for 2022–2025), with further growth anticipated to reach USD 85.3 billion by 2030 (23.4% CAGR for 2025–2030). In the Chinese market, the China Additive Manufacturing Industry Development Research Report released by the China Additive Manufacturing Industry Alliance indicates sustained expansion of the domestic AM industry scale over the past four years. Notably, industrial-grade AM equipment accounted for 55% of total sales in China’ s market during 2021, demonstrating the nation’ s robust capabilities and promising prospects in this sector.

Table2-1 The scale of China's additive manufacturing industry

Time	2017	2018	2019	2020	2021	2022	2023	2024
Scale	96	120	158	195	262	330	410	500

Concurrently, the development prospects of the additive manufacturing (AM) industry appear equally promising. According to projections from authoritative research reports such as the Wohlers Report and the China Additive Manufacturing Industry Development Research Report, China's AM industry is expected to maintain a minimum compound annual growth rate (CAGR) of 25.21%, compared to the global projection exceeding 30%. These remarkable growth indicators not only reveal the substantial market potential and expansive development trajectory of the AM industry, but also signify its emergence as a key engine for driving sustained global economic growth.

3. Market research

In metal additive manufacturing (AM), Selective Laser Melting (SLM) stands as one of the most representative and intensively researched processes, and also ranks among the most commercially scaled AM technologies. SLM enables the monolithic fabrication of components with complex internal structures unattainable through conventional manufacturing techniques.

Table3-1 Current metal 3D printing enterprises

Establish	Name	Country	Primary technology	Yearly income(USD)	Remark
1989	3DSystems	American	SLM/SLS	557 million	-
1905	EOS	German	SLM/SLS	-	-
2002	Concept Laser	German	SLM	-	Acquired by GE
2010	SLM	German	SLM	-	-

	Solutions				
2004	Shining 3D	China	SLM/SLS	66.96 million	-
2011	Bright Laser	China	SLM/SLS	5.98 million	-
2009	Farsoon	China	SLM	5.73 million	-

In metal additive manufacturing (AM), Selective Laser Melting (SLM) stands as one of the most representative and intensively researched processes, and also ranks among the most commercially scaled AM technologies. SLM enables the monolithic fabrication of components with complex internal structures unattainable through conventional manufacturing techniques.

Despite substantial advancements in metal AM, process standardization remains challenging, particularly regarding yield assurance for large-scale components. Significant deviations in mechanical properties and geometric accuracy may occur even when identical equipment processes identical parts. This variability stems primarily from the predominant use of open-loop or semi-closed-loop control in current metal AM processes, coupled with inadequate monitoring technologies for intermediate stages such as melt pool temperature, powder spreading quality, and layer-wise formation. Consequently, existing AM systems lack efficient integration of intelligent trajectory tracking, real-time defect recognition, adaptive feedback, and dynamic process adjustment. Implementing comprehensive closed-loop monitoring could substantially enhance part success rates and reduce production costs for enterprises.

Concurrently, China has increasingly prioritized the development of in-process monitoring technologies for metal AM. Multiple governmental bodies including the Ministry of Industry and Information Technology (MIIT) have introduced supportive policies:The Ministry of Science and Technology's 2022 Annual Project Application Guide for the National Key R&D Program "Additive Manufacturing and Laser Manufacturing" (March 2022) specifically emphasizes developing online monitoring and quality evaluation technologies for laser powder bed fusion, including efficient in-situ quality assessment, feature recognition, and parameter control methods.Made in China 2025 explicitly identifies AM as a key domain for advancing intelligent manufacturing equipment, smart production processes, and breakthrough technology development.MIIT's Notice on Declaring 2018 Industrial Technology Foundation Public Service Capability Improvement and Industry Quality Common Technology Promotion Projects (August 2018) proposed subsidies up to ¥3 million per qualified entity for quality control and evaluation of metal 3D printing powders and components.

Thus, advancing SLM melt pool temperature monitoring and online powder spreading detection technologies directly aligns with national AM development strategies and satisfies MIIT's criteria for specialized, sophisticated, distinctive, and innovative "Little Giant" enterprises, indicating substantial growth potential.

4. Research and analysis

In the domain of Selective Laser Melting (SLM) printing equipment, Wohlers Report 2021 authoritatively indicates that four Chinese brands featured prominently among the top ten global SLM equipment manufacturers by sales volume in 2020, demonstrating China's formidable competitiveness in this sector. Specifically, China's Farsoon (3rd), Guangdong Hanbang Laser Technology (4th), Eplus3D (5th), and Bright Laser Technologies (BLT, 7th) secured these rankings, showcasing their exceptional market performance and technological prowess.

Among international manufacturers, Germany's EOS maintained its leading position through technological superiority and brand influence, followed by US-based GE Additive in second place. Furthermore, Germany's

TRUMPF (6th), SLM Solutions (8th), Italy's Sisma (9th), and the Netherlands' Additive Industries (10th) claimed their respective rankings through distinctive competitive advantages, collectively shaping a multifaceted global competitive landscape for SLM equipment.

It should be noted that sales volume of Selective Laser Melting (SLM) equipment does not comprehensively reflect a brand's technological capabilities. Taking domestic brands as examples, Guangdong Hanbang Laser Technology and Eplus3D have achieved substantial sales through specialized market penetration in footwear, medical, and construction sectors. However, these companies primarily target lower-tier market segments, resulting in less prominence in technological sophistication. In contrast, Bright Laser Technologies (BLT) and Farsoon operate at the technological forefront in China with high market expectations. Notably, BLT's strategic focus on SLM contract manufacturing services in recent years has reduced its external equipment sales, further demonstrating the non-linear correlation between sales volume and technical prowess.

Regarding international SLM equipment manufacturers, Germany's EOS and SLM Solutions represent industry benchmarks with leading technological positions in the SLM domain. These brands have gained global recognition and trust through exceptional technical capabilities and diverse application expertise.
This report presents a comparative analysis of technical specifications for high-end SLM systems (build volume ≥ 400×400mm) from four representative domestic and international suppliers: Farsoon, BLT, EOS, and SLM Solutions. Detailed technical parameters are summarized in Tables 4-1 and 4-2.

Table 4-1 Comparison of key technical parameters of major domestic and foreign brands

Brand model	Maximum molding size	Powder Layer Thickness	Laser	Materials	System	Laser Spot Diameter	Speed
Farsoon FS621M	620mm×620mm×1100mm	0.02-0.1 mm	Single-laser (1000W) Four-laser (500W)	Six categories and 14 kinds of metal powders	MakeStar	0.09-0.2 mm	15.2 m/s
Bright LaserBLT-S1000	1200mm×600mm×1500mm	0.02-0.1 mm	Eight-laser (500W) Ten-laser (500W) Twelve-laser (500W)	Six types of self-developed powders	BLT-MCS	-	7m/s
EOS EOS M 400-4	400mm×400mm×400mm	-	Four-laser (400W)	Four categories and 11 kinds of metal powders	EOSTATE	0.1mm	7m/s
SLM Solutions NXG X II 600	600mm×600mm×600mm	0.02-0.2 mm	Twelve-laser (1000W)	No restrictions on the types of theories	MPM LPM	0.08-0.16 mm	10m/s

Table 4-2 Comparative Analysis of Process Monitoring Systems of Major SLM Brands at Home and Abroad

Process monitoring system	Farsoon MakeStar	Bright Laser BLT-MCS	EOS EOSTATE	SLM Solutions MPM、LPM
Monitoring items and technologies	equipment temperature, oxygen content, piston position, printing height	Part self-inspection, powder quality monitoring, 3D reconstruction, stress field monitoring	Production process data, powder bed monitoring, implementation process monitoring of the melting point of the molten pool, and non-destructive testing of the finished products	Molten pool monitoring, laser power monitoring
Monitoring technology	Equipment data stream Analysis record	Machine vision Artificial intelligence	Optical tomography, machine vision, beam path axis sensors	The technology has not been disclosed
Closed-loop control	Non-closed-loop control type	Automatic closed-loop control, feedback and repair are carried out for the height compensation of the fabricated parts and the powder coating defects	Non-closed-loop control type	Non-closed-loop control type

All four aforementioned SLM equipment manufacturers incorporate process monitoring systems, yet each exhibits critical limitations including incomplete monitoring capabilities and an inability to implement closed-loop control. In summary, even the most advanced domestic and international SLM equipment producers have not yet developed comprehensive and efficient additive manufacturing process monitoring systems. Crucially, none have achieved closed-loop feedback control utilizing real-time process monitoring data — representing a pivotal technological frontier in current additive manufacturing development.

Significant challenges persist during part formation in Selective Laser Melting (SLM) due to the process's inherent complexity. The technique involves intricate physical, chemical, and metallurgical interactions that frequently induce defects including balling phenomena, porosity, and microcracking. Compounding these issues, cyclic high-energy laser irradiation generates extreme thermal gradients, while rapid solidification shrinkage of the moving melt pool under strong interfacial constraints — coupled with transient non-equilibrium cyclic solid-state phase transformations — produces substantial residual stresses that often manifest as severe part distortion and stress-corrosion cracking. Furthermore, powder spreading defects critically compromise product integrity through recoater blade anomalies (curling, lifting, mechanical damage), contamination trails, and powder bed fragmentation, collectively contributing to fissure formation, non-uniform layer deposition, and material contamination.

The uncontrollable nature of these formation issues compromises dimensional accuracy while rendering components inadequate in both processability and service performance. SLM equipment exhibits critically low

success rates when manufacturing high-performance parts — validation data indicates yield rates of 60-70% for domestic systems versus 80-90% for international counterparts when producing precision components. This yield deficiency increases material costs by in excess of 70% for high-performance SLM-fabricated parts, establishing low formation success rates as the most critical development bottleneck in current SLM additive manufacturing technology.

