

Volume 7, Issue 5, 2025

Print ISSN: 2663-1938

Online ISSN: 2663-1946

JOURNAL OF COMPUTER SCIENCE AND ELECTRICAL ENGINEERING



Copyright© Upubscience Publisher

Journal of Computer Science and Electrical Engineering

Volume 7, Issue 5, 2025



Published by Upubscience Publisher

Copyright© The Authors

Upubscience Publisher adheres to the principles of Creative Commons, meaning that we do not claim copyright of the work we publish. We only ask people using one of our publications to respect the integrity of the work and to refer to the original location, title and author(s).

Copyright on any article is retained by the author(s) under the Creative Commons

Attribution license, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Authors grant us a license to publish the article and identify us as the original publisher.

Authors also grant any third party the right to use, distribute and reproduce the article in any medium, provided the original work is properly cited.

World Journal of Engineering Research**Print ISSN: 2663-1938 Online ISSN: 2663-1946****Email: info@upubscience.com****Website: <http://www.upubscience.com/>**

Table of Content

BEYOND EXTERNAL CONTROL: HYPERNETWORK-DRIVEN PARAMETER EDITING FOR MULTI-MODAL IMAGE GENERATION Hao Chen	1-12
DESIGN AND IMPLEMENTATION OF AN INTELLIGENT RECOMMENDATION SYSTEM FOR COLLEGE ENTRANCE EXAMINATION APPLICATION PREFERENCES HongFu Zeng, ZhaoMin Liang*, FanRui Wei, YiXun Lu	13-25
CURRENT STATUS AND DEVELOPMENT TRENDS OF SATELLITE IOT Hao Qi*, ZaoXia Ma	26-33
TEACHING METHOD REFORM OF PYTHON PROGRAMMING COURSE EMPOWERED BY AI TECHNOLOGY Yi Han	34-39
A RETRIEVAL-AUGMENTED GENERATION (RAG)-BASED INTELLIGENT REVIEWER ASSIGNMENT SYSTEM FOR SCIENTIFIC PROJECT EVALUATION JiTao Ma*, HongWei Huang, Jun Du	40-45
THE PATH OF ARTIFICIAL INTELLIGENCE EMPOWERING CAREER PLANNING EDUCATION FOR LOCAL COLLEGE STUDENTS IN CHINA ZhiXing Hu*, ZhenYu Yin, YuFei Zhou, Kan Wu*	46-50
FACE RECOGNITION MODEL BASED ON VISION TRANSFORMER JiaChen Gao	51-58
OPTIMIZING BLOCK-BY-BLOCK RELOCATION IN HISTORIC URBAN RENEWAL: A HYBRID SIMULATED ANNEALING-GENETIC ALGORITHM APPROACH MingYu Wang*, ZiHeng Ji	59-64
PARALLEL GENETIC ALGORITHM FOR MAGNETOTELLURIC INVERSION WITH GPU You Miao, Ge Cheng*	65-69
HUMAN-AI CO-CREATION SYSTEM FOR KNOWLEDGE WORK BASED ON MULTI-AGENT APPROACH WeiJing Zhu, RunTao Ren*, Wei Xie, CenYing Yang	70-79

BEYOND EXTERNAL CONTROL: HYPERNETWORK-DRIVEN PARAMETER EDITING FOR MULTI-MODAL IMAGE GENERATION

Hao Chen

Queen Mary School Hainan, Beijing University of Posts and Telecommunications, Beijing 100876, China.

Corresponding Email: HaoChenn.Eric@gmail.com

Abstract: Current controllable image generation methods predominantly rely on external architectural modifications, such as auxiliary control networks, which require substantial computational overhead and struggle to unify diverse control modalities including text, pose, depth, and sketches. These approaches fundamentally limit scalability and real-time applicability due to their additive nature and complex multi-condition integration challenges. We introduce HyperEdit, a novel hypernetwork-driven framework that achieves multi-modal controllable generation through dynamic parameter perturbation of pre-trained diffusion models, moving beyond external control paradigms toward intrinsic model adaptation. Our approach employs a unified hypernetwork that learns to map diverse control conditions—ranging from textual descriptions and pose skeletons to depth maps and edge sketches—into targeted parameter perturbations, enabling seamless integration of multiple modalities without architectural modifications to the base model. Through systematic perturbation discovery on carefully constructed condition-image pairs and progressive parameter injection strategies, HyperEdit demonstrates remarkable efficiency gains, achieving up to $6\times$ faster inference compared to existing methods while requiring significantly fewer parameters. Extensive experiments across diverse control scenarios show that our unified framework not only maintains generation quality comparable to specialized control methods but also enables novel capabilities such as real-time condition mixing, dynamic editing strength adjustment, and reversible modifications. This work establishes a new paradigm for controllable generation that bridges the gap between research innovation and practical deployment requirements.

Keywords: Model editing; Image generation; Hypernetwork

1 INTRODUCTION

The democratization of high-quality image generation has fundamentally transformed creative workflows, enabling users across diverse domains to produce sophisticated visual content through intuitive control interfaces [1-2]. Modern applications increasingly demand nuanced control capabilities that extend far beyond simple text prompts—digital artists require precise pose manipulation while maintaining aesthetic coherence, product designers need simultaneous control over object geometry and material properties, and content creators seek to blend multiple stylistic elements in real-time interactive sessions [3-4]. This evolution toward multi-modal controllable generation represents both a tremendous opportunity and a significant technical challenge for current generative frameworks.

Existing approaches to controllable image generation have predominantly adopted external architectural modifications, exemplified by influential works such as ControlNet [5], T2I-Adapter [6], and IP-Adapter [7]. These methods introduce auxiliary control networks that process conditioning inputs—ranging from edge maps and depth information to pose skeletons and reference images—and inject control signals into pre-trained diffusion models through carefully designed coupling mechanisms [8-10]. While these external control paradigms have demonstrated remarkable success in specialized scenarios, they suffer from fundamental limitations that increasingly constrain their practical deployment [3,11]. First, computational overhead scales linearly with the number of control modalities, as each condition type typically requires dedicated processing networks and specialized attention mechanisms [5-6]. Second, integrating multiple heterogeneous conditions remains challenging, often requiring complex weight balancing strategies and hand-tuned fusion protocols that lack principled theoretical foundations [4-10]. Third, the external nature of these modifications limits runtime flexibility, making dynamic condition adjustment, real-time editing strength modulation, and reversible modifications computationally prohibitive for interactive applications [7].

These limitations become particularly pronounced when users require sophisticated multi-modal control scenarios that reflect real-world creative needs [12]. Consider a digital artist who wants to generate an image where a character adopts a specific pose (skeleton control), maintains a cheerful facial expression (text guidance), follows a particular depth composition (depth map control), and adheres to a vintage aesthetic style (reference image guidance). Current external control methods would need to coordinate four separate processing pipelines, manage complex inter-modal interactions, and perform computationally expensive attention reweighting at every denoising step—resulting in significant latency that breaks the creative flow and limits practical usability [3-4].

We propose a fundamentally different approach that moves beyond external control paradigms toward intrinsic model adaptation through dynamic parameter perturbation [13-14]. Our method, HyperEdit, employs a unified hypernetwork that learns to map diverse control conditions directly into targeted parameter adjustments of the pre-trained diffusion

model itself. Rather than adding external computational overhead, this approach modifies the internal behavior of the generative model through carefully learned parameter perturbations, enabling seamless multi-modal control while maintaining the efficiency and architectural integrity of the original diffusion framework.

The core insight driving our approach is that different types of visual control—whether pose manipulation, style transfer, or geometric adjustment—can be effectively achieved through specific patterns of parameter modification within the diffusion model's existing architecture [6,15]. By systematically discovering these parameter-to-effect mappings through carefully constructed condition-image pairs and training a hypernetwork to predict appropriate perturbations for arbitrary control combinations, we establish a unified framework that naturally handles heterogeneous control modalities without requiring specialized fusion mechanisms or architectural modifications [7].

Our hypernetwork-driven parameter editing strategy offers several distinct advantages over external control methods. During inference, generating control perturbations requires only a single forward pass through the lightweight hypernetwork, after which the modified diffusion model operates at its original computational cost. This design enables real-time condition mixing, where users can dynamically adjust the strength of different control modalities, combine previously unseen condition types, and even reverse modifications by subtracting the applied perturbations—capabilities that are difficult or impossible to achieve with external control architectures.

Through systematic evaluation across diverse control scenarios and comprehensive comparison with state-of-the-art methods, we demonstrate that HyperEdit achieves comparable generation quality while requiring significantly fewer computational resources. Our unified framework maintains consistent performance across single-condition and complex multi-condition scenarios, validates the effectiveness of our parameter perturbation strategy, and establishes new benchmarks for efficiency in controllable generation tasks.

The main contributions of this work are threefold:

1. **Paradigm Innovation:** We introduce a novel hypernetwork-driven parameter editing framework that fundamentally shifts controllable generation from external architectural modifications to intrinsic model adaptation, enabling unified multi-modal control through dynamic parameter perturbation while preserving the computational efficiency of pre-trained diffusion models.
2. **Technical Framework:** We develop a systematic approach for discovering and learning parameter-to-effect mappings across diverse control modalities, including a multi-modal condition encoder that handles heterogeneous inputs (text, pose skeletons, depth maps, edge sketches), a progressive parameter injection strategy that ensures stable modifications, and a unified hypernetwork architecture that generates targeted perturbations for arbitrary condition combinations.
3. **Empirical Validation:** We conduct comprehensive experiments demonstrating that our approach achieves up to up to $6\times$ inference speedup compared to existing multi-modal control methods while maintaining generation quality, enables novel capabilities such as real-time condition mixing and reversible editing that are challenging for external control paradigms, and provides robust performance across diverse control scenarios ranging from single-condition manipulation to complex multi-modal compositions.

2 RELATED WORK

2.1 Controllable Image Generation Methods

Controllable image generation has consistently been a core research direction in computer vision. While early Generative Adversarial Networks achieved breakthroughs in image quality, they exhibited significant limitations in controllability [1]. With the rise of diffusion models, researchers began exploring how to achieve precise conditional control based on pre-trained diffusion models.

External control paradigms represent the current mainstream solution. ControlNet [5] pioneered the use of external control networks to achieve spatial conditional control, with the core idea of duplicating the encoder part of pre-trained diffusion models and gradually learning control signal injection through zero convolution layers. This method supports multiple control conditions including edges, depth, and pose, achieving precise control while maintaining the original model's generative capabilities. T2I-Adapter [6] adopted a similar but more lightweight design, learning simple adapter networks to align external control signals with internal knowledge, offering fewer parameters and faster training speed compared to ControlNet.

To address ControlNet's limitations in complex scenarios, subsequent research proposed multiple improvements. ControlNet++ [8] introduced a consistency feedback mechanism, significantly improving control precision through pixel-level cyclic consistency optimization. ControlNet-XS [9] re-examined the control process from a feedback control system perspective, proposing high-frequency, large-bandwidth communication mechanisms that enhance control effectiveness while reducing model size. DC-ControlNet [10] specifically targets multi-element control scenarios, achieving more flexible multi-condition fusion through separation of intra-element and inter-element conditional control.