Implementing comprehensive melt pool monitoring and control systems in SLM equipment is paramount for overcoming low forming success rates. According to Wohlers Report 2021, integrating melt pool monitoring reduces non-essential material costs by ⩾ 60% and energy consumption by ⩾ 50%. Furthermore, deploying a multi-dimensional process control system incorporating both melt pool and powder spreading monitoring can elevate SLM success rates to ⩽ 90%, while reducing material waste by >80% and energy costs by >55%. Consequently, developing closed-loop control systems with integrated melt pool and powder bed monitoring capabilities substantially enhances profit margins within the SLM additive manufacturing industry.

5. Research result

5.1 Temperature Monitoring Technology

Contemporary melt pool temperature monitoring devices predominantly employ conventional techniques — including thermocouple direct measurement, two-color pyrometry, and CCD/infrared thermography—all exhibiting significant limitations. Thermocouple direct measurement provides relatively high accuracy (±5°C) but captures only single-point data, failing to monitor the full melt pool thermal profile. Two-color pyrometry suffers from low accuracy (>±50°C) due to emissivity uncertainties in determining true melt pool temperatures, coupled with prohibitive implementation costs ($15k-$35k). Meanwhile, CCD and infrared thermography exhibit inadequate response speeds (100-500ms) for dynamic laser processes. Our integrated photoelectric-thermal imaging fusion technology overcomes these constraints by achieving sub-millisecond response (0.5ms) while enabling spatiotemporal mapping of thermal evolution dynamics along laser trajectories and characterization of interlayer thermal history during overlapping deposition processes.

Table5.1-1Contemporary melt pool temperature monitoring devices predominantly employ conventional techniques

Type	direct thermocouple	CCD thermography	infrared thermography	two-color pyrometry	Photodiode pyrometry
Accuracy	±0.4%	±2.1%	±2.0%	±2.0%	±1.35%
Temperature	-200~2800°C	-50~1400°C	150~1600°C	400~2200°C	300~4000°C
Times	0~5s	15fps	>25HZ	5ms~99.99s	3.6μs
Measure	Point	Area	Area	Point	Point
Cost(RMB)	1000~2000	20000~50000	10000~20000	10000~30000	1000~2000

5.2 Powder Spreading Detection Technology

Contemporary powder spreading inspection systems predominantly utilize conventional non-destructive testing (NDT) methodologies — including magnetic particle testing (MT), liquid penetrant testing (PT), and radiographic testing (RT)—each constrained by significant operational limitations. MT is exclusively applicable to ferromagnetic materials and necessitates magnetic particle application during inspection. PT requires sequential surface

application of cleaning agents, penetrants, and developers, while RT demands radiation-generating equipment and produces hazardous ionizing radiation. To overcome these constraints, next-generation powder bed monitoring systems employ AI-driven machine vision technology. This innovative approach utilizes industrial-grade CMOS area-scan cameras to capture comprehensive powder bed images for computational processing, enabling multi-material defect detection with enhanced feature extraction capabilities for richer flaw characterization.

Table5.2-1 Powder Spreading Detection Technology

Type	Magnetic particle	liquid penetrant	Ultrasonic waves	Radiographic	Machine vision
Advantages	Lossless Simple operation Low cost	No restrictions on the materials	No restrictions on the materials Simple operation	Available images Accurate	More information
Disadvantages	Only for ferromagnetic materials	Complicated process	Immature	Radiation	Low accuracy

6. Prospects for further research development

Competitors include products from German companies such as EOS, SLM Solutions, Huashu High-Tech, and Platinum Tech. Among these, EOS's EOS M 400-4 lacks a melt pool temperature monitoring system and cannot perform temperature closed-loop control; The NXG X II 600 from SLM Solutions, the FS-721M from Huashu High-Tech, and the EOS M 400-4 all lack powder bed quality monitoring functionality. While the BLT-S1000 from Platinum Technology can monitor powder bed quality, it does not form a closed-loop system and cannot provide timely closed-loop feedback or adjust process parameters in response to detected defects.

Regarding current scientific research on process monitoring in SLM, the article "Online Detection of Melt Pool Temperature in Selective Laser Melting Metal Forming" published in the March 2020 issue (Volume 47, Issue 3) of China Laser utilized a PIN-type photodiode to detect melt pool radiation and constructed a composite amplification circuit to measure the photocurrent, effectively detecting the small-amplitude, rapidly changing radiation signals from the melt pool. The study successfully obtained the radiation spectrum information of the melt pool in the 540–660 nm wavelength range for a 90% Cu – 10% Sn alloy powder material during the SLM forming process, fully demonstrating the feasibility of using photodiodes to collect radiation signals and obtain melt pool temperature information. In the October 2021 issue of the Journal of Aeronautics, Volume 42, Issue 10, titled "A Review of Intelligent Monitoring and Process Control of Defects in Laser Selective Melting Additive Manufacturing," it is summarized that in terms of SLM feedback control, the primary focus is on considering the dimensional and shape-specific characteristics of the component during the process planning stage, setting paths and process parameters to achieve quality control. Research on real-time feedback control based on monitoring signals is limited and requires further exploration. One of the primary trends in current research on intelligent monitoring and real-time feedback control for SLM processes is the development of full-process monitoring and real-time quality control technologies for full-sized components. Currently, most research on SLM process monitoring and control focuses on single-scan processes or small-sized simple test specimens. How to achieve full-process monitoring and control for the manufacturing of large-sized components in industrial production remains a significant challenge and is also an important future research direction.

Given the significant limitations of products currently developed by domestic and international companies, which feature relatively simple melt pool temperature detection capabilities and lack closed-loop control functionality for powder spreading quality, we have developed a closed-loop temperature measurement device based on optoelectronic fusion thermal imaging online detection and a closed-loop powder spreading defect detection device based on machine vision detection technology. For the closed-loop temperature measurement device, it not only enables real-time monitoring of the melt pool temperature but also performs closed-loop control of the melt pool temperature. For the closed-loop powder bed defect detection device, it uses artificial intelligence algorithms to extract various defect features of the powder bed in real time and adjusts process parameters based on defect feature information.

7. Conclusions

This study achieves significant technological breakthroughs in process monitoring. For melt pool temperature detection, the implementation of photoelectric sensor-fused thermal imaging processing technology enhances both response speed and measurement accuracy. This system effectively captures the thermal evolution dynamics along laser trajectories and interlayer thermal history during overlapping processes, resolving limitations inherent in conventional thermometry such as single-point measurement, low precision, and slow response.

Regarding powder spreading quality inspection, the adoption of AI-driven machine vision enables comprehensive powder bed imaging for computational analysis. This methodology facilitates multi-material defect detection while enhancing feature extraction capabilities to acquire richer flaw information, thereby achieving real-time monitoring and closed-loop control of powder bed quality.

These technological innovations substantially improve SLM forming success rates while reducing material consumption by 60% and energy costs by 55% according to validation data. Consequently, they generate enhanced profit margins for the SLM additive manufacturing industry. Future development will increase R&D investment to further advance system performance and competitive positioning.

CONFLICT OF INTERESTT

The authors declare no conflicts of interest relevant to this study.

References

[1] Hao Fei, Zhu Jinyao, Shang Yufeng, Hao Yanan & Xiao Jinghua.(2025) Implementation and Research of Variable Constant Gratings Based on Patterned Cutting Method University physics experiment, 38 (03), 1-6. Doi: 10.14139 / j.carol carroll nki cn22-1228.2025.03.001.

[2] Li Sha, Gao Wei, Chen Lei & Liang Xiaoli.(2025). Trajectory Planning Method for Complex Intersection Line Cutting of Laser Processing robots. Laser, 46 (6), 226-231. The doi: 10.14016 / j.carol carroll nki JGZZ. 2025.06.226.

[3] Yang Zhiquan, Cao Jinghu, Lei Na, Xia Bifeng, Zhou Zhichao & Zhang Xiuli.(2025) The Application and Practice of Laser Cutting Technology in the Inspection of Hot-rolled Finished Products Metallurgy and quality standardization, 63 (03), 44, 48, doi: 10.26915 / j.i ssn1003-0514.2025.03.012.

[4]Peiying Bian,Ali Jammal,Kewei Xu,Fangxia Ye,Nan Zhao & Yun Song.(2025).A Review of the Evolution of Residual Stresses in Additive Manufacturing During Selective Laser Melting Technology..Materials (Basel, Switzerland),18(8),1707-1707.

[5](2025). Application of Laser Cutting Machines in the New Energy Industry. Modern Manufacturing,(02),58-59.

[6] Zhou Biao.(2024). Research and Application of Process Parameters for Laser Cutting of Carbon Steel Plates. China Machinery,(36),90-93.

[7] Cui Qi, Li Ming & Yang Rong.(2024). Exploration of Project-driven Laser Processing Training Courses China's modern education equipment, (23), 61-63. The doi: 10.13492 / j.carol carroll nki cmee. 2024.23.031.

[8] Xiong Yanfei & Liu Dengbang.(2024). Research on Visual Image Target Annotation of Laser Cutting Robots. Laser and Infrared,54(12),1864-1870.

[9] Zhou Hao, Deng Miaowen, Yang Yuanzheng & Shi Junlei.(2025). The influence of laser cutting speed on the magnetic properties of iron-based amorphous alloys. Hot working process, 54 (05), 46-50. Doi: 10.14158 / j.carol carroll nki. 1001-3814.20223541.

[10] Jiang Huaqiao, Cui Yinhui & Zhang Sen.(2024). Research and Application Progress of Laser Processing Technology in the Field of Aviation Structural Components. Application of laser, 44 (6), 70-79. The doi: 10.14128 / j.carol carroll nki. Al. 20244406.070.

[11] Wei Zhuo & Zhang Hong.(2024). A Review of Data-Driven Optimization Methods for Laser Cutting Process Parameters. Mechanical and Electrical Engineering Technology,53(04),1-5+78.

[12] Qian Jun, He Yong, Xu Danhong & Zhang Junjie.(2024). Research on Laser processing Technology and in vitro biocompatibility of Biological hydrogels. Electrical Discharge Machining and Dies,(02),43-47.

[13] Xu Zhaohua, Zheng Zhicong, Gu Ruiyu, Sheng Hui & Zhang Kai.(2024) Research and Application of Five-Axis Linkage Laser Precision Cutting Method Based on Reverse Engineering Modern manufacturing engineering, (4), 135-139. The doi: 10.16731 / j.carol carroll nki. 1671-3133.2024.04.018.

[14] Yan Furong, Wang Chenxi & Gao Zhiyuan.(2024). Research on Structural Design and Control System of Laser Cutting Device. Electronic production, 32 (08), 101-103. The doi: 10.16589 / j.carol carroll nki cn11-3571 / tn. 2024.08.019.

[15] Chen Zhenqiang.(2024). Practical Application of Modern Laser Technology in Aviation Mechanical Processing. China Machinery,(07),2-5.

[16]Jibing Chen,Yong She,Xinyu Du,Yanfeng Liu,Yang Yang & Junsheng Yang.(2024).Influence of oxygen content on selective laser melting leading to the formation of spheroidization in additive manufacturing technology..RSC advances,14(5),3202-3208.

[17] Liu Liu, Jiang Huannian, Fei Yongyun & Zhao Yang.(2023). Design of a Control System for a one-to-Two Loading and Unloading Device in Laser Cutting. Automation Applications,64(23),35-37.

[18] Qiao Yongqiang, Li Renge, Li Zhiyu & Wen Qingnian.(2023). Analysis of the Processing Principle and Structural Composition of CNC Laser Cutting Machines. China Equipment Engineering,(23),72-74. In Practice. China

Machinery,(07),2-5.
[19]Gisario Annamaria,Barletta Massimiliano & Veniali Francesco.(2022).Correction to: Laser polishing: a review of a constantly growing technology in the surface finishing of components made by additive manufacturing.The International Journal of Advanced Manufacturing Technology,123(3-4),1401-1401.
[20]Annamaria Gisario,Massimiliano Barletta & Francesco Veniali.(2022).Laser polishing: a review of a constantly growing technology in the surface finishing of components made by additive manufacturing.The International Journal of Advanced Manufacturing Technology,120(3-4),1433-1472.
[21]Mohamed, Omar Ahmed & Xu, Wei.(2021).Comment about lack of sufficient data on “A prediction model for finding the optimal laser parameters in additive manufacturing of NiTi shape memory alloy” by Mehrpouya et al. [The International Journal of Advanced Manufacturing Technology 105.11 (2019): 4691-4699.Progress in Additive Manufacturing,7(2),1-8.
[22] Wu Weihui, Ma Gengxiong, Wang Di, Ma Xuyuan & Liu Linqing.(2022). Construction and Experiment of Real-time Powder Mixing Gradient Material SLM Forming System Laser Technology,46(04),492-498.
[23] Feng Shuying & Zhang Huimei.(2020). Research Progress of Selective Laser Sintering. Jiangxi chemical, (4), 56-57. Doi: 10.14127 / j.carol carroll nki jiangxihuagong. 2020.04.018.
[24](2019).Technology - Additive Manufacturing; Investigators at Ruhr University Bochum Discuss Findings in Additive Manufacturing [Cavitation Erosion Resistance of 316l Austenitic Steel Processed By Selective Laser Melting (Slm)].Journal of Technology,1030-.
[25] Zhang Chunyu, Chen Xianshuai & Sun Xuetong.(2020). The Development of Metal 3-D Printing Manufacturing Technology. Laser Technology,44(03),393-398.
[26] Sun Jianfeng, Yang Yongqiang & Yang Zhou.(2016). Research on Surface Roughness of Ti6Al4V by Selective Laser Melting Based on Powder characteristics. Chinese Journal of Lasers,43(07),104-113.
[27] Zhu Yanqing, Shi Jifu, Wang Leilei, Zhong Liuwen, Li Yujian & Xu Gang.(2015). Current Development Status of 3D Printing Technology. Manufacturing Technology and Machine Tools,(12),50-57.
[28] Zang Jialun, Sun Yucheng, Li Chuang & Li Zhiyong.(2015). Domestic Casting Rapid Prototyping Technology and Application. China Foundry Equipment and Technology,(04),1-5.
[29] Yu Hao.(2015). Overview of 3D Printing Development Strategies in the European Union and Asia. New Materials Industry,(05),25-30.
[30] Song Changhui, Yang Yongqiang, Wang Yunda, Yu Jiakuo & Mai Shuzhen.(2014). Research on Laser Selective melting Forming Process and Properties of CoCrMo alloy. Chinese Laser,41(06),58-65.

Global Academic Frontiers, Vol. 3, Issue 3, 2025, pp. 99-119 Print ISSN 2995-5688 Online ISSN 2995-570X
DOI: https://doi.org/10.5281/zenodo.17018925

A Study of the Impact of ESG on Corporate Carbon Performance -- Based on the mediating effect of New quality productivity

Ying Wu [1] Yu Liao [1*] Dai-Yun Li [2]

[1] College of School of Economics, Wuhan Textile University, Wuhan 430200, China
[2] College of School of Economics and Management, Guangzhou Institute of Science and Technology, Guangzhou 510000, China
*Corresponding author Email: 825946265@qq.com
Received 17 June 2025; Accepted 12 July 2025; Published 1 September 2025

Abstract: To investigate whether corporate ESG can improve corporate carbon performance and facilitate the achievement of China's "dual-carbon" objective, we utilize panel data from China's A-share-listed companies spanning 2015 to 2022. We develop new quality productivity indicators through principal component analysis and empirically assess the impact and mechanisms of ESG on corporate carbon performance using bidirectional fixed-effects, mediated-effects, and moderated-effects models. The influence and mechanism of ESG on corporate carbon performance are experimentally analyzed utilizing a two-way fixed effects model, a mediation effects model, and a moderating effects model. Research indicates that (i) strong ESG performance can markedly enhance corporate carbon performance; this conclusion remains valid following various robustness and endogeneity assessments, including substituting core explanatory variables, lagging explanatory variables by one period, and augmenting the standard error of clustering at the city level. (ii) Mechanism analysis indicates that ESG performance enhances the carbon performance of firms both directly and indirectly by fostering advancements in new quality productivity. (iii) The examination of moderating effects indicates that company financialization enhances the favorable influence of ESG on corporate carbon performance. Heterogeneity study indicates that ESG exerts a more pronounced influence in the eastern area, among non-state-owned firms, and within non-heavily polluting enterprises. Consequently, to advance the low-carbon transformation of the economy, it is essential to stratify the policy design. Incentives for ESG can be enhanced for non-state-owned enterprises, manufacturing sectors, and the eastern region, while technology subsidies or differentiated assessment criteria are required for state-owned enterprises, heavily polluting industries, and the central and western regions to address the structural disparities in ESG practices and to encourage non-manufacturing industries to pursue synergistic avenues of digital transformation and ESG integration.

Keywords: ESG, Carbon performance, New quality productivity, Mediating effect, Moderating effect

I. INTRODUCTION

In recent years, as global climate change escalates and resource limitations tighten, advancing the green and low-carbon transformation of firms has emerged as a crucial option for nations to address the environmental issue. In this context, and in reaction to the problems presented by climate change, the Chinese government committed at the 2020 United Nations General Assembly to reach peak carbon emissions by 2030 and attain carbon neutrality by 2060. This objective indicates a transition from the conventional high-carbon emission paradigm to a more

environmentally friendly, low-carbon, and sustainable development framework for China and the globe. Throughout this shift, ESG—Environmental, Social, and business Governance—has gained prominence as a critical metric for assessing business sustainability and social responsibility. Enterprises must not only seek to maximize economic benefits but also improve resource efficiency and minimize environmental impact through innovation, social responsibility, and governance optimization. This is particularly crucial for listed companies, which are expected to lead in advancing low-carbon transformation and high-quality economic development. Nonetheless, throughout implementation, the limitations of the conventional productivity growth model have increasingly become apparent, hindering effective support for the achievement of carbon objectives. The swift advancement of the information technology revolution, encompassing artificial intelligence, big data, digital transformation, and green technology, has significantly altered the essence of productivity. Currently, China's economy has transitioned from high-speed growth to high-quality development, rendering traditional productivity metrics insufficient to adequately capture the new impetus and potential of economic advancement. Consequently, the notion of "new quality productivity" has arisen, which not only enhances and broadens the conventional understanding of productivity but also addresses the demands of economic advancement in the contemporary day. Can good ESG governance enhance company success in reducing carbon emissions?

Presently, ESG research indicates a progression from disjointed analysis to systematic integration, with its theoretical framework transcending a singular focus on financial performance to establish a three-dimensional dynamic coupling of environmental responsibility, social accountability, and governance efficacy[1]. Scholars widely concur that ESG practices generate long-term value through enhanced capital allocation, improved supply chain synergies, and the promotion of low-carbon technology; yet, short-term cost pressures and delayed rewards remain important grounds of contention. Particularly in emerging markets, the unique attributes of the policy landscape and market mechanisms have resulted in notable disparities in the efficiency of the "compliance-innovation" transformation of ESG inputs, while the fragmentation of the global ESG rating system has exacerbated the ambiguity surrounding corporate decision-making. Research on carbon performance is transitioning from static accounting to dynamic capacity development. The initial emphasis on emission measurement has transitioned to the interactive analysis of driving mechanisms and value creation: on one hand, the correlation between carbon emission intensity and financial performance has transcended the conventional “cost-burden theory,” with the technology compensation effect and green premium mechanism increasingly elucidating the symbiotic phenomenon of “emission reduction and profitability”; on the other hand, the novel avenues of supply chain synergy for carbon emission reduction and the empowerment of digital technology are transforming corporate carbon management[2]. Conversely, emerging avenues like supply chain synergy and digital technology empowerment are transforming the strategic framework of enterprise carbon management. Nonetheless, the technical impediment of low-carbon transformation in heavy industries, the insufficient carbon accounting capabilities of small and medium-sized firms (SMEs), and the fragmentation of the global carbon market remain challenges that the academic community must address.Research on company financialization elucidates the "double-edged sword" effect of capital allocation. Financial asset holdings can alleviate financing limitations and furnish capital reserves for green investments; yet, they may also inhibit genuine innovation resources due to arbitrage preferences, a conflict that is especially evident in the framework of shareholder value maximization. The policy environment significantly influences the trajectory of financialization; a bank-dominated financial system may diminish disincentives for innovation, but a competitively neutral policy can limit the regulatory arbitrage of state-owned enterprises (SOEs). The present study focus has transitioned to the processes by which technical governance solutions, including blockchain and ESG financial instruments, can mitigate the inclination towards "deconcentration."