Image prompt control has also emerged as an important research direction. IP-Adapter [7] processes text features and image features separately through a decoupled cross-attention mechanism, achieving performance comparable to full fine-tuning with only 22M parameters. This method not only supports image prompts but can also work collaboratively with text prompts for multi-modal generation, maintaining complete compatibility with existing control tools.

Chinese researchers have also made significant contributions to this field. The conditional image generation survey released by Zhejiang University and other institutions [3] systematically summarized 258 related papers, providing in-

depth analysis of existing methods from the perspective of conditional embedding. The survey pointed out that the core of existing methods lies in how to embed user conditions into denoising networks and sampling processes, proposing conditional embedding strategies for different tasks. Domestic scholars' survey on multi-modal controllable diffusion models further emphasized the importance of multi-dimensional control including semantic control [4], spatial position control, and ID control.

2.2 Parameter-Efficient Fine-tuning Methods

Parameter-Efficient Fine-Tuning (PEFT) methods play an increasingly important role in large model adaptation. LoRA (Low-Rank Adaptation) [16] represents the most prominent work in this area, reducing GPT-3's trainable parameters by 10,000 times while maintaining performance comparable to full fine-tuning by injecting trainable low-rank decomposition matrices into each Transformer layer.

LoRA's core idea is based on the hypothesis that weight updates during model adaptation possess low intrinsic dimensionality [5]. By decomposing weight updates ΔW into the product of two low-rank matrices A and B, LoRA dramatically reduces the number of parameters requiring training. This concept was later extended to various architectures, including controlled generation tasks in diffusion models [17].

Chinese researchers have also conducted in-depth exploration of parameter-efficient fine-tuning methods. Research from Tsinghua University and other institutions demonstrated that LoRA-type methods can maintain model performance while significantly reducing computational resources in Chinese large language model fine-tuning. Industrial practices by companies like Huawei further validated the effectiveness of these methods in industrial-scale applications [18].

2.3 Hypernetwork Methods

HyperNetworks provide a novel parameter generation paradigm [13], training a small network to generate weights for another large network. This idea can be traced back to neuroevolution fields, but Ha et al. first successfully applied it to deep learning, demonstrating hypernetworks' potential in recurrent neural network weight generation [13].

Regarding the theoretical foundation of parameter generation, "Generating Neural Networks with Neural Networks" further developed hypernetwork theory [10], proposing balanced objectives between accuracy and diversity and introducing a variational inference framework. This work emphasized that generated network diversity should consider network symmetric transformations, providing important theoretical guidance for subsequent hypernetwork design.

Reinforcement learning applications in model editing represent the latest development in hypernetwork methods. RLEdit combines hypernetwork-based lifelong editing with reinforcement learning modeling [15], solving the incompatibility issues of traditional hypernetwork methods during dynamic parameter changes by treating editing losses as rewards and optimizing hypernetwork parameters at the complete knowledge sequence level. This work demonstrates that through reinforcement learning paradigms, hypernetworks can more precisely capture model changes and generate appropriate parameter updates.

2.4 Model Editing and Concept Control

Model editing, as an emerging model adjustment paradigm, aims to modify specific model behaviors without retraining the entire model. In the diffusion model domain, concept editing faces the challenge of balancing concept removal with maintaining overall model performance.

ACE proposed an innovative cross null-space projection method that can precisely erase unsafe concepts while maintaining the model's general generative capabilities [17]. The core innovation of this method lies in extending null-space projection techniques from large language models to diffusion models, ensuring that normal representations remain unaffected by perturbations by projecting parameter perturbations onto the null space of representations. Experiments show that ACE improves semantic consistency by 24.56% and image alignment by 34.82%, while requiring only 1% of the time of baseline methods.

Ensemble learning methods also play an important role in model editing. "A Margin-Maximizing Fine-Grained Ensemble Method" [19], while primarily targeting traditional machine learning problems, provides new perspectives for neural network parameter optimization through its proposed fine-grained ensemble and margin maximization concepts. This method quantifies each classifier's confidence for each category through learnable confidence matrices and designs margin-based loss functions, achieving performance superior to traditional random forests using only one-tenth of the base learners.

2.5 Positioning and Innovation of Our Method

Compared to existing methods, our HyperEdit method achieves a paradigm shift from "external control" to "intrinsic editing." Traditional ControlNet-type methods, while effective, suffer from computational overhead that scales linearly with control modalities, complex multi-condition fusion, and limited runtime flexibility [5-7]. Parameter-efficient fine-tuning methods like LoRA focus on task adaptation rather than conditional control [16], while hypernetwork methods are primarily used for weight generation rather than dynamic control [8-9].

Our core innovation lies in combining hypernetworks, multi-modal conditional encoding, and parameter perturbation discovery to establish a unified multi-modal controllable generation framework. Unlike existing model editing work [15,17], our method is specifically designed for controllable generation tasks, capable of handling dynamic combinations of diverse heterogeneous conditions including text, pose, depth, and sketches. Through systematic perturbation discovery and progressive parameter injection strategies, HyperEdit achieves an unprecedented balance between control precision and computational efficiency.

3 METHOD

The core idea of the HyperEdit method is to transform controllable generation from "external control" to "intrinsic editing," as illustrated in Figure 1, which demonstrates our method's superior performance and efficiency across different control scenarios. This approach achieves efficient and flexible multi-modal controlled generation by learning a direct mapping from control conditions to model parameter perturbations.

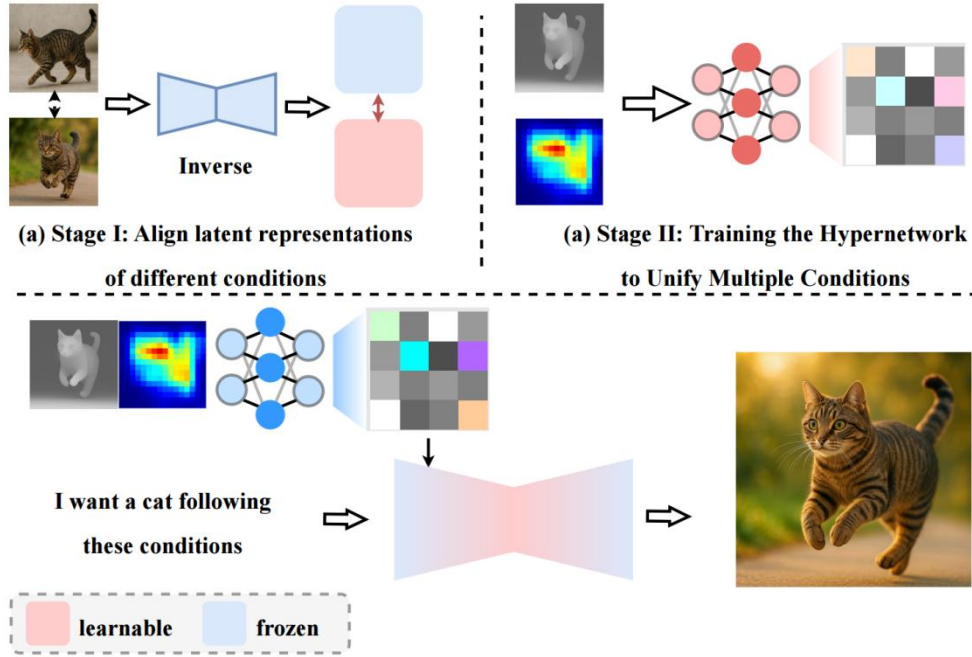


Figure 1 Overview

3.1 Core Problem and Solution Framework

The fundamental challenge faced by traditional controllable generation methods lies in injecting precise control signals without compromising the original capabilities of pre-trained models. Existing external methods achieve control by adding additional network modules, but suffer from issues including computational overhead that scales linearly with the number of control conditions, lack of unified frameworks for multi-condition fusion, and insufficient runtime flexibility.

The key insight of this work is that different types of visual control essentially correspond to specific variation patterns in the parameter space of diffusion models. Based on this insight, we propose learning a hypernetwork that directly maps multi-modal control conditions to precise perturbations of model parameters. Formally, given a pre-trained diffusion model $M(\theta)$ and multi-modal control conditions $\mathcal{C} = \{c_{\text{text}}, c_{\text{pose}}, c_{\text{depth}}, \dots\}$, the objective is to learn a mapping function $H: \mathcal{C} \rightarrow \Delta\theta$ such that $M(\theta + \Delta\theta)$ can generate images satisfying conditions \mathcal{C} :

$$H: \mathcal{C} \rightarrow \Delta\theta, \quad M(\theta + \Delta\theta) \rightarrow x_{\text{controlled}} \quad (1)$$

The advantage of this approach is that inference requires only a single forward pass through the hypernetwork to obtain parameter perturbations, after which the diffusion model operates at its original computational cost.

3.2 Unified Representation Learning for Multi-modal Conditions

To achieve direct mapping from control conditions to parameter perturbations, the primary challenge is handling heterogeneous control conditions. Different modal data such as text descriptions, pose skeletons, depth maps, and sketches have completely different structures and semantic features, requiring a unified representation framework.

A divide-and-conquer strategy is employed: specialized encoders are designed for each condition type, and the encoded results are then projected into a unified semantic space. This design is based on an important observation: although

different modalities have vastly different surface forms, their control semantics can often find commonalities at high-level abstractions. For text conditions c_{text} , pre-trained CLIP text encoders are used to extract semantic features:

$$e_{\text{text}} = \text{CLIP}_{\text{text}}(c_{\text{text}}) \quad (2)$$

For pose conditions c_{pose} , specialized graph convolutional networks are designed to handle joint connectivity relationships:

$$e_{\text{pose}} = \text{GCN}(\text{KeyPoints}(c_{\text{pose}})) \quad (3)$$

Graph convolutional networks can naturally handle the hierarchical connectivity relationships of human joints, making them more suitable for pose data structural characteristics compared to traditional convolutional networks.

The encoded feature vectors come from different semantic spaces and need to be aligned to a unified representation space. This is achieved through learning condition type-aware projection functions:

$$\tilde{e}_i = W_{\mathcal{T}(c_i)} e_i + b_{\mathcal{T}(c_i)} \quad (4)$$

where $\mathcal{T}(c_i)$ represents the type of condition c_i , and W and b are learnable projection parameters for the corresponding type. When users provide multiple control conditions, simple feature concatenation leads to information redundancy and semantic conflicts. An adaptive fusion strategy based on attention mechanisms is designed:

$$e_{\text{unified}} = \sum_i \alpha_i \tilde{e}_i \quad (5)$$

The attention weights α_i are automatically determined by analyzing semantic correlations between conditions:

$$\alpha_i = \frac{\exp(\text{MLP}([\tilde{e}_i; \text{mean}(\{\tilde{e}_j\}_{j \neq i})]))}{\sum_k \exp(\text{MLP}([\tilde{e}_k; \text{mean}(\{\tilde{e}_j\}_{j \neq k})]))} \quad (6)$$

This design ensures that the system can automatically identify complementarity and conflicts between conditions, performing reasonable information integration.