Investigations on the relationship between ESG and carbon performance are in a phase of theoretical refinement. Current findings primarily emphasize the connection between technological innovation, such as clean technology research and development, and institutional design, like carbon information disclosure. However, they overlook three critical issues: first, the inconsistency of ESG ratings results in distorted market signals, diminishing enterprises' incentives to lower emissions; second, there exists a risk of disruptions in the tracing of supply-chain carbon data, which compromises the efficacy of comprehensive value-chain management; and third, "greenwashing" practices obscure the true nature of carbon performance through strategic disclosure, rendering such behavior inadequate for achieving emission reductions. Third, "greenwashing" practices obscure the fundamental nature of emission reduction by deliberate information dissemination. These vulnerabilities underscore the necessity of establishing a verifiable and traceable transmission chain for ESG-carbon performance.The emergence of new quality productivity theory offers a fresh perspective for carbon performance studies. The framework situates productivity evolution within a three-dimensional coordinate system comprising technological revolution, factor reconfiguration, and institutional innovation, highlighting the synergistic development of green technological innovation, digital empowerment, and business ecological reconfiguration[3]. The fundamental advancement consists of demonstrating the intrinsic connection between the reduction of carbon emission intensity and the improvement of total factor productivity; however, empirical validation at the micro-firm level remains inadequate, particularly regarding the specific dynamics of the technology-institution-capital triadic interaction, which requires further analysis.The control of ESG-carbon performance through financialization is intricate. The short-term arbitrage incentive may undermine long-term emission reduction investments in ESG; nevertheless, instruments like green bonds and carbon futures can offer risk mitigation and financial backing for the low-carbon transition. The origin of this contradiction resides in the selection of corporate strategic emphasis—when financialization supports technological advancement instead of dissociating from the organization, it can create a "capital-technology-emission reduction" enhancing loop[4]. The difficulty of how policy design may effectively direct financial resources toward significant emission reduction regions has emerged as a prevalent issue in both theory and practice.There are considerable deficiencies in the current literature regarding multidimensional integration, dynamic mechanisms, and local adaptation: ESG research inadequately addresses the influence of non-financial factors; Carbon performance studies fail to adequately respond to emerging trends such as new quality productivity; and there is a scarcity of empirical evidence to substantiate the distinctive trajectory of ESG-financialization-emission reduction among firms in emerging markets. This study aims to address these theoretical blind spots[5].

The potential marginal contributions of this paper are as follows: Initially, the advancement of the theoretical framework. Limited domestic research on carbon emission reduction pertains to new quality productivity, and even fewer studies examine the correlation between ESG and carbon performance, lacking a systematic theoretical framework and empirical validation; therefore, there is a need for the development of the local context[6]. International scholars often isolate individual elements from ESG or perpetuate the CSR research framework to examine carbon-related issues, while the notion of "new quality productivity" remains underrepresented in their academic discourse, rendering their conclusions inapplicable to the Chinese context. This research offers empirical evidence for the localization of the new quality productivity hypothesis, based on data from Chinese A-share listed enterprises[7]. Third, the enhancement of heterogeneity analysis. This paper examines heterogeneity across four dimensions: ownership characteristics, industry type, geographic location, and pollution levels, elucidating the varying effects of ESG performance on carbon performance in distinct contexts, and offering micro-evidence of firms' enhancement of new quality productivity. The examination of the moderating influence[8]. This study presents corporate financialization as a moderating variable, elucidates its function in the relationship between ESG

performance and carbon performance, and offers a novel theoretical viewpoint for mitigating economic "deconcentration."

II. THEORETICAL ANALYSIS

2.1 Sustainable Development Theory

The creation and progression of the sustainable development theory arise from humanity's methodical contemplation of the adverse externalities associated with industrial civilization. The concept originated from the ecological and environmental warnings presented in Silent Spring during the 1960s, but it has undergone three significant developmental phases to achieve global acceptance. The 1972 United Nations Conference on the Human Environment (UNCHE) was the inaugural event to integrate environmental concerns into the international political agenda[9], presenting the interdependent relationship between the right to development and environmental protection, thereby establishing a dualistic equilibrium between ecology and economy. " In 1987, the Brundtland Report defined sustainable development as "meeting the needs of the present without compromising the ability of future generations to meet their own needs." In 1987, the Brundtland Report defined sustainable development as "meeting the needs of the present without compromising the ability of future generations to meet their own needs," established the principle of intergenerational equity, and incorporated social equity into the theoretical framework. In 1992, the Earth Summit in Rio adopted Agenda 21, which advanced the transition from concept to action, highlighting the collaboration of government, enterprises, and society, during which enterprises were expressly recognized as the primary custodians of environmental responsibility.

As we entered the 21st century, the theoretical framework evolved to encompass the interplay between technological innovation and institutional transformation. The 17 Sustainable Development Goals (SDGs) outlined in the 2015 UN 2030 Agenda for Sustainable Development integrated specific environmental metrics, such as clean energy and climate action, with corporate governance structures and stakeholder engagement for the first time, signifying a shift from macro-policy considerations to micro-enterprise practices in sustainable development. This signifies the integration of sustainable development from the macro policy level to micro enterprise practices. Within this theoretical framework, the ESG (Environment, Society, Governance) system has effectively transformed into the operational framework for sustainable development goals at the corporate level: the environmental dimension directly mitigates carbon emission intensity per unit of output through the advancement of clean technology and carbon footprint management[10]; the social dimension indirectly curtails carbon leakage within the value chain by promoting supply chain greening and fostering low-carbon behaviors among employees; and the governance dimension implements a carbon management system. The governance aspect creates a continual improvement process by implementing a carbon management system and disclosing carbon emission data[11]. This triad of pathways positions ESG as a crucial link between corporate micro-decision-making and macro carbon neutrality objectives, while also offering robust theoretical basis for the empirical analysis of the influence of ESG performance on carbon outcomes[12].

2.2 Corporate Social Responsibility Theory

The development of corporate social responsibility (CSR) philosophy illustrates the transition of global business from a profit-centric model to a value-oriented one. The ideological lineage originates from Howard Bowen's notion of "businessmen's social responsibility" in the 1950s, which underscored the necessity for corporate decision-making to transcend economic objectives and address social expectations, thereby establishing the foundation for the subsequent connection between social responsibility and environmental concerns[13]. In the 1960s, Milton

Friedman's "Shareholder Priority Theory" incited vigorous debates, compelling the academic community to develop a more systematic explanatory framework—the CSR pyramid model introduced by Archie Carroll in 1979, which explicitly incorporated environmental responsibility into corporate ethics. At the close of the 20th century, Edward Freeman's stakeholder theory elucidated environmental responsibility, underscoring the significance of corporate accountability to governmental environmental regulations. At the conclusion of the 20th century, Edward Freeman's stakeholder theory elucidated environmental responsibility, underscoring that corporate adherence to governmental environmental regulations and community ecological expectations directly influences its legitimacy[14]. Furthermore, the carbon footprint, as a measurable indicator of environmental externalities, has emerged as a pivotal metric for assessing corporate environmental responsibility. In the 21st century, Michael Porter's theory of shared value advocates for the evolution of corporate social responsibility (CSR) into a strategic framework, encouraging the conversion of social challenges into business opportunities, such as minimizing carbon emission costs via clean technology research and development, meeting environmental obligations, and augmenting competitive advantages, alongside the dual benefit mechanism of "carbon emission reduction - efficiency enhancement."

The double dividend mechanism of "carbon emission reduction - efficiency improvement" establishes a theoretical connection between the ESG (environmental, social, governance) framework and carbon performance. The ESG framework can be viewed as an indexed expansion of CSR theory. The environmental (E) dimension pertains to enhancements in carbon performance, achieved by diminishing carbon emission intensity per unit of output value through technological advancements, including investments in renewable energy and the decarbonization of production processes. The social (S) dimension mitigates carbon leakage within the value chain via green supply chain management and low-carbon product innovation. The governance (G) dimension is contingent upon the board of directors' oversight of climate issues and the disclosure of carbon-related information. The governance (G) dimension depends on the board of directors' oversight of climate monitoring, carbon information disclosure, and various institutional frameworks to guarantee the ongoing achievement of emission reduction objectives. According to this rationale, CSR theory elucidates the inherent motivation behind ESG practices that enhance carbon performance—specifically, the internalization of environmental externalities via resource reallocation, technological innovation, and institutional transformation. Furthermore, it highlights the potential moderating influence of financialization: when firms excessively prioritize short-term profits from financial assets, they may undermine long-term ESG-related low-carbon investments. It also indicates the potential moderating influence of financialization: when companies excessively seek short-term gains from financial assets, they may displace ESG-related long-term low-carbon investments, undermine the resource foundation for fulfilling environmental responsibilities, and consequently hinder advancements in carbon performance. This theoretical approach establishes the basis for empirically investigating the relationship between ESG and carbon performance, as well as the constraints of financialization.

2.3 Stakeholder Theory

The development of stakeholder theory has significantly demonstrated the transition from exclusive shareholder value maximization to inclusive value co-creation. The term of "stakeholder," introduced by the Stanford Research Institute in 1963, denotes any group that influences or is influenced by the achievement of business objectives; nevertheless, a rigorous analytical framework had not yet been established at that time. In 1984, R. Edward Freeman, in his book "Strategic Management: A Stakeholder Approach," revolutionized the perception of the firm as a nexus connecting shareholders, employees, consumers, suppliers, and the public. In 1984, R. Edward Freeman, in "Strategic Management: A Stakeholder Approach," pioneered the concept of enterprises as a contractual network linking shareholders, employees, consumers, suppliers, communities, governments, and other

diverse entities. He argued that long-term value creation depends on balancing the rights and interests of all stakeholders, establishing a foundation for comprehending the multi-dimensional driving mechanism of ESG (Environmental, Social, Governance) - Environmental Responsibility (E This establishes the basis for comprehending the various driving forces of ESG (environmental, social, and governance). Environmental responsibility (E) addresses governmental environmental regulations and community ecological welfare; social responsibility (S) fulfills employee rights and consumer ethical consumption standards; and governance responsibility (G) upholds investor trust through transparent decision-making[15]. Collectively, these three dimensions are essential for companies to attain legitimacy and support from stakeholders. In the 1990s, Mitchell et al. proposed the theory of stakeholder salience. The categorization of stakeholders according to the three dimensions of power, legitimacy, and urgency elucidates the varied influences of distinct actors on corporate ESG practices: governments compel corporations to enhance their carbon performance via mandatory measures such as carbon quota regulations and carbon taxes; institutional investors leverage the power of "voting with their feet" based on ESG ratings; and consumers shape the market through eco-friendly purchasing behaviors. Consumers engage in eco-friendly purchasing behavior to provide market incentives. This multi-tiered stakeholder pressure network compels enterprises to transition their carbon management from passive compliance to proactive innovation.

In the 21st century, stakeholder theory has become intricately linked with the Sustainable Development Goals (SDGs), and initiatives like the UN's Principles for Responsible Investment (PRI) mandate that companies disclose carbon emissions data and permit stakeholder oversight, thereby transforming carbon performance into a fundamental metric of a company's responsiveness to stakeholders' climate expectations. The ESG framework functions as an implementation tool for stakeholder theory: companies mitigate carbon emissions in their production processes via investments in photovoltaic technology and carbon capture methods (environmental dimension), address regulatory pressures and consumer preferences for low-carbon products through supply chain carbon inventory and product carbon labeling (social dimension), and meet investor demands for ESG investments by establishing sustainability committees and enhancing carbon disclosure systems (governance dimension). This multi-dimensional interaction process converts external stakeholder pressure into internal incentive for carbon governance, ultimately enhancing carbon performance[16]. This theoretical framework elucidates the trajectory of ESG practices aimed at enhancing carbon performance—specifically, acquiring resources and legitimacy through the identification, coordination, and fulfillment of the climate demands of principal stakeholders—while also offering insights into the moderating influence of corporate financialization. When corporations excessively allocate resources to financial assets, they may jeopardize their investment in the low-carbon transformation of their core operations, resulting in diminished investment in both the low-carbon transformation of their business and their workforce. When companies excessively prioritize financial asset allocation, they may jeopardize their investment in the low-carbon transformation of their core operations, leading to deficiencies in stakeholder relations, including employee technical training and community collaboration on environmental protection, thereby diminishing the effectiveness of ESG in enhancing carbon performance.

2.4 Signaling Theory

The progression of signaling theory elucidates the underlying rationale of business behavioral strategies in information-asymmetric markets, and its significance to ESG practices and carbon performance fundamentally reflects the principles of information economics within sustainable development. The theory emerged from the labor market model introduced by Michael Spence in 1973, which posits that individuals with private information convey signals to the external environment through observable actions to mitigate information costs. This principle was later incorporated into capital structure analysis by Stephen Ross in the 1980s, who developed a framework for examining dividend policy and debt financing as indicators of corporate quality. In the late 1990s, the emergence of

socially responsible investing prompted a theoretical shift towards the value signaling function of non-financial disclosure. Companies communicated their dedication to sustainable development to the capital market by publishing environmental reports and revealing carbon footprint data. The efficacy of these signals was contingent upon their verifiability and cost differentiation: significant polluters who misleadingly promoted a low-carbon transition would be more impactful than firms that inaccurately claimed a low-carbon transition. The efficacy of this signal is contingent upon its verifiability and cost differentiation: companies with significant pollution will incur substantial regulatory penalties and reputational damage if they misrepresent a low-carbon transition, whereas firms that authentically adopt clean technology innovations will benefit from reduced financing costs by enhancing their ESG ratings[17]. This "split-equilibrium" mechanism has facilitated credible signaling of ESG performance to the capital market, enabling the assessment of a company's capacity to manage its carbon footprint. In the 21st century, signaling theory has been increasingly aligned with the climate change agenda, with carbon performance data (e.g., carbon emissions per unit of revenue, carbon-neutral roadmap) serving as a fundamental conduit for environmental signals. This data not only affects investors' evaluations of climate risk management capabilities but also yields significant financial returns through mechanisms such as premiums on green bond issuance and reductions in carbon tariffs. The governance aspect of the ESG framework, including the creation of a sustainability committee and third-party verification of carbon reports, markedly enhances the credibility of environmental signals by improving the standardization and transparency of information disclosure. Concurrently, stakeholder engagement in the social dimension, such as community environmental dialogues and supply-chain carbon emissions inspections, broadens the audience for these signals, enabling non-financial entities like consumers and suppliers to modify their cooperation strategies in alignment with ESG signals, thereby establishing a foundation for environmental policy formulation and the advancement of a green economy[18].

The ESG signals enable non-financial stakeholders, including customers and suppliers, to modify their collaboration strategies, so generating a multi-faceted incentive for enhancing carbon performance. Increased financialization may distort the signaling mechanism: an over-reliance on short-term gains from financial assets may compel management to reduce long-term ESG-related investments, resulting in "greenwash" noise in carbon performance signals. Additionally, the incentives for statement modification induced by financialization may lead to selective disclosure, such as exaggerating the progress of Scope 3emissions reductions while concealing the extent of those reductions. Simultaneously, financialization may result in selective disclosure (e.g., overstating advancements in Scope 3 emissions reduction while concealing Scope 1 emissions statistics), so undermining the relationship between ESG indicators and actual carbon performance[19]. The moderating effect indicates that financialization may diminish the substantive improvement of carbon performance through ESG practices due to resource crowding out, while also intensifying the capital market's misperception of the efficacy of low-carbon transition via signal distortion. This dual mechanism offers a theoretical framework for empirically investigating the moderating pathways of corporate financialization in the "ESG-Carbon Performance" relationship. This dual mechanism offers a theoretical framework for empirically investigating the moderating role of company financialization in the relationship between ESG and carbon performance.

2.5 New-Quality Productivity Theory

The new quality of productive forces constitutes a revolutionary theoretical framework within the context of socialist political economy with Chinese characteristics, embodying the qualitative advancement of productive forces, with its essence rooted in the dialectical unity of disruptive technological innovation and the comprehensive reconfiguration of production factors[20]. From a philosophical standpoint, it transcends the epistemological constraints of conventional linear progress in productive forces, instigating a quantitative reconfiguration of production factors through an intergenerational technological leap, and effecting a structural transition from

"factor-driven" to "paradigm-driven." This qualitative transformation is marked by the non-linear increase in total factor productivity, which reconfigures the principles of value creation within the triadic interplay of technological revolution, industrial transformation, and institutional innovation, thereby establishing advanced productive forces that signify a historical epoch.

The theoretical framework is anchored in the modern interpretation of Marxism's contradictory dynamics between productive forces and production relations. Its kinetic mechanism delineates a threefold breakthrough: firstly, a fundamental advancement in the technological foundation, evident in the reconfiguration of the production function through cutting-edge scientific and technological clusters; secondly, a revolutionary transformation of the factor structure, marked by the emergence of new factors such as data, knowledge, and intelligence, which alleviate the diminishing marginal returns of traditional factors; and thirdly, a systematic reconfiguration of industrial forms, facilitating the transition of the industrial system from mechanical to ecological division of labor, thereby elevating industrial production into an advanced productive force. The third aspect is the systematic reconfiguration of industrial structure, facilitating the transition of the industrial system from mechanical labor division to ecological synergy. This transformation engenders the synergistic evolution mechanism of "technology-industry-system," facilitating a transition in productive forces from quantitative accumulation to qualitative transformation, ultimately achieving a paradigm shift in the mode of production[21].

The structural dimension of the new quality productivity comprises three theoretical pillars: first, the paradigm shift in the driving force of innovation, characterized by a bidirectional feedback loop between advancements in fundamental research and the innovation of application scenarios; second, the time-space compression effect on resource allocation, wherein digital technology mitigates physical barriers to the flow of production factors; and third, the multidimensional leap in value creation, marked by the profound integration of sustainability, digitization, and human-centric values. Their development adheres to the principle of dialectical negation, encompassing the retention of the rational core of traditional productive forces while simultaneously discarding developmental constraints, ultimately leading to the establishment of the material foundation for a new form of human civilization[22]. The fundamental propositions that require elucidation in the present theoretical framework are: how the intergenerational transformation of productive forces, instigated by the technological revolution, can catalyze adaptive modifications in production relations; how the value creation logic of novel elements can reshape the paradigm of economic governance; and how the benefits of the socialist system can be converted into a systematic assurance for the enhanced quality of productive forces. This necessitates the development of a cohesive theoretical framework that connects the behavior of micro-entities, the progression of meso-industries, and the macro-institutional context, in order to elucidate the historical inevitability and practical trajectory of the qualitative transformation of productive forces from the standpoint of historical materialism.