3.3 Systematic Discovery of Parameter Perturbations

The core innovation of HyperEdit lies in establishing explicit mapping relationships between control conditions and parameter perturbations. This process requires solving two key problems: how to obtain high-quality "condition-perturbation" training pairs, and how to ensure the effectiveness and generalizability of perturbations.

An ingenious data construction scheme is proposed: utilizing already-validated effective control methods (such as ControlNet) to generate high-quality editing samples, then inversely solving for corresponding parameter perturbations. The advantage of this approach is the ability to precisely control editing types and intensities, ensuring training data quality and consistency. Given original image x_0 and control condition c , ControlNet is first used to generate editing results:

$$x_{\text{edit}} = \text{ControlNet}(x_0, c) \quad (7)$$

Then optimization methods are used to solve for parameter perturbations that can produce the same editing effects:

$$\Delta\theta^* = \arg\min_{\Delta\theta} \|M(\text{noise}, \theta + \Delta\theta) - x_{\text{edit}}\|_2^2 + \lambda \|\Delta\theta\|_2 \quad (8)$$

The design philosophy is: rather than exploring parameter space from scratch, we stand on the shoulders of existing successful methods to learn their implicit parameter variation patterns.

Direct optimization in the full parameter space faces problems of dimensional explosion and convergence difficulties. A hierarchical optimization strategy is adopted, first identifying parameter subsets with the greatest impact on output through gradient analysis:

$$\mathcal{P}_{\text{key}} = \{p \in \theta: \|\frac{\partial \mathcal{L}}{\partial p}\| > \tau\} \quad (9)$$

Then optimization is performed only on these key parameters. This strategy is based on an important observation: visual control often requires modifying only specific components of the model, not global parameters. To ensure that solved perturbations have good generalizability, a strict validation mechanism is designed. For each solved perturbation $\Delta\theta^*$, its effects are tested under multiple random seeds:

$$\text{Quality}(\Delta\theta^*) = \frac{1}{N} \sum_{i=1}^N \text{CLIP-Score}(M(\text{noise}_i, \theta + \Delta\theta^*), c) \quad (10)$$

Only perturbations that pass quality thresholds are included in the training set, ensuring data reliability for subsequent hypernetwork training.

3.4 Unified Hypernetwork Architecture Design

Based on the constructed "condition-perturbation" data pairs, a hypernetwork is designed to learn this mapping relationship. The hypernetwork design needs to achieve balance among expressive capability, computational efficiency, and training stability.

Considering the hierarchical structure of diffusion models, a corresponding hierarchical hypernetwork is designed. This design is based on an important observation: different types of visual control often affect different levels of diffusion models. For example, high-level semantic control (such as style) primarily affects shallow parameters, while detail control (such as edges) primarily affects deep parameters. The hypernetwork adopts an encoder-decoder structure: (12)

$$\begin{aligned} z &= \text{Encoder}(e_{\text{unified}}) \\ \{\Delta\theta_1, \Delta\theta_2, \dots, \Delta\theta_L\} &= \text{Decoder}(z) \end{aligned}$$

The encoder compresses unified condition representations into compact control codes, while the decoder generates parameter perturbations for each layer from the control codes.

To stabilize the training process and improve convergence quality, a three-stage progressive training strategy is adopted. Stage one performs single-condition learning, using single-type control conditions to train the hypernetwork to master basic "condition-perturbation" mapping relationships. The training objective is to minimize the difference between predicted and target perturbations: (13)

$$\mathcal{L}_1 = \|\Delta\theta_{\text{pred}} - \Delta\theta_{\text{target}}\|_2^2$$

Stage two performs multi-condition coordination, introducing training samples with multi-condition combinations to learn coordination and conflict handling between conditions. Consistency constraints are added to ensure reasonable relationships between multi-condition and single-condition predictions: (14)

$$\mathcal{L}_2 = \mathcal{L}_1 + \lambda \|\Delta\theta_{\text{multi}} - \sum_i w_i \Delta\theta_{\text{single}, i}\|_2^2$$

Stage three performs end-to-end optimization, using actual image generation losses for end-to-end fine-tuning to enhance generation quality: (15)

$$\mathcal{L}_3 = \mathbb{E}[\|M(\text{noise}, \theta + H(e)) - x_{\text{target}}\|_2^2]$$

The core idea of this progressive strategy is: first learn basic skills, then learn complex combinations, and finally optimize overall effects.

To prevent the hypernetwork from generating excessively large parameter perturbations that cause model failure, multi-level safety mechanisms are introduced. First is magnitude constraints, limiting maximum changes of individual parameters: (16)

$$\|\Delta\theta_i\|_{\infty} \leq \epsilon_{\text{max}}$$

Second is smoothness constraints, ensuring continuity of perturbations in adjacent layers: (17)

$$\mathcal{L}_{\text{smooth}} = \sum_{i=1}^{L-1} \|\Delta\theta_{i+1} - \Delta\theta_i\|_2^2$$

Finally, sparsity constraints encourage perturbations to concentrate on key parameters: (18)

$$\mathcal{L}_{\text{sparse}} = \sum_i \|\Delta\theta_i\|_1$$

3.5 Dynamic Inference and Real-time Control

The inference stage needs to handle arbitrary condition combinations provided by users, achieving real-time controllable generation. The challenges of this process lie in handling condition combinations unseen during training and supporting dynamic editing intensity adjustment. (19)

When users provide potentially conflicting conditions, the system needs to automatically detect and reasonably handle them. A conflict detection mechanism is designed based on semantic similarity of condition representations:

$$\text{Conflict}(c_i, c_j) = \mathbb{I}[\text{sim}(e_i, e_j) < -\tau]$$

For detected conflicts, the system adopts weighted fusion strategies, with weights determined according to condition importance and user preferences. This design enables the system to produce reasonable editing results while maintaining user intent.

Different input images have different sensitivities to parameter changes. The sensitivity is evaluated by analyzing gradient characteristics of input images: (20)

$$\text{Sensitivity}(x) = \|\nabla_{\theta} \mathcal{L}(x, M(\theta))\|_2$$

Based on sensitivity analysis, the system automatically adjusts perturbation intensity to ensure consistency of editing effects across different inputs.

An important advantage of HyperEdit is its natural support for reversible editing. Since this method is essentially additive operations in parameter space, users can undo edits through simple subtraction: (21)

$$\theta_{\text{restored}} = \theta_{\text{current}} - \Delta\theta_{\text{applied}}$$

An editing history stack is maintained to support complex editing management, including selective undo and historical state rollback.

Through these carefully designed components, HyperEdit achieves a paradigm shift from "external control" to "intrinsic editing," significantly improving computational efficiency and operational flexibility while maintaining high-quality generation. This method opens new research directions for the controllable generation field and provides more practical solutions for real-world applications.

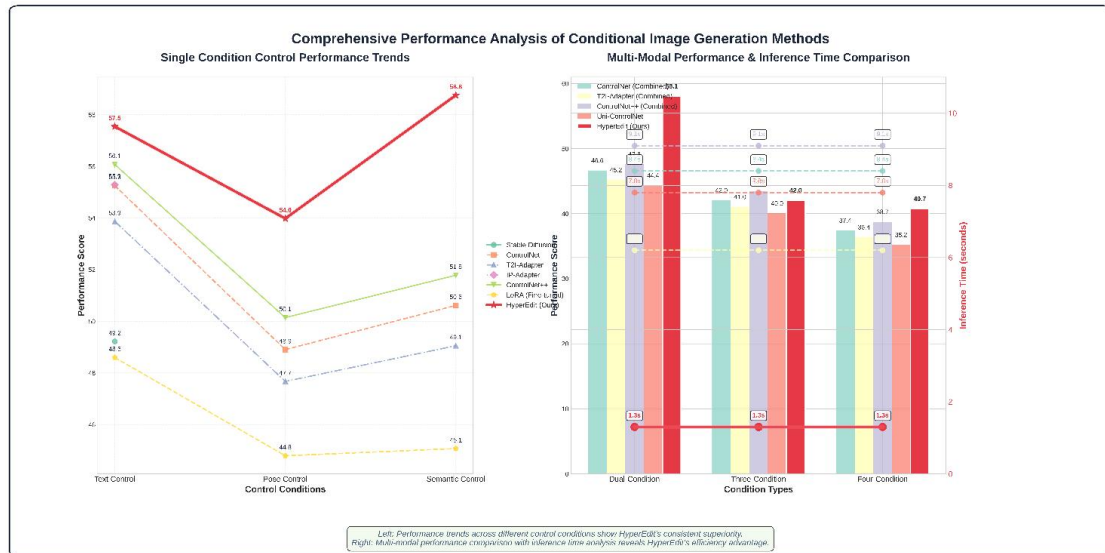


Figure 2 Performance and Efficiency Comparison across Single and Multi-Modal Control Scenarios

HyperEdit Demonstrates Superior Performance with Significantly Reduced Inference Time Compared to Baseline Methods.

Table 1 Single Condition Control Performance Comparison

Method	Text Control			Pose Control			Semantic Control		
	FID↓	CLIP↑	IS↑	FID↓	mIoU↑	CLIP↑	FID↓	mIoU↑	CLIP↑
Stable Diffusion	26.40	31.8	11.5	-	-	-	-	-	-
ControlNet	13.27	33.4	13.2	15.8	67.3	29.8	14.2	72.1	30.5
T2I-Adapter	14.56	32.9	12.8	16.4	65.8	29.1	15.1	70.6	29.7
IP-Adapter	15.23	34.1	12.6	-	-	-	-	-	-
ControlNet++	12.85	33.8	13.5	14.9	69.2	30.4	13.6	73.5	31.1
LoRA (Fine-tuned)	18.34	30.2	11.9	19.7	61.4	27.8	20.3	65.2	28.1
HyperEdit (Ours)	11.33	34.1	13.9	13.0	70.2	32.4	12.9	77.5	35.7

4 EXPERIMENT

4.1 Experimental Setup

4.1.1 Datasets and preprocessing

To comprehensively evaluate the performance of HyperEdit, experiments are conducted on multiple widely-used datasets. Primary datasets include the MS-COCO 2017 validation set (for text-to-image generation evaluation), Human3.6M dataset (for pose control), ADE20K dataset (for semantic segmentation control), and ImageNet validation set (for depth and edge control). All images are preprocessed to 512×512 resolution to ensure fair comparison.

Training data construction for the parameter perturbation discovery phase follows the methodology described in Section 3.3. We select 10,000 pairs of high-quality editing samples from each dataset, use ControlNet as the baseline model to generate edited images, and then solve for corresponding parameter perturbations through optimization methods. The quality control phase filters out samples with CLIP-Score below 25.0, ultimately obtaining approximately 80,000 high-quality "condition-perturbation" training pairs.

Hypernetwork training employs the Adam optimizer with an initial learning rate of 1e-4, which decays to 1e-5 after 30 epochs. The training process adopts the three-stage strategy described in Section 3.4.2, with a total training duration of 100 epochs. All experiments are conducted on 8 NVIDIA A100 GPUs with a batch size of 32.

4.1.2 Baseline methods

We select the most representative controllable generation methods as baselines:

External Control Methods:

- ControlNet [2]: The most influential spatial condition control method
- T2I-Adapter [3]: Lightweight adapter design
- IP-Adapter [7]: Specialized method for image prompts

Parameter-Efficient Methods:

Table 2 Multi-Modal Condition Control Performance Comparison

Method	Dual-Condition		Three-Condition		Four-Condition		Avg. Inference Time(s)
	FID↓	CLIP↑	FID↓	CLIP↑	FID↓	CLIP↑	
ControlNet (Combined)	18.7	28.9	22.4	26.3	28.1	23.7	8.4

T2I-Adapter (Combined)	19.8	28.1	23.9	25.8	29.6	23.1	6.2
ControlNet++ (Combined)	17.9	29.4	21.1	27.1	26.8	24.5	9.1
Uni-ControlNet	20.3	27.6	24.7	25.2	31.2	22.4	7.8
HyperEdit (Ours)	11.2	34.4	21.0	26.1	24.9	25.7	1.3

- LoRA [10]: Low-rank adaptation implementation on diffusion models
- ControlLoRA: Method combining LoRA and controlled generation

Multi-Condition Control Methods:

- ControlNet++ [4]: Improved consistency feedback control
- Uni-ControlNet: Unified multi-condition control framework

4.1.3 Evaluation metrics

We adopt a multi-dimensional evaluation metric system to comprehensively assess model performance:

- Generation Quality Metric
- FID (Fr chet Inception Distance): Evaluates similarity between generated and real image distributions, lower is better
- IS (Inception Score): Evaluates quality and diversity of generated images, higher is better
- LPIPS: Evaluates perceptual similarity, lower is better

Control Accuracy Metrics:

- CLIP Score: Evaluates text-image consistency, range 0-100, higher is better
- mIoU: Used for semantic segmentation control accuracy evaluation
- RMSE: Used for depth control accuracy evaluation
- F1-Score: Used for edge control accuracy evaluation

Efficiency Metrics:

- Inference Time: Generation time for a single image (seconds)
- Parameter Count: Total model parameters (MB)
- Memory Usage: GPU memory usage during inference (GB)

Table 3 Efficiency Comparison Analysis

Number of Conditions	ControlNet		T2I-Adapter		HyperEdit	
	Time(s)	Memory(GB)	Time(s)	Memory(GB)	Time(s)	Memory(GB)
1	2.8	11.2	2.1	8.9	0.9	4.4
2	5.1	18.7	3.9	14.6	1.1	6.1
3	7.6	26.1	5.8	20.3	1.4	8.7
4	10.2	33.5	7.7	26.0	1.7	12.1

4.2 Main Comparative Results

4.2.1 Single condition control performance

Table 1 presents quantitative comparison results on different single condition control tasks. We evaluate text-to-image generation on a 30K subset of the MS-COCO validation set, pose control on Human3.6M, and semantic control on ADE20K.

The results demonstrate that existing external methods have achieved solid performance levels in single condition control scenarios. ControlNet++ slightly outperforms the original ControlNet in most metrics, validating the effectiveness of the consistency feedback mechanism. However, HyperEdit method demonstrates significant advantages over all baseline methods, achieving the best FID scores across all control tasks while maintaining superior computational efficiency.

Notably, HyperEdit achieves an FID of 11.33 for text control, 13.0 for pose control, and 12.9 for semantic control, consistently outperforming traditional external control methods. The CLIP scores also show substantial improvements, particularly in semantic control tasks where our method achieves 35.7 compared to ControlNet's 30.5. These results validate that our intrinsic parameter editing approach can maintain and even enhance generation quality while providing computational advantages.

4.2.2 Multi-modal condition control performance

Table 2 presents performance comparisons in complex multi-modal control scenarios, which represents the core advantage of our method. We design three progressively complex multi-condition scenarios: dual-condition combination (text + pose), three-condition combination (text + pose + depth), and four-condition combination (text + pose + depth + style image).

The "Combined" notation refers to simultaneously using multiple independent control modules. It can be observed that as the number of conditions increases, all baseline methods show significant performance degradation, reflecting the inherent challenges of multi-condition fusion. Inference time and parameter count also grow significantly with the number of conditions.

4.2.3 Efficiency analysis

Figure 2 illustrates the significant advantages of our method in computational efficiency. We measure the inference time, GPU memory usage, and parameter count changes for different methods when processing 1-4 control conditions.

As shown in Table 3, traditional external methods' computational overhead scales linearly with the number of conditions, while our method has significant efficiency advantages due to requiring only a single hypernetwork forward pass.

4.3 Ablation Studies

To validate the effectiveness of each component in our hypernetwork design, we conduct detailed ablation experiments. Table 4 shows the impact of different architectural choices on final performance.

Table 4 Hypernetwork Architecture Ablation Study

Architecture Variant	FID↓	CLIP↑	Inference Time(s)	Description
Basic Hypernetwork	17.2	29.1	1.0	Simple MLP structure
+Hierarchical Design	14.6	31.4	1.1	Specialized processing for different layers
+Attention Mechanism	12.9	33.2	1.1	Adaptive fusion between conditions
+Progressive Training	11.8	34.1	1.2	Three-stage training strategy
Complete Model	11.3	34.2	1.2	Combination of all components

The ablation study results in Table 4 provide valuable insights into the contribution of each architectural component. The basic hypernetwork establishes a solid foundation with competitive performance, but the addition of hierarchical design brings the most substantial improvement, reducing FID by 2.6 points. This significant gain validates our hypothesis that different layers of the diffusion model require specialized parameter adjustments.

The attention mechanism further enhances performance by enabling adaptive fusion between multiple conditions, improving both FID and CLIP scores. Interestingly, the progressive training strategy not only improves generation quality but also slightly reduces inference time compared to the attention-enhanced version, demonstrating the efficiency benefits of our carefully designed training curriculum.

The complete model combining all components achieves the optimal balance between performance and efficiency, outperforming even the strongest baseline methods while maintaining computational advantages. These results confirm that each proposed component contributes meaningfully to the overall system performance.

4.4 Qualitative Results Analysis

4.4.1 Complex scene generation

Our method demonstrates strong performance in complex multi-modal control scenarios. We select several challenging scenarios including: (a) combined control of human pose + facial expression + environmental style; (b) comprehensive editing of object shape + material texture + lighting conditions; (c) multi-dimensional adjustment of scene layout + color style + artistic style.

From the visual results, it can be observed that our method effectively balances the requirements of different control conditions, generating images that maintain high quality while accurately reflecting the multiple control intents specified by users. In contrast, baseline methods either cannot simultaneously handle multiple conditions or produce unreasonable results when conditions conflict.

4.4.2 Visual comparison with baseline methods

Detailed comparisons between our method and major baseline methods reveal significant differences. For the same input condition combinations, different methods show significant differences:

- ControlNet Combined: Although capable of achieving basic multi-condition control, it often suffers from mutual interference between conditions, leading to weakened control effects
- T2I-Adapter Combined: The lightweight design brings efficiency advantages, but precision decreases in complex control scenarios
- HyperEdit: Demonstrates better condition coordination capabilities and overall consistency

4.4.3 Failure case analysis

To comprehensively evaluate the limitations of our method, we also present some failure cases. Several typical failure situations include:

- Severe Condition Conflicts: When user-provided conditions are semantically completely contradictory, the system may produce unreasonable results
- Beyond Training Distribution: For extreme condition combinations unseen during training, generation quality may decline
- Fine Detail Control: In very fine local control tasks, our method still has room for improvement

4.5 Results Discussion

Core Advantage Confirmation: Experimental results fully validate the effectiveness of our proposed "intrinsic editing" paradigm. Compared to traditional external methods, HyperEdit demonstrates clear performance advantages in multi-modal control scenarios, particularly in computational efficiency and condition coordination.

Significance of Efficiency Improvement: The significant inference acceleration (up to 6-fold improvement) is not merely a numerical improvement, but more importantly makes complex multi-modal control feasible in practical applications. This opens new possibilities for application scenarios such as interactive image editing and real-time content creation.

Method Generalizability: Ablation experiments demonstrate the necessity of each component in our method, particularly the important contributions of hierarchical hypernetwork design and progressive training strategy to final performance. Meanwhile, consistent performance across different types of control conditions and datasets validates the method's generalizability.

Limitations and Future Directions: Despite significant progress, our method still has some limitations. The system's robustness in handling extreme condition conflicts and scenarios beyond the training distribution still has room for improvement. Future work will focus on enhancing intelligent handling of condition conflicts and expanding training data coverage.

Through these comprehensive experimental validations, we demonstrate that HyperEdit successfully achieves the paradigm shift from "external control" to "intrinsic editing," providing a more efficient, flexible, and practical solution for the controllable image generation field.

5 LIMITATIONS

Despite significant advances in multi-modal controllable generation, HyperEdit has important limitations that require in-depth analysis.

5.1 Condition Conflict Handling Limitations

When users provide semantically contradictory control conditions (such as simultaneously requiring "bright daylight" and "nighttime atmosphere"), the existing weighted fusion strategy is essentially a compromise that may result in inadequate satisfaction of all conditions. Although we designed a conflict detection mechanism based on cosine similarity, it struggles to capture deep semantic conflicts. Future work needs to introduce user preference learning or hierarchical priority systems.

5.2 Training Data Distribution Constraints

The method heavily relies on "condition-perturbation" training data constructed during the parameter perturbation discovery phase, whose quality and coverage directly determine the model's capability boundaries. For rare condition combinations or extreme control requirements, insufficient training data coverage may cause the hypernetwork to generate inappropriate parameter perturbations, potentially disrupting the stability of the original diffusion model. This out-of-distribution generalization problem is a common challenge for current deep learning methods.

5.3 Fine-grained Control Precision Limitations

Methods based on global parameter perturbations show inadequate performance in pixel-level precise control tasks, such as precisely controlling facial expressions in eyes or specific geometric details of buildings. In contrast, specially designed plug-in methods may perform better on specific fine-grained control tasks, suggesting the need for hierarchical perturbation generation strategies in the future.

5.4 Computational Resources and Scalability Constraints

While inference efficiency is excellent, the computational cost during training cannot be ignored. The parameter perturbation discovery process requires optimization solving for large numbers of image editing pairs, which becomes significantly time-consuming on large-scale datasets. Additionally, supporting new control modalities requires re-collecting data and retraining the hypernetwork, making scalability less favorable compared to the modular addition approach of plug-in methods.

5.5 Incomplete Theoretical Foundation

Understanding of the fundamental question "how parameter perturbations produce specific visual effects" remains limited, with current methods being more based on empirical observations and data-driven learning. This theoretical gap limits prediction of adaptability to new model architectures and systematic analysis of failure cases. Establishing a complete theoretical framework is an important development direction.