Table 1. New Quality Productivity Index Calculation System

Factors	Subfactor	Indicators(ratio)	Description of indicator values	Weights
Labor force	Labor	R&D salaries	Research and development expenses - salaries and wages/operating income	28
		R&D staff	Number of R&D staff / Number of employees	4
		High-skilled workers	Number of people with bachelor's degree or above / Number of employees	3

Production tool				
	Materialized labor (Objects of labor)	Fixed assets	Fixed assets/total assets	2
		Manufacturing costs	(Subtotal cash outflows from operating activities + depreciation of fixed assets + amortization of intangible assets + provision for impairment - cash paid for purchases of goods and services - wages paid to and for employees) / (Subtotal cash outflows from operating activities + depreciation of fixed assets + amortization of intangible assets + provision for impairment)	1
	Hard technology	R&D depreciation and amortization	R&D expenses - depreciation and amortization/operating income	27
		R&D lease payments	Research and development expenses - rental expenses/operating income	2
	Soft technology	R&D direct investment	R&D expenses - direct inputs/operating income	28
		Intangible assets	Intangible assets/total assets	3
		Total asset turnover	Operating income/average total assets	1
		Inverse equity multiplier	Owners' equity/total assets	1
New-Quality productivity				100

3 RESEARCH DESIGN

3.1 Modeling

In order to test the research hypothesis H1, this paper first constructs the following regression model (1):

$$CP_{i,t} = \alpha_0 + \alpha_1 ESG_{i,t} + \sum Controls + \sum Firm + \sum Year + \varepsilon_{i,t} (1)$$

In model (1), CP is an explanatory variable, representing corporate carbon performance; ESG is an explanatory variable, representing the development of corporate social responsibility; Controls is a series of control variables; Firm and Year are fixed for company and year, respectively; ε is a random perturbation term; and the subscripts, i and t, stand for the individual firm and time, respectively;

In order to further explore the mechanism of ESG on carbon emission performance, based on the previous theoretical analysis, and according to the suggestion of the three-step approach to the testing of the transmission

mechanism, this paper constructs the transmission mechanism model (2) and (3) to test the hypotheses H2a and H2b;

$$NPRO_{i,t} = \beta_0 + \beta_1 ESG_{i,t} + \sum Controls + \sum Firm + \sum Year + \varepsilon_{i,t} (2)$$

$$CP_{i,t} = \gamma_0 + \gamma_1 ESG_{i,t} + \gamma_2 NPRO_{i,t} + \sum Controls + \sum Firm + \sum Year + \varepsilon_{i,t} (3)$$

In models (2) and (3), NPRO is the mediating variable, which indicates new quality productivity, and the rest of the variables have the same meaning as above;

In order to test hypothesis H3, corporate financialization is chosen as the moderating variable and model (4) is constructed

$$CP_{i,t} = \delta_0 + \delta_1 ESG_{i,t} + \delta_2 FINRATIO_{i,t} + \delta_3 ESG_{i,t} \times FINRATIO_{i,t} + \sum Controls + \sum Firm + \sum Year + \varepsilon_{i,t} (4)$$

In model (4), FINARTIO is the moderating variable that indicates the degree of financialization of the firm, $ESG_{i,t} \times FINRATIO_{i,t}$ represents the interaction term between ESG and the degree of financialization, and the rest of the variables have the same meaning as above.

Based on the aforementioned theoretical analysis and model design, the analysis makes the following assumptions:

H1: Good ESG performance can significantly improve a firm's carbon performance.

H2a: Effective ESG governance enhances firms' new quality productivity levels.

H2b: NQP plays a significant mediating effect between ESG performance and carbon emission performance.

H3: There is a significant positive moderating effect of corporate financialization in the impact of ESG performance on carbon emission performance.

3.2 Selection of variables

(1) Corporate carbon performance

In this paper, we refer Yu He [23] to define corporate carbon performance (CP) as the ratio of operating income (RMB 10,000) to total carbon emissions (tons), i.e., operating income per unit of carbon emissions. Among them, the total carbon emissions are calculated as combustion and fugitive emissions + production process emissions + waste emissions + emissions due to land use change (forest to industrial land).

(2) ESG rating data

Combining the relevant literature and the actual situation of the country, this paper chooses the ESG data of CSI as the core explanatory variables, and replaces the CSI ESG data with Bloomberg ESG data in the robustness test.

(3) New quality productivity

Referring to Jia Song [24] firms' new quality productivity measures for the construction of the indicator system. In addition, instrumental variables are added to this dataset for matching, and the chosen instrumental variable is total factor productivity of enterprises.

(4) Financialization of enterprises

The ratio of profits from financial channels such as investment income, gains and losses from changes in fair value and other comprehensive income of non-financial enterprises to operating profit is used as an indicator to measure the degree of financialization of enterprises. Profit from financial channels = (investment income + gain

from changes in fair value + loss from other comprehensive income), and the degree of financialization = profit from financial channels / operating profit.

(5) Control variables

Referring to previous studies on carbon performance, this paper selects the following control variables: firm size, firm age, gearing ratio, equity concentration, return on net assets, gross sales margin, cash ratio, and operating income growth rate, and the definitions of each variable are shown in the following table.

Table 2. Definition of variables

Variable type	Variable name	Variable Description
Explanatory variable	ESG performance (ESG)	CSI Disclosure Corporate ESG Score
Explained variable	Corporate Carbon Performance (CP)	Operating income per unit of carbon emissions
Intermediary variable	New Prime Productivity (NPRO)	Referring to Song Jia et al. (2024) enterprise new quality productivity measurement method for the construction of the indicator system, details are shown in Table
Moderator variable	Financialization of enterprises (FINRATIO)	(Investment income + Gain on fair value changes + Loss on other comprehensive income - Operating profit)/Operating profit
Control variable	Enterprise size (Size)	Natural logarithm of total assets for the year
	FirmAge	ln(current year - year of incorporation + 1)
	Gearing ratio (Lev)	Total liabilities at the end of the year / Total assets at the end of the year
	Shareholding Concentration (Top 10)	Number of shares held by top ten shareholders / Total number of shares
	Return on net assets (ROA)	Net profit / average balance of total assets
	Gross profit margin on sales (GProfit)	(Operating Revenue - Operating Costs) / Operating Revenue
	Cashflow	Net cash flows from operating activities / total assets
	Revenue growth rate (Growth)	Operating income for the current year / Operating income for the previous year - 1

3.3 Data sources and processing

Based on the content of this paper's research and the availability of data, this paper selects the financial report form data of A-share listed enterprises and the ESG rating data of CSI from 2015 to 2022, and performs the following processing on the relevant sample data: ① Eliminate the samples of ST and *ST listed enterprises with poor operation; ② Eliminate the samples with missing ESG indexes; ③ Eliminate samples of the financial industry and the real estate industry; ④ Perform Winsor2 tail reduction processing. The final sample size of 14,046 is obtained. Among them, ESG data are from Wind database, corporate carbon performance is calculated based on the collected

carbon emissions, new quality productivity is calculated using entropy method with reference to Song Jia's method, and the rest of control variables are from CSMAR database.

4 EMPIRICAL ANALYSIS

4.1 Descriptive statistics

The results of descriptive statistics show that the mean value of corporate carbon performance (CP) is 0.470, and the standard deviation is 0.529, indicating that the carbon performance of different enterprises varies greatly, and the carbon performance of some enterprises is significantly higher or lower than the mean value, with the minimum value of 0.00486 and the maximum value reaching 18.46, showing a wide distribution of extreme values.The mean value of ESG performance is 4.055, and the standard deviation is 1.120, with a minimum value of 1 and a maximum value of 7.750, reflecting significant differences in ESG performance. The mean value of firm size (Size) is 22.69 with a standard deviation of 1.386, and the minimum and maximum values are 17.64 and 28.64, respectively, indicating that the sample firms span a wide range of sizes.

Table 3. Descriptive statistics

Variables	Sample Size	Average Value	(Statistics) Standard Deviation	Upper Quartile	Minimum Value	Maximum Value
CP	14,046	0.470	0.529	0.529	0.005	18.460
ESG	14,046	4.055	1.120	1.120	1.000	7.750
Size	14,046	22.690	1.386	1.386	17.640	28.640
Lev	14,046	0.449	0.201	0.201	0.008	2.290
ROA	14,046	0.034	0.075	0.075	-1.130	1.285
GProfit	14,046	0.270	0.173	0.173	-0.862	0.964
Cashflow	14,046	0.052	0.074	0.074	-1.686	2.222
Growth	14,046	0.381	15.990	15.990	-0.965	1.878
Top10	14,046	55.420	15.120	15.120	1.310	101.200
FirmAge	14,046	3.036	0.275	0.275	1.792	4.025

4.2 Benchmark regression analysis

The results of the regression between ESG performance and corporate carbon performance are shown in the table, where column (1) does not include control variables, column (2) does not fix individual and time effects, and column (3) includes the first two. In the baseline regression analysis, the effect of ESG performance on corporate carbon performance (CP) is significantly positive, with a coefficient of 0.025 before the inclusion of control variables, which increases to 0.031 after the inclusion of control variables, and both are significant at the 1% level, indicating that the enhancement of ESG performance of enterprises can significantly improve their carbon performance. The coefficients of firm size (Size) range from 0.024 to 0.046 and are significant at the 1% or 5% level, suggesting that larger firms usually have better carbon performance, possibly due to their resource and technology advantages. The coefficient of Return on Assets (ROA) is significantly positive in some models, suggesting that more profitable firms may invest more resources in carbon reduction.

Table 4. Benchmark regression analysis

	(3)
	CP
ESG	0.031***
	(0.001)
Size	0.046**
	(0.033)
Lev	0.061
	(0.366)
ROA	0.484***
	(0.000)
GProfit	-0.738***
	(0.000)
Cashflow	-0.142
	(0.152)
Growth	0.000**
	(0.042)
Top10	-0.003**
	(0.020)
FirmAge	1.653***
	(0.000)
Firm	Yes
Year	Yes
_cons	-5.388***
	(0.000)
N	14046
adj. R^2	0.191

Note: ***, ** and * indicate significant at the 1%, 5% and 10% levels, respectively, estimated using firm-level clustering robust standard errors with p-values in parentheses, below.

4.3 Robustness Tests

In the robustness test, the effect of ESG performance on corporate carbon performance (CP) remains significantly positive and both are significant at the 1% level, further validating the reliability of the conclusion that ESG performance has a positive effect on carbon performance. Column (1) incorporates the use of city-level clustering standard errors into the main regression to further ensure the reliability of the results. Column (2) uses Bloomberg ESG data (ESG_Bloomberg) as an alternative core explanatory variable, at which point the coefficient coefficient is 0.007, also significant at the 1% level, indicating consistency of findings across ESG data sources. The coefficient of column (3) lagged one-period ESG performance (L.ESG) is 0.144 and is significant at the 1% level, indicating that the impact of ESG performance on carbon performance is persistent, and that firms' long-term ESG investments lead to sustained carbon performance improvement. The introduction of control variables (CV) and firm fixed effects (Firm) and year fixed effects (Year) further enhances the robustness of the model. Overall, the results of the robustness test further consolidate the conclusions of the benchmark regression analysis, indicating that the positive impact of ESG performance on carbon performance is robust across different data sources, time

horizons, and model settings, which provides a strong empirical support for firms to achieve carbon emission reduction and sustainable development by improving ESG performance.