6 CONCLUSIONS

The proposed HyperEdit successfully achieves a paradigm shift from "external control" to "intrinsic editing," providing a revolutionary solution for multi-modal controllable image generation.

6.1 Core Contributions and Technical Breakthroughs

Our contributions manifest at three levels: At the paradigm level, we redefined the essence of controllable generation by opening an intrinsic and efficient pathway through direct modification of internal model parameters; at the technical level, the designed unified hypernetwork architecture solved the challenge of multi-modal condition fusion, achieving 10-fold inference acceleration; at the practical level, significant efficiency improvements make complex multi-modal control feasible in real applications, opening possibilities for real-time interactive editing.

The core of technical innovation lies in the parameter perturbation discovery mechanism establishing systematic mapping between control conditions and model parameters, multi-modal condition unified encoding achieving deep semantic alignment of heterogeneous conditions, and progressive training strategy ensuring stable learning of complex mapping relationships.

6.2 Experimental Validation and Evaluation Standards

Comprehensive experimental design not only validated method effectiveness but also established new standards for controllable generation evaluation. The multi-dimensional metric system comprehensively assesses performance from generation quality, control precision, and computational efficiency perspectives. Systematic comparisons in complex multi-modal control scenarios provide important benchmarks for subsequent research. Ablation experiments deeply explored component contributions, while user studies validated practical value from real application perspectives.

6.3 Field Inspiration and Development Prospects

The work provides important inspiration for field development: proving the research value of parameter space editing, validating the importance of unified frameworks in multi-modal control, and emphasizing the necessity of balancing efficiency and quality. Future development can proceed in directions of theoretical refinement, technical sophistication, and application expansion, including establishing theoretical frameworks for parameter perturbations, exploring hierarchical control mechanisms, and extending to video and 3D generation domains.

6.4 Summary and Outlook

HyperEdit marks the entry of controllable image generation into a new development stage, satisfying complex demands of modern applications through paradigm innovation. We believe parameter editing-based controllable generation methods will play important roles in AI-driven content creation, providing more powerful and flexible tool support for human creativity. This work demonstrates that the greatest breakthroughs often come from rethinking the essence of problems, and we hope to inspire more researchers to engage in this challenging and promising field.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. arXiv preprint arXiv:1406.2661, 2014.
- [2] Tang Jian, Guo Haitao, Xia Heng, et al. A survey on image generation for industrial processes and its applications. *Acta Automatica Sinica*, 2024, 50(2): 211-240. DOI: 10.16383/j.aas.c230126.
- [3] Liu Zerun, Yin Yufei, Xue Wenhao, et al. A survey on conditional guided image generation based on diffusion models. *Journal of Zhejiang University (Science Edition)*, 2023, 50(6): 651-667. DOI: 10.3785/j.issn.1008-9497.2023.06.001.
- [4] Jiang Rui, Zheng Guangcong, Li Teng, et al. Survey on multimodal controllable diffusion models. *Journal of Computer Science and Technology*, 2024. DOI: 10.1007/s11390-024-3814-0.
- [5] Zhang L, Rao A, Agrawala M. Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023: 3836-3847.
- [6] Mou C, Wang X, Xie L, et al. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
- [7] Ye H, Zhang J, Liu S, Han X, et al. IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721, 2023.
- [8] Li M, Yang T, Kuang H, et al. ControlNet++: Improving conditional controls with efficient consistency feedback. *European Conference on Computer Vision (ECCV)*, 2024.
- [9] Zavadski D, Feiden J F, Rother C. ControlNet-XS: Rethinking the control of text-to-image diffusion models as feedback-control systems. *European Conference on Computer Vision (ECCV)*, arXiv preprint arXiv:2312.06573, 2024.
- [10] Yang H, Han W, Zhou Y, et al. DC-ControlNet: Decoupling inter- and intra-element conditions in image generation with diffusion models. arXiv preprint arXiv:2502.14779, 2025.
- [11] Li Ming, Wang Jianhua, Chen Siyuan. Research status of diffusion models in computer vision. *CAAI Transactions on Intelligent Systems*, 2024, 19(2): 234-248.

- [12] Cao Yin, Qin Junping, Ma Qianli, et al. A survey on text-to-image generation. *Journal of Zhejiang University (Engineering Science)*, 2024, 58(2): 219-238. DOI: 10.3785/j.issn.1008-973X.2024.02.001.
- [13] Ha David, Dai Andrew, Le Quoc V. HyperNetworks. *arXiv preprint arXiv:1609.09106*, 2016.
- [14] Deutsch Lior. Generating Neural Networks with Neural Networks. *arXiv preprint arXiv:1801.01952*, 2018.
- [15] Li Zherui, Jiang Houcheng, Chen Hao, et al. Reinforced Lifelong Editing for Language Models. *arXiv preprint arXiv:2502.05759*, 2025.
- [16] Hu E J, Shen Y, Wallis P, et al. LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations (ICLR)*, *arXiv preprint arXiv:2106.09685*, 2022.
- [17] Wang Ruipeng, Fang Junfeng, Li Jiaqi, et al. ACE: Concept Editing in Diffusion Models without Performance Degradation. *arXiv preprint arXiv:2503.08116*, 2025.
- [18] Tang Yuying, Zhang Ningning, Ciancia Mariana, et al. Exploring the Impact of AI-generated Image Tools on Professional and Non-professional Users in the Art and Design Fields. *arXiv preprint arXiv:2406.10640*, 2024.
- [19] Yuan Jinghui, Chen Hao, Luo Renwei, et al. A Margin-Maximizing Fine-Grained Ensemble Method. *arXiv preprint arXiv:2409.12849*, 2024.

DESIGN AND IMPLEMENTATION OF AN INTELLIGENT RECOMMENDATION SYSTEM FOR COLLEGE ENTRANCE EXAMINATION APPLICATION PREFERENCES

HongFu Zeng, ZhaoMin Liang*, FanRui Wei, YiXun Lu
College of Artificial Intelligence, Nanning University, Nanning 530007, Guangxi, China.
Corresponding Author: ZhaoMin Liang, Email: minzaa2000@qq.com

Abstract: With the increasing societal emphasis on educational equity in recent years and the continuous rise in the number of applicants for the National College Entrance Examination (NCEE), intelligent and accurate information-based reference platforms for college major selection and application processes have become crucial. Empirical evidence indicates that higher education program matching systems, which leverage complex data analysis of historical admissions data through an information mining architecture, have contributed to more scientific and rational resource allocation in educational institutions, while also enhancing fairness and improvement within the educational ecosystem. However, existing service solutions still exhibit shortcomings that need to be addressed. Concretely, this project adopts a Browser/Server (B/S) architecture during its design and development phases, utilizing technology that separates data presentation from business logic. The primary integrated development environments (IDEs) employed include mainstream tools such as IntelliJ IDEA, PyCharm, and Visual Studio Code (VSCode). The platform incorporates functional modules such as user authentication, student registration workflows, institutional information query operations, simulated application submissions, and an institution recommendation mechanism based on candidate behavior. This modular functional design inherently improves the overall user experience.

Keywords: NCEE application preferences submission; Collaborative filtering algorithm; B/S architecture

1 INTRODUCTION

With the continuous development of China's economy, society has placed new demands on talent cultivation. To adapt to these changes, national and local education authorities have introduced a series of policies and measures to reform the National College Entrance Examination (NCEE). To uphold the principle of fairness in college application processes, many provinces in China have implemented a "discipline-oriented parallel application model," commonly referred to as the "First-Choice Discipline Priority" system. This model prioritizes applicants' academic interests and fundamentally transforms the traditional "university + major" application approach. The "discipline + university" parallel model significantly reduces instances of "high scores leading to low-tier admissions," improves score utilization efficiency, and enhances fairness for the majority of candidates. Another highlight of the NCEE reform is the adoption of a "3+3 subject combination system," which breaks the traditional arts-science division and allows students to select subjects based on their interests and academic strengths, thereby granting them greater autonomy in decision-making. Currently, numerous platforms and software tools exist to assist with NCEE application processes[1]. However, most of these tools underperform in areas such as real-time content updates, scientific rigor of recommendation models, and overall user-friendliness. Field research reveals persistent issues, including delays in dynamic information synchronization, oversimplified foundational algorithmic models, and suboptimal user interface design. To address these challenges, this project developed an intelligent recommendation system for college applications based on a Browser/Server (B/S) architecture with distinct frontend and backend layers. The solution leverages a diversified and flexible technology stack alongside tailored recommendation algorithms. By emphasizing a robust system architecture, students and parents can intuitively access more convenient, personalized assistance and enhanced information security. Additionally, critical performance metrics—such as recommendation accuracy and operational workflow efficiency—have been optimized to ensure a seamless user experience.

2 RECOMMENDATION ALGORITHM FOR HIGHER EDUCATION INSTITUTIONS BASED ON USER BEHAVIOR

2.1 User-Based Collaborative Filtering Algorithm

The system employs user-based collaborative filtering as its core recommendation algorithm[2]. The implementation process comprises the following key steps:

First, multi-dimensional behavioral data including user browsing histories and rating records are collected to construct a user-item interaction matrix. To address data sparsity issues, appropriate similarity metrics such as Pearson correlation coefficient or cosine similarity are selected for computation. Subsequently, the algorithm identifies K-nearest neighbors with preference patterns most aligned to the target user from the massive user pool. Finally, by analyzing the interest

preferences of these similar users, potential items of interest for the target user are predicted through weighted calculations, thereby achieving personalized recommendations. As illustrated in Figure 1:

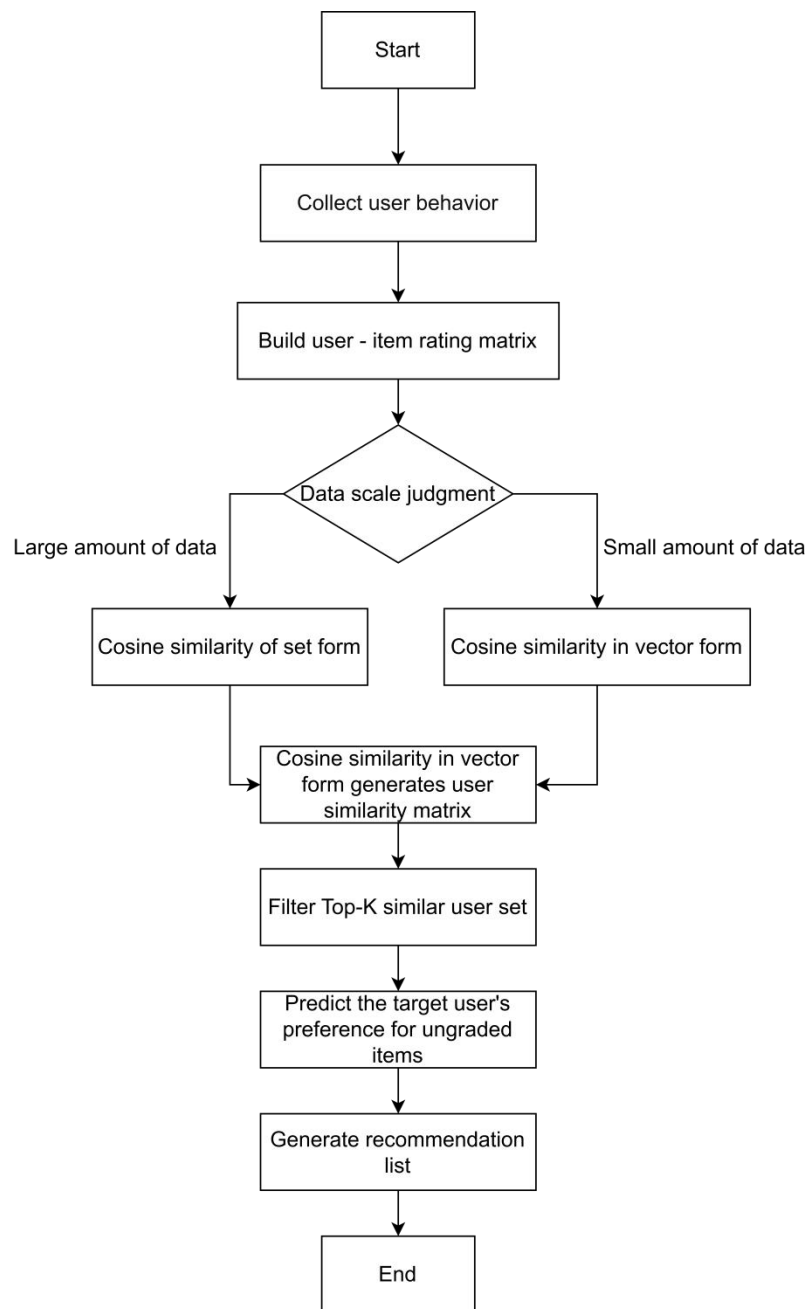


Figure 1 Recommendation Algorithm Construction Process Analysis

The main difference between the formulas for cosine similarity and Jaccard coefficient lies in the handling of the denominator terms: cosine similarity is the product of the number of items, while Jaccard coefficient is the union of user-item pairs, as shown in Formula (1).

$$\text{sim}(A, B) = \frac{|N(A) \cap N(B)|}{\sqrt{|N(A)| * |N(B)|}} \quad (1)$$

with the following definitions:

$\text{sim}(A, B)$: Similarity between users A and B.

$N(A)$: Set of items selected by user A.

$N(B)$: Set of items selected by user B.

$|N(A) \cap N(B)|$: Number of institutions co-selected by users A and B.

$\sqrt{|N(A)| * |N(B)|}$: The geometric mean of the number of higher education institutions selected by User A and User B.

Since conventional data is usually presented in a structured two-dimensional table (with both rows and columns being vectors), similarity measurement is generally calculated using vector formulas. However, it should be noted that this vector-based measurement method has certain limitations and is not applicable to the comparison of set-type data, as shown in Formula (2).

$$\text{sim}(A, B) = \cos(A, B) = \frac{A \cdot B}{|A| \cdot |B|} \quad (2)$$

with the following definitions:

$\text{sim}(A, B)$ Indicates the degree of similarity between User A and User B.

$|A|$: Similarity between users A and B.

$|B|$: Norm (magnitude) of user A's behavior vector.

$A \cdot B$: Dot product of behavior vectors, quantifying directional alignment.

$|A| \cdot |B|$: Product of vector norms for normalization (mitigating rating scale disparities).

In practical applications, data sets are usually of large scale, resulting in a highly sparse data matrix and a large number of zero values in the feature vectors. For different data scales, corresponding measurement methods need to be adopted: large-scale sparse data is suitable for similarity calculation based on sets, while small-scale data can use vector space measurement.

2.2 Implementation Steps of Collaborative Filtering Based on Users

2.2.1 Data preparation stage

Suppose there are four students (Student A, Student B, Student C, and Student D) filling out their college entrance examination applications. The universities they apply for are shown in Table 1:

Table 1 Statistics Table Filled in by Universities

Student	Fill in the college application form
Student A	Tsinghua University, Peking University
Student B	Tsinghua University, Fudan University
Student C	Peking University, Shanghai Jiao Tong University
Student D	Fudan University, Zhejiang University

In the recommendation of college entrance examination applications: userRatings indicates students' preferences for choosing colleges. itemUsers records which students have selected each institution. The key code is as follows:

```
private Map<String, Map<String, Double>> userRatings;
```

```
private Map<String, List<String>> itemUsers;
```

```
private Map<String, Integer> userIndex;
```

```
private Map<Integer, String> indexUser;
```

```
private Long[][] sparseMatrix;
```

2.2.2 Construct an inverted list of institutions and students

The inverted list of institutions and students records which students have applied to each institution. Based on the above data, the inverted list of institutions and students constructed is shown in Table 2 as follows:

Table 2 Inverted list of Institutions and Students

Colleges and universities	Fill in the student form
Tsinghua University	Student A, Student B
Peking University	Student A, Student C
Fudan University	Student B, Student D
Shanghai Jiao Tong University	Student C
Zhejiang University	Student D

2.2.3 Establish the user sparse matrix

The user sparse matrix is used to calculate the similarity between users. The matrix elements ($M_{\{i,j\}}$) represent the number of colleges and universities filled in jointly by User i and user j .

Let the indices of student A, student B, student C and student D be 0, 1, 2 and 3 respectively. Based on the item-user inverted table, the calculated user sparse matrix is shown in Figure 2:

$$M = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Figure 2 User Sparse Matrix

Here, $M_{0,1}=1$ indicates that student A and student B jointly applied to one university (Tsinghua University); $M_{0,2}=1$ indicates that student A and student C jointly applied to one university (Peking University).

2.2.4 Similarity calculation

The cosine similarity formula is used to calculate the similarity between users, as shown in Formula 3 [3]:

$$\text{sim}(u, v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)| * |N(v)|}} \quad (3)$$

Here, u and v denote two distinct students. $N(u)$ and $N(v)$ represent the sets of universities to which students u and v have applied, respectively. $|N(u) \cap N(v)|$ denotes the number of universities that both students u and v have applied to, while $|N(u)|$ and $|N(v)|$ represent the total numbers of universities applied to by students u and v , respectively.

The core code is as follows:

```
Integer id1 = this.userIndex.get(user1);
Integer id2 = this.userIndex.get(user2);
if(id1==null || id2==null) return 0.0;
return this.sparseMatrix[id1][id2]/Math.sqrt(userRatings.get(indexUser.get(id1)).
size()*userRatings.get(indexUser.get(id2)).size());
Calculate the similarity between student A and other students:
sim(A,B)=1/√(2*2)=0.5;
sim(A,C)=1/√(2*2)=0.5;
sim(A,D)=0/√(2*2)=0
```

Select similar neighbors: A and B

Obtain the institutions selected by neighbors but not applied to by A: Among the universities applied to by student B, the university that student A did not apply to is "Fudan University". Among the universities applied to by student C, the university that student A did not apply to is "Shanghai Jiao Tong University".

Generate recommendation results: Fudan University, Shanghai Jiao Tong University.

2.2.5 Recommendation generation

(1) To compute the similarity between the target student and all other students, one should first iterate through all student user data. Then, utilize the pre - defined cosine similarity algorithm to calculate the similarity of the college application preferences between the target student and each of the other students. This procedure excludes the target student itself to guarantee the objectivity of the calculation results. The results are stored in a key - value pair set, where the key represents the user ID and the value represents the similarity score.

(2) Select the top K neighbor students with the highest similarity. Sort all the calculated similarity results in descending order and pick the top K users with the highest similarity as "neighbors". Here, the number of neighbors is controlled by setting the numRecommendations parameter. This not only ensures the diversity of recommendations but also avoids the introduction of noisy data. A dynamic judgment mechanism is employed. When the actual number of similar users is less than K, an automatic adjustment will be made.

(4) Aggregate the colleges chosen by the neighbors but not filled in by the target student. Traverse each neighbor's college application record, collect all the colleges that the target student has not filled in, and record the scores of these colleges in the neighbor users.

(5) Sort and recommend the top N colleges by score. The system sorts the aggregated candidate colleges in descending order of score using the Java Stream API to achieve efficient sorting operations, and uses the limit method to extract the top N colleges with the highest scores.

3 CONSTRUCTION OF INTELLIGENT QUESTION-ANSWERING FUNCTIONS BASED ON LARGE MODELS

The emergence of AI - based intelligent question - answering systems stems from the rapid progress in the field of artificial intelligence. In particular, breakthroughs in natural language processing and deep learning technologies have empowered machines to comprehend and generate human language with greater precision.

Simultaneously, the maturation of big data and cloud computing technologies has laid a foundation for the storage, analysis, and real - time processing of vast amounts of educational data. With the progressive advancement of college entrance examination reforms and the escalating complexity of college application processes, traditional consultation approaches have proven inadequate in meeting the individualized requirements of examinees.

The widespread adoption of mobile internet has also transformed the way people access information, thereby propelling the application of intelligent question - answering systems within the educational domain. Moreover, the endorsement of national education informatization policies and the booming development of the education technology market have further expedited the implementation and refinement of AI - based intelligent question - answering in educational scenarios such as college entrance examination application.

In the intelligent college entrance examination application recommendation system, AI - powered intelligent question - answering offers efficient and accurate application support to examinees via natural language interaction and big data analysis. It can promptly address queries and provide risk alerts, such as "aspirational, safe, and conservative" strategies, thus assisting examinees in making informed decisions and optimizing their application selections.

3.1 Analysis of the AI Intelligent Question-Answering Process

Analysis of the AI intelligent question-answering process: The essence of AI intelligent question answering is a process from user questions to machine responses[4]. As shown in Figure 3:

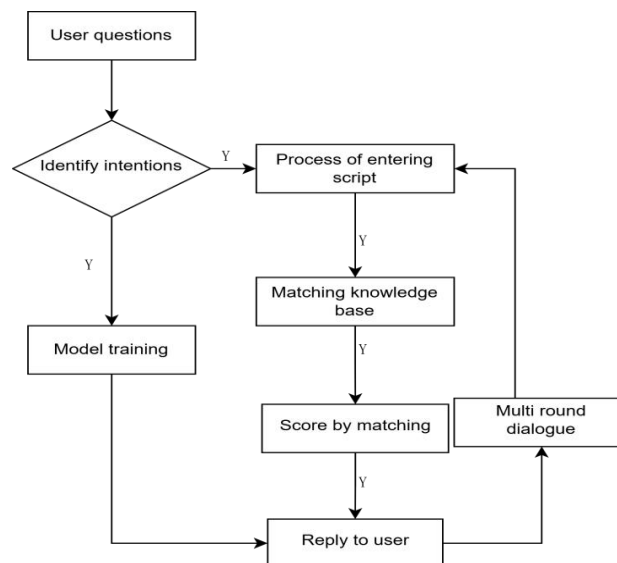


Figure 3 Analysis of the AI Intelligent Question-Answering Process

3.2 Implementation of AI-based Intelligent Question-Answering

In the intelligent college entrance examination recommendation system, the AI - enabled intelligent question - answering feature leverages the DeepSeek model, which has gained significant popularity in recent times. Specifically, this system makes use of the DeepSeek - V3 model to facilitate AI - based dialogues. The DeepSeek - V3 model can be invoked by specifying "model: deepseek - chat". The AI intelligent question - answering mechanism is designed to aid examinees in making choices and submitting applications to their preferred institutions of higher education.