Table 5. Robustness test

	(1) CP	(1) CP	(1) CP
ESG	0.080 ***		
	(0.000)		
ESG_Bloomberg		0.007***	
		(0.000)	
L.ESG			0.144***
			(0.009)
CV	Yes	Yes	Yes
Firm	Yes	Yes	Yes
Year	Yes	Yes	Yes
City	Yes	No	No
Constant	-43.001 ***	-1.806 ***	44.024 ***
	(0.000)	(0.000)	(0.000)
N	4199	14046	14046
adj. R^2	0.474	0.136	0.078

Note: Column (1) uses standard errors for city-level clustering.

4.4 Tests for mediating effects

In the mediation effect test, the direct effect of ESG performance on corporate carbon performance (CP) is significantly positive, with a coefficient of 0.031 and significant at the 1% level, indicating that ESG performance can directly enhance corporate carbon performance. Meanwhile, the effect of ESG performance on new quality productivity (NPRO) is also significantly positive, with a coefficient of 0.038, and significant at the 5% level, indicating that ESG performance can indirectly improve carbon performance by enhancing the productivity and technological innovation capacity of enterprises. The direct effect of new quality productivity (NPRO) on carbon performance is also significantly positive, with a coefficient of 0.017 and significant at the 1% level, further validating the mediating role of NPRO between ESG performance and carbon performance. The results of the mediation effect test indicate that ESG performance not only directly improves carbon performance, but also indirectly improves carbon performance by promoting NPRO.

Table 6. Mediating effects test

	(2) NPRO	(3) CP
ESG	0.038**	0.031 ***
	(0.050)	(0.000)
NPRO		0.017 ***
		(0.003)

Size	0.198***	0.047**
	(0.000)	(0.030)
Lev	-1.154***	0.061
	(0.000)	(0.365)
ROA	-2.478***	0.507***
	(0.000)	(0.000)
GrossProfit	0.093	-0.733***
	(0.470)	(0.000)
Cashflow	2.552***	-0.156
	(0.000)	(0.123)
Growth	-0.001	0.000**
	(0.241)	(0.040)
Top10	-0.003**	-0.003**
	(0.035)	(0.018)
FirmAge	-0.560***	1.643***
	(0.000)	(0.000)
_cons	3.006***	-5.463***
	(0.000)	(0.000)
Firm/Year	Yes	Yes
N	14046	14046

4.5 Moderating effects test

In the moderating effect test, the direct effect of ESG performance on corporate carbon performance (CP) is significantly positive (with coefficients ranging from 0.030 to 0.040, all significant at the 1% level), suggesting that an increase in corporate ESG performance can directly improve carbon performance. The moderating effect of the degree of corporate financialization (FINRATIO) is reflected by the interaction term (ESG × FINRATIO), whose coefficients range from 0.015 to 0.018 and are significant at the 1% level, suggesting that the higher the degree of corporate financialization, the stronger the contribution of ESG performance to carbon performance. This result suggests that firms with a high degree of financialization may obtain more resources (e.g., investment income or cash flow) through financial channels, thus more effectively transforming ESG inputs into practical actions for carbon reduction. In addition, the coefficient of the financialization variable (FINRATIO) itself is significantly positive (0.020 to 0.024), suggesting that moderate financialization may indirectly support carbon performance enhancement by enhancing firms' financial flexibility. The moderating effect test reveals the positive and reinforcing role of corporate financialization in the relationship between ESG and carbon performance, suggesting that financialization not only directly supports carbon performance, but also amplifies the positive impact of ESG. The implication for corporate managers is that when promoting ESG strategies, they need to make reasonable use of financialization tools to optimize resource allocation, but they need to be wary of the risk of over-financialization that may lead to de-realization, so as to achieve a balance between environmental benefits and financial health.

Table 7. Moderating effects test

	(1)	(2)
	CP	CP
ESG	0.030***	0.040***
	(0.000)	(0.000)

FINRATIO	0.020***	0.024***
	(0.000)	(0.000)
ESG x FINRATIO		0.018***
		(0.000)
Size	0.044***	0.040***
	(0.014)	(0.004)
Lev	-0.017	0.018
	(0.055)	(0.744)
ROA	0.327***	0.432***
	(0.073)	(0.000)
GrossProfit	-0.558***	-0.560***
	(0.088)	(0.000)
Cashflow	-0.138*	-0.125
	(0.080)	(0.125)
Growth	0.000***	0.000***
	(0.005)	(0.005)
Top10	-0.004***	-0.004***
	(0.000)	(0.000)
FirmAge	1.552***	1.569***
	(0.041)	(0.000)
_cons	0.351***	-4.990***
	(0.000)	(0.000)
Firm/Year	Yes	Yes
N	13694	13694
adj. R^2	0.186	0.206

4.6 Heterogeneity test

The results of the heterogeneity test indicate that the effect of corporate ESG performance on carbon performance is significantly different among different types of enterprises, industries and regions. Specifically, ESG enhancement in non-state-owned enterprises has a significant contribution to carbon performance (β=0.062, p<0.01), while the role of ESG in state-owned enterprises fails the test of significance (β=-0.023, p=0.297), which may originate from the fact that state-owned enterprises are bound by the stronger administrative objectives, and the market transformation efficiency of ESG inputs is lower; enterprises in heavy pollution industries ESG performance is not significantly associated with carbon performance (β=0.001, p=0.963), while the carbon reduction effect of ESG in non-heavily polluted industries is significant (β=0.048, p<0.01), implying that the marginal utility of ESG in highly polluted industries is weakened by the technology path dependence or the squeeze on environmental compliance costs; looking at the sub-industries, the ESG of manufacturing enterprises has a significant effect on the improvement of carbon performance (β=0.054, p<0.01), but non-manufacturing enterprises show a negative effect (β=-0.062, p<0.05), which may be related to the fact that non-manufacturing industries (e.g., services) lack physical emission reduction grips, and the linkage mechanism between ESG inputs and carbon performance has not yet been matured; at the regional level, the carbon performance gain of ESG in the eastern region is significant (β=0.041, p& lt;0.01), while the effect is not significant in the central and western regions, reflecting that the more perfect

market mechanism and green financial support system in developed regions amplify the environmental benefits of ESG.

This result suggests that policy design needs to be stratified: ESG incentives can be strengthened for non-state-owned enterprises, manufacturing industries, and the eastern region, while state-owned enterprises, heavily polluting industries, and the central and western regions need to be supported by technology subsidies or differentiated assessment standards to bridge the structural gap in ESG practices, and to promote the non-manufacturing industries to explore synergistic paths between digital transformation and ESG.

Table 8. Heterogeneous groupings1

	CP			
	Nationalized business	Non-state enterprise	Service industry	Non-manufacturing industry
ESG	-0.023	0.062***	0.054***	-0.062**
	(0.297)	(0.000)	(0.000)	(0.027)
Controls	Yes	Yes	Yes	Yes
_cons	-8.203***	-3.340***	-4.998***	-7.565***
	(0.000)	(0.000)	(0.000)	(0.000)
N	5908	7776	10951	3095
adj. R^2	0.164	0.237	0.214	0.154

Table 9. Heterogeneous groupings2

	CP	
	Heavily polluting enterprises	Non-heavily polluting enterprises
ESG	0.001	0.048***
	(0.963)	(0.000)
Controls	Yes	Yes
_cons	-5.187***	-5.393***
	(0.000)	(0.000)
N	4087	9959
adj. R^2	0.227	0.181

Table 10. Heterogeneous groupings3

	CP		
	Eastern part	Western part	Central section
ESG	0.041***	0.008	0.003
	(0.000)	(0.688)	(0.870)
Controls	Yes	Yes	Yes
_cons	-5.121***	-4.388***	-7.500***
	(0.000)	(0.000)	(0.000)
N	9415	2552	2079
adj. R^2	0.185	0.226	0.191

5 CONCLUSIONS AND POLICY RECOMMENDATIONS

This paper demonstrates through empirical analysis that corporate ESG performance significantly enhances carbon performance improvement via multiple pathways. The ESG framework actively encourages the implementation of low-carbon technologies and enhances resource efficiency in enterprises by integrating environmental accountability, social collaboration, and governance enhancement, exemplified by the closed loop of emission reduction via investments in clean energy and management of supply chain carbon footprints. New quality productivity serves as a crucial intermediary between the two, while technological innovation and factor reorganization (e.g., industrial internet, circular economy model) substantially diminish carbon emission intensity per unit of output, thereby validating the transmission logic of "ESG inputs - new quality productivity leap - optimization of carbon performance." The transmission logic is confirmed. Simultaneously, the financialization of businesses enhances the environmental advantages of ESG via green financial instruments and oversight of capital markets. Heterogeneity analysis indicates that ESG significantly improves carbon performance, particularly in non-state-owned enterprises, manufacturing sectors, and eastern enterprises, highlighting the synergistic effects of market-oriented mechanisms, industrial traits, and regional resource endowments. This indicates that ESG functions not merely as a compliance instrument, but also as a systematic innovation catalyst for facilitating low-carbon transition, offering empirical proof that enhanced productivity can aid in achieving the "dual-carbon" objective.