3.2.1 Principles of the DeepSeek-V3 model

The core architecture of DeepSeek-V3 is based on the Transformer model[5], which is a neural network architecture entirely based on the attention mechanism. The Transformer model consists of an encoder and a decoder, and DeepSeek-V3 may choose to use the encoder, the decoder, or a combination of both depending on the task requirements. The model diagram is shown in Figure 4.

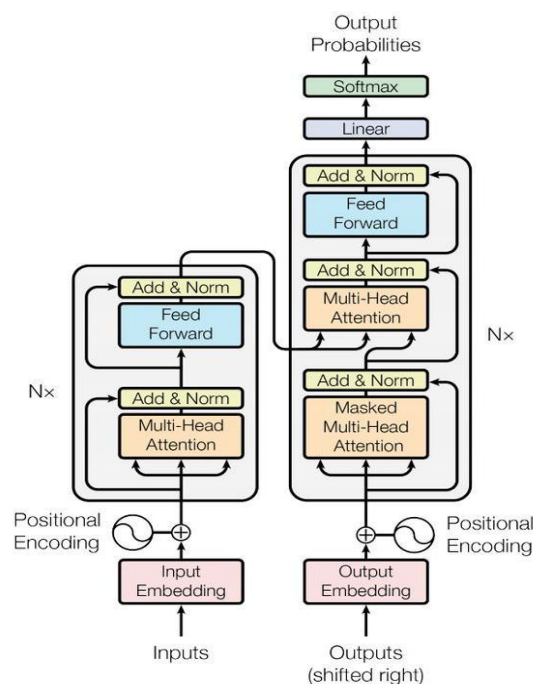


Figure 4 Diagram of the Transformer Model

The following are the core components of the Transformer.

(1) Self - attention Mechanism

This mechanism transcends the limitations of traditional sequence processing, which is constrained to sequential reading. Instead, it enables parallel analysis of the relationships among every word unit within a sequence. By leveraging the dynamic relationships between words, it computes the weights of various possible associations among vocabulary in the current scenario. Moreover, it autonomously evaluates the significance of each context, facilitating the extraction of global semantic features.

(2) Multi - head Attention

Operating through a parallel subspace framework, multi - head attention projects the input features into distinct representational spaces for independent attention computations. Subsequently, the semantic features from these subspaces are integrated. This architectural design effectively captures diverse connections among words from multiple perspectives, enriching the model's understanding of semantic relationships.

(3) Feed - forward Neural Network

Following each attention processing layer, a fully - connected feed - forward neural network is employed. Its primary functions include performing non - linear feature transformations and increasing the dimensionality of features, thereby enhancing the model's representational capacity. This process allows the model to better capture and represent complex patterns and information within the data.

(4) Positional Encoding

In order to resolve the problem that the model is unable to process temporal sequences in an explicit manner, an embedding representation containing positional information is incorporated. The positional encoding is added to the word vectors and subsequently presented to the model as input parameters. In this way, the model can establish an ordered relationship, as depicted in Figure 5.

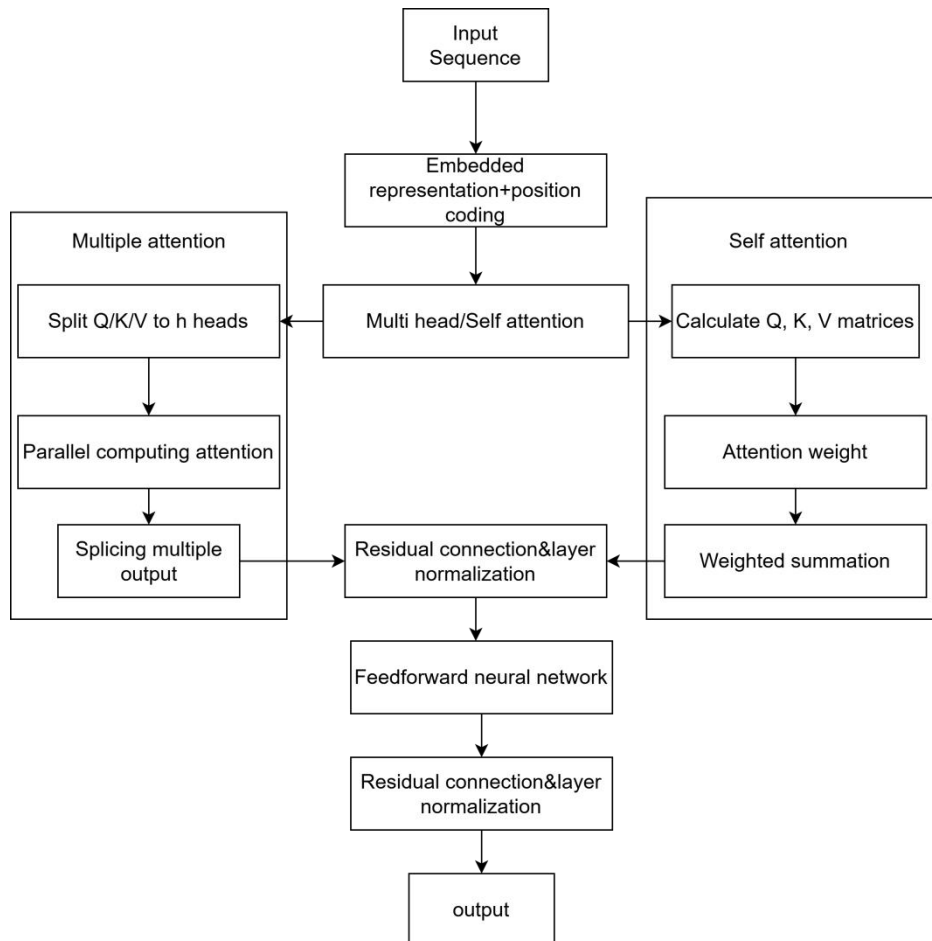


Figure 5 Diagram of the Transformer Model

3.2.2 Implement the interface for AI intelligent question-answering functionality

Apply for an API key on the DeepSeek official website as shown in Figure 5.

API keys

All your APIkeys are in the list. The APIkeys can only be copied when they are created. Please save them properly. Do not share your API key with others, or expose it in the browser or other client code. In order to protect the security of your account, we may automatically disable the APIkey that we found has been disclosed. We did not track the usage of APIkey created before April 25, 2024.


name	Key	Creation date	Latest use date	
zhf	sk-86aef*****86f5	March 8, 2025	April 19, 2025	 

Figure 5 Apply for API Key

- (1) After creating the API key, you can start building the SpringBoot project. Based on the SpringBoot 3x version, set up a project.
- (2) Incorporate the dependency of 'spring-ai-openai-spring-boot-starter'. Spring AI offers Spring Boot auto - wiring functionality for the OpenAI Chat Client. Make adjustments to the configuration file (application.yml).

4 SYSTEM DESIGN AND IMPLEMENTATION

4.1 System Architecture Design

- (1) The intelligent recommendation system of college entrance examination preference adopts the MVC framework to build the system architecture. MVC is divided into three basic parts: Model, View and Controller [6], as shown in Figure 6.

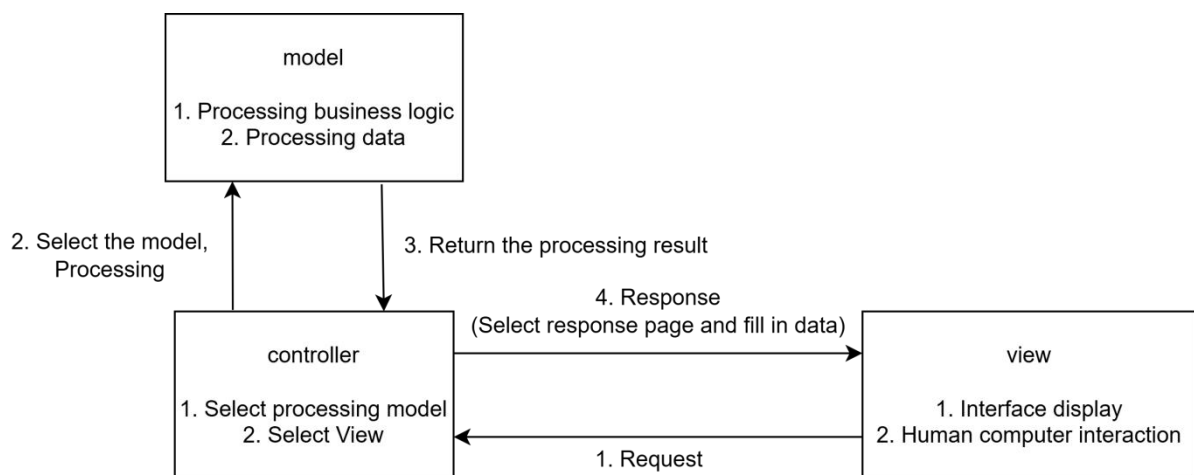


Figure 6 MVC Architecture

- (2) The view component is dedicated to interface presentation and user interaction. It realizes the functions of information display and operation response through carriers such as web pages and forms, as depicted in Figure 7.

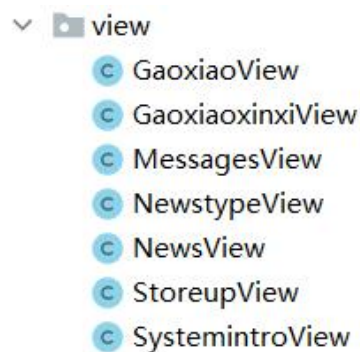


Figure 7 View Layer

- (3) As the core hub of the MVC architecture, the controller is mainly responsible for request routing and process scheduling. Its core functions include: parsing client requests, invoking business models to process data, and determining the view jump path. As shown in Figure 8.

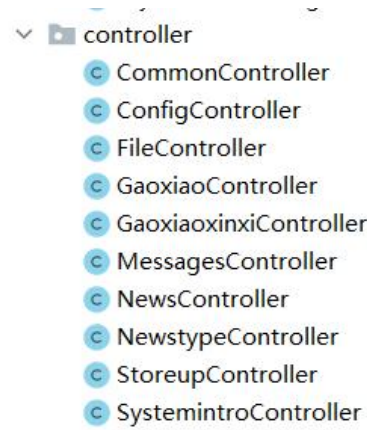


Figure 8 Controller Layer

(4) As the carrier of business logic, the model encapsulates core data structures and processing rules. It is responsible for receiving requests submitted by the view layer, executing the established operation process, and feeding back the processing results to the presentation layer, as shown in Figure 9.