The empirical evidence necessitates policy interventions calibrated to enterprise heterogeneity and China's institutional realities. To address the systemic 'discrimination in access to financing' against non-SOEs revealed in our study, provincial governments should collaborate with the People's Bank of China (PBOC) to implement a dual-track financing mechanism, integrating mandatory quotas with risk mitigation tools. First, introduce *ESG Credit Allocation Mandates* requiring regional commercial banks to dedicate no less than 15% of their annual green loan portfolios to non-SOEs with accredited BBB+ or higher ESG ratings (certified by PBOC-recognized agencies like Sino-Securities ESG). Second, establish provincial *Green Bond Guarantee Pools*—modeled on Zhejiang's 2023 pilot—where governments cover 50% of underwriting risks for qualified private enterprises issuing transition-themed green bonds, directly aligning with Category B ("Transition Activities") of PBOC's *Green Bond Endorsed Projects Catalogue (2021)*. Concurrently, the China Banking and Insurance Regulatory Commission (CBIRC) should adjust risk-weighting factors for ESG loans to non-SOEs (e.g., reducing capital reserve requirements by 0.5x for BBB+-rated exposures), incentivizing banks to overcome collateral-driven lending biases. To ensure scalability, this system should integrate with the national *Corporate ESG Database* launched by the Ministry of Ecology and Environment in 2024, enabling automated eligibility verification and reducing compliance costs by ~30% based on Suzhou Industrial Park's sandbox testing.

Capitalizing on the unique ecological endowment of central/western provinces, we propose a "Carbon-Plus Ecosystem" framework that converts natural capital into tangible compliance benefits and market advantages. Core to this approach is establishing provincial-level Eco-Carbon Trading Platforms in renewable-rich regions (e.g., Gansu's wind corridors, Sichuan's hydropower basins), where enterprises can offset up to 30% of mandatory carbon quotas by investing in verified ecological projects—including grassland restoration, desert greening, or biodiversity conservation — with carbon sequestration volumes quantified via the National Forestry and Grassland Administration's *Methodology for Forest Carbon Sink Accounting (2023)*. To maximize policy synergy, these offsets should be automatically convertible into tradable China Certified Emission Reductions (CCERs) under the national carbon market, while provincial governments provide additional fiscal incentives such as land-use priority rights for participating firms (e.g., 20% faster approval for industrial land in Inner Mongolia's zero-carbon zones) and value-added tax rebates equivalent to 15% of eco-investment costs (anchored in Article 9 of the State Council's *Western*

Development Revitalization Plan 2021-2030). Crucially, this mechanism embeds a dynamic calibration feature where the offset cap (initially 30%) adjusts annually based on regional carbon intensity reduction targets from the Ministry of Ecology and Environment's *Provincial Carbon Budget Allocation Scheme*, preventing market distortion. Implementation-wise, the National Development and Reform Commission (NDRC) should integrate this platform with its "Ecological Asset Voucher" pilot in Yunnan/Gansu, using blockchain for real-time auditing of carbon-ecosystem equivalence.

To resolve the critical data gap in service-sector emissions—where current national carbon inventories cover <15% of digital economy actors—we advocate a two-pronged strategy combining mandatory granular accounting standards with shared digital infrastructure. Immediate priority should be given to formulating industry-tailored carbon footprint guidelines under ISO 14064 framework within 24 months, specifically targeting cloud computing, fintech, and logistics sectors, with compliance mandated for enterprises exceeding RMB 1 billion annual revenue (covering ~80% of sector emissions). These standards must enforce Scope 3+ measurement requiring hyperscalers like Alibaba Cloud and Tencent Cloud to disclose embedded emissions of downstream user activities — e.g., calculating per-API-call carbon intensity via standardized PUE-to-CO_2 conversion factors certified by the National Institute of Metrology. Concurrently, a National Green Cloud Audit Platform should be launched under the Ministry of Industry and Information Technology (MIIT), consolidating real-time energy data from all tier-IV+ data centers and automating verification through AI-powered carbon ledger systems (validated by Shenzhen's pilot reducing reporting errors by 40%). To incentivize adoption, a 15% corporate income tax rebate should be granted to firms deploying MIIT-certified AI energy optimization tools—mirroring Tencent's Net Zero Accelerator model—with eligibility conditional on achieving >20% year-on-year carbon productivity gains as measured against the platform's benchmarks. Crucially, this infrastructure must interlink with the NDRC's "Eastern Data Western Computing" project, directing low-carbon data flows to renewable hubs (e.g., Guizhou's hydro-powered clusters) while enforcing PUE $\leqslant$1.25 thresholds through the Revamped Data Center Grading Scheme (2025).

Data Availability Statement

Data are available in specialized financial databases, processed by the authors.

References

[1] Luo, L., & Tang, Q. (2023). The real effects of ESG reporting and GRI standards on carbon mitigation: International evidence. Business Strategy and the Environment, 32(6), 2985–3000.

[2] Cong Y, Zhu C, Hou Y, Tian S and Cai X (2022) Does ESG investment reduce carbon emissions in China?. *Front. Environ. Sci.* 10:977049. doi: 10.3389/fenvs.2022.977049

[3] Chu, H., Niu, X., Li, M., & Wei, L. (2025). Research on the impact of new quality productivity on enterprise ESG performance. International Review of Economics & Finance, 99, 104009. https://doi.org/10.1016/j.iref.2025.104009

[4] Xie, Y. (2024). The impact of ESG performance on corporate sustainable growth from the perspective of carbon sentiment. Journal of Environmental Management, 367, 121913. https://doi.org/10.1016/j.jenvman.2024.121913

[5] Lu, J., & Li, H. (2024). The impact of ESG ratings on low carbon investment: Evidence from renewable energy companies. Renewable Energy, 223, 119984. https://doi.org/10.1016/j.renene.2024.119984

[6] Geng, Y., Zheng, Z., Yuan, X., & Jiménez-Zarco, A. I. (2025). ESG performance and total factor productivity of enterprises: The role of digitization. Research in International Business and Finance, 77, 102920.

[7] Xue, Q., Jin, Y., & Zhang, C. (2024). ESG rating results and corporate total factor productivity. International Review of Financial Analysis, 95, 103381. https://doi.org/10.1016/j.irfa.2024.103381

[8] Huang, R., Zhu, Z., Ruan, R., & Lou, X. (2024). Linking low-carbon practices with ESG performances: Exploration evidence from the configurational perspective. Journal of Cleaner Production, 435, 140532.

[9] Lee, C. L., & Liang, J. (2024). The effect of carbon regulation initiatives on corporate ESG performance in real estate sector: International evidence. Journal of Cleaner Production, 453, 142188. https://doi.org/10.1016/j.jclepro.2024.142188

[10] Yuan, Y., Dai, H., & Ma, J. (2025). The Impact of Corporate ESG Performance on Supply Chain Resilience: A Mediation Analysis Based on New Quality Productive Forces. *Sustainability*, *17*(10), 4418. https://doi.org/10.3390/su17104418

[11] Parris, T. M., & Kates, R. W. (2003). CHARACTERIZING AND MEASURING SUSTAINABLE DEVELOPMENT. Annual Review of Environment and Resources, 28(Volume 28, 2003), 559-586. https://doi.org/https://doi.org/10.1146/annurev.energy.28.050302.105551

[12] Springett, D. (2003). Business conceptions of sustainable development: A perspective from critical theory. Business Strategy and the Environment, 12(2), 71–86. https://doi.org/10.1002/bse.353

[13] Lindgreen, A. and Swaen, V. (2010), Corporate Social Responsibility. International Journal of Management Reviews, 12: 1-7. https://doi.org/10.1111/j.1468-2370.2009.00277.x

[14] Lee, M.-D.P. (2008), A review of the theories of corporate social responsibility: Its evolutionary path and the road ahead. International Journal of Management Reviews, 10: 53-73. https://doi.org/10.1111/j.1468-2370.2007.00226.x

[15] Chan, E. Y. (2025). Moral Signaling in Startups: How ESG Claims Shape Stakeholder Judgments and Ethical Legitimacy. Journal of Business Ethics. https://doi.org/10.1007/s10551-025-06039-0

[16] Talan, G., Sharma, G. D., Pereira, V., & Muschert, G. W. (2024). From ESG to holistic value addition: Rethinking sustainable investment from the lens of stakeholder theory. International Review of Economics & Finance, 96, 103530. https://doi.org/10.1016/j.iref.2024.103530

[17] Bafera, J., & Kleinert, S. (2022). Signaling Theory in Entrepreneurship Research: A Systematic Review and Research Agenda. *Entrepreneurship Theory and Practice*, *47*(6), 2419-2464. https://doi.org/10.1177/10422587221138489

[18] Fu, L., Boehe, D. M., & Orlitzky, M. O. (2021). Broad or Narrow Stakeholder Management? A Signaling Theory Per

spective. *Business & Society*, *61*(7), 1838-1880. https://doi.org/10.1177/00076503211053018

[19] Lee, M. T., Raschke, R. L., & Krishen, A. S. (2022). Signaling green! Firm ESG signals in an interconnected environment that promote brand valuation. Journal of Business Research, 138, 1–11. https://doi.org/10.1016/j.jbusres.2021.08.061

[20] Fan Gao. (2023). The logic behind the proposal of “new quality productivity,” its multidimensional connotations, and its significance for the times. Review of Political Economy., 14 (06), 127-145.

[21] Wen Zhou & Lingyun Xu. (2023). On New Quality Productivity: Its Characteristics and Key Focus Areas. Reform, (10), 1-13.

[22] Zheng Xu, Linhao Zheng & Mengyao Cheng. (2023). The intrinsic logic and practical concepts of new quality productivity empowering high-quality development. Contemporary Economic Research, (11), 51-58.

[23] Yu He, Qingliang Tang & Kaitian Wang. (2017). Carbon Performance and Financial Performance. Accounting Research, (02), 76-82+97.

[24] Jia Song, Jinchang Zhang & Yi Pan. (2024). The Impact of ESG Development on Corporate New Quality Productivity: Empirical Evidence from Chinese A-Share Listed Companies. Contemporary Economic Management.*, 46* (06), 1-11. doi:10.13253/j.cnki.ddjjgl.2024.06.001.

Global Academic Frontiers

Volume 3 • Issue 3 • September 2025

ISSN 2995-5688

09 >

9 772995 568254

Free Copy