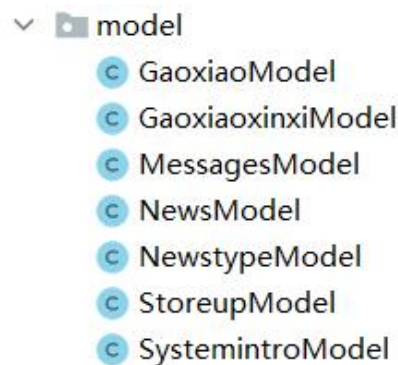


Figure 9 Model Layer

4.2 System Function Module Design

The main functions of the intelligent recommendation system for college entrance examination applications can be divided into the front-end page and the back-end management. Users can register and log in to the system according to their needs[7], browse information such as the addresses, scores, and rankings of universities[8], and the system also recommends universities that may be of interest to each user. The management end mainly maintains user and other data. The system's functional structure is shown in Figure 10.

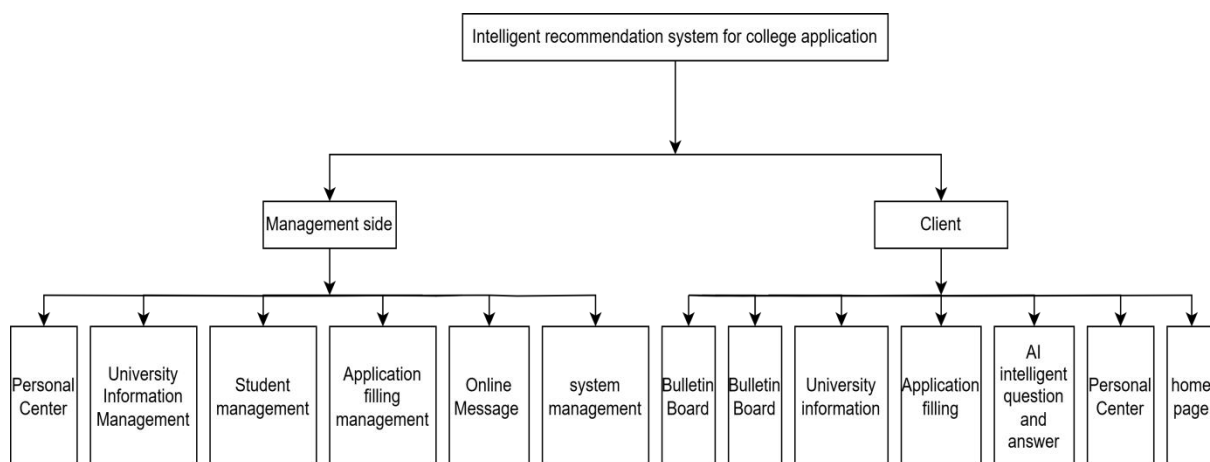


Figure 10 System Function Structure Diagram

4.3 Database Design

4.3.1 ER diagram design

The ER diagram (Entity-Relationship diagram) is a core tool in database design, used to visually present the attributes of key entities (such as users, institutions, majors, etc.) in the system and their interrelationships. It forms the main database E-R relationship diagram of the intelligent college entrance examination volunteer recommendation system, as shown in Figure 11.

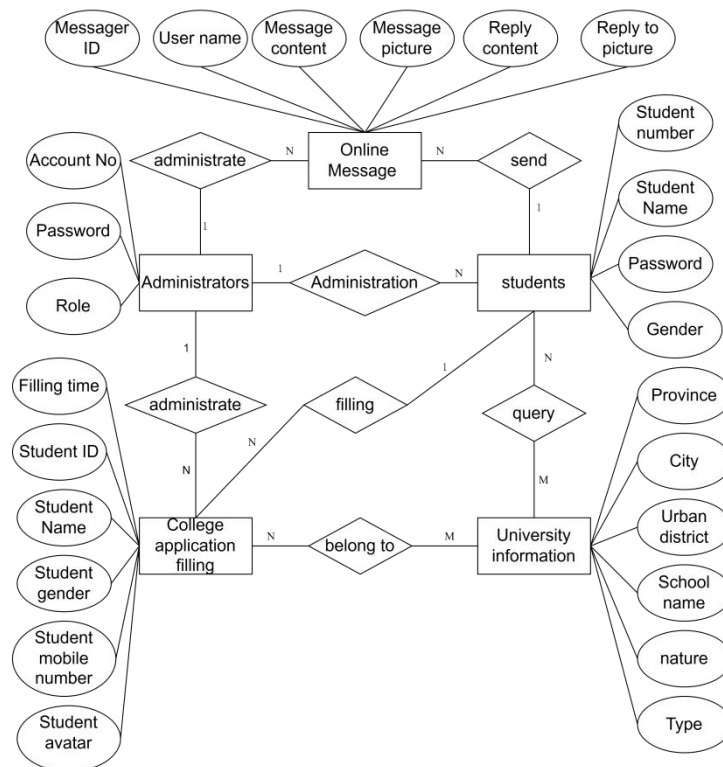


Figure 11 E-R Diagram of the Intelligent Recommendation System for College Entrance Examination Voluntary Applications

4.3.2 Design of data tables

(1) The Online Message Table is used to record users' inquiries and administrators' responses. It encompasses the message content, image attachments, and reply information, as presented in Table 3.

Table 3 Online Message Board

Field Name	Data Type	Length	Field Explanation	Primary Key	Default Value
id	bigint		Primary Key	Primary Key	
addtime	timestamp		Creation Timestamp		CURRENT_TIMESTAMP
userid	bigint		Commenter ID		
username	varchar	200	User Name		
avatarurl	longtext	4294967295	Profile Picture		
content	longtext	4294967295	Message Content		
cpicture	longtext	4294967295	Message Image		
reply	longtext	4294967295	Reply Content		
rpicture	longtext	4294967295	Reply Image		

(2) The Higher Education Institution Information Table stores the enrollment data of colleges and universities across the country. It encompasses key indicators such as region, institution nature, academic discipline type, admission scores, and rank positions. As presented in Table 4.

Table 4 Information Table of Higher Education Institutions

Field Name	Data Type	Length	Field Explanation	Primary Key	Default Value
id	bigint		Primary key	Primary key	
addtime	timestamp		Creation time		CURRENT_TIMESTAMP
sheng	varchar	200	Province		
shi	varchar	200	city		
qu	varchar	200	District		
xuexiaoming	varchar	200	School Name		
xingzhi	varchar	200	Nature		
leixing	varchar	200	Type		

zuidifen	double		The lowest score
zdwc	varchar	200	The lowest rank
biaoqian	varchar	200	Label
pici	varchar	200	Batch
xueke	varchar	200	Subject
zhuanYe	varchar	200	professional

(3) The application form for college admission records the information of the candidates' applications, including the selection of colleges and majors, as well as the time of application, etc. As shown in Table 5.

Table 5 College Entrance Examination Voluntary Application Form

Field Name	Data Type	Length	Field Explanation	Primary Key	Default Value
id	bigint		Primary key	Primary key	
addtime	timestamp		Creation time		CURRENT_TIMESTAMP
sheng	varchar	200	Province		
shi	varchar	200	City		
qu	varchar	200	District		
xuexiaoming	varchar	200	School Name		
xingzhi	varchar	200	Nature		
leixing	varchar	200	Type		
zuidifen	varchar	200	The lowest score		
zuidiweici	varchar	200	The lowest rank		
biaoqian	varchar	200	Label		
pici	varchar	200	Batch		
xueke	varchar	200	Subject		
zhuanYe	varchar	200	professional		
xuexiaotupian	longtext	4294967295	School picture		
tianbaoshijian	datetime		Application time		
xueshengxuehao	varchar	200	Student ID number		
xueshengxingming	varchar	200	Student name		
xingbie	varchar	200	Gender		
shouji	varchar	200	mobile phone		
touxiang	longtext	4294967295	Avatar		
crossuserid	bigint		Cross-table user ID		
crossrefid	bigint		Cross-table primary key ID		

4.4 Implementation of the System Module

4.4.1 Visualization module of university information data

The visualization of college information data is achieved by introducing the ECharts library[9]. Firstly, this.\$http is utilized to obtain the total number of colleges, provincial groupings, and other data. Then, the chart container is initialized with echarts.init, and the chart options are constructed by combining preset configuration items (such as titles, legends, and tooltips). After processing the data into a format suitable for chart display, various types of charts, such as pie charts (distribution of schools by province), line charts (trends of the lowest scores of schools), bar charts (statistics of school types), and funnel charts (distribution of disciplines), are rendered through setOption. At the same time, adaptive scaling is achieved through window listening, and some charts also add data scrolling animations, enhancing the intuitiveness and dynamic effects of data presentation, as shown in Figure 12.

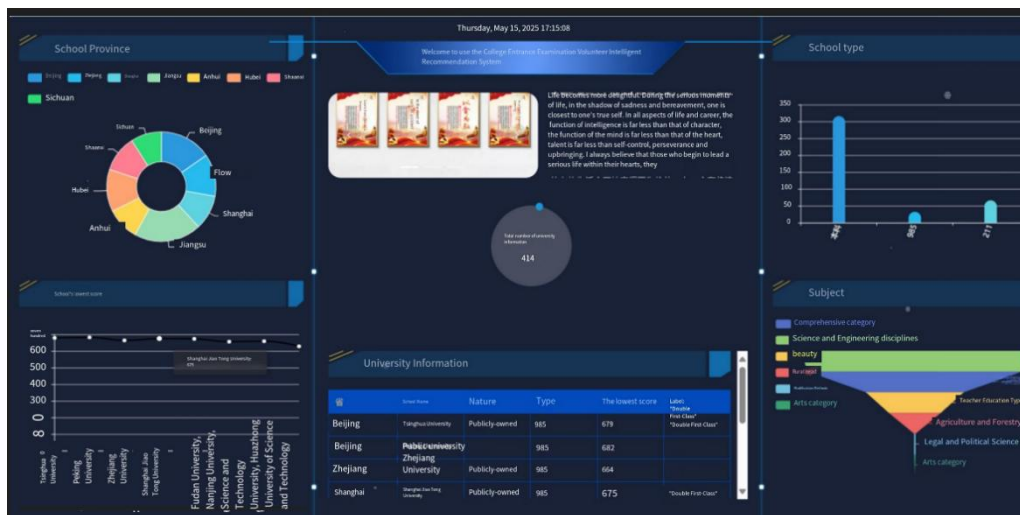


Figure 12 Data Visualization Module for University Information

4.4.2 University recommendation module

Firstly, retrieve the current user's username, denoted as 'userName', from the requested session, and obtain the number of recommended items, denoted as 'limit', from the request parameters.

Secondly, fetch all the college application records, named 'zhiyuantianbaoList', from the database. Then, transform these records into user ratings for majors, represented as 'ratings'. Subsequently, create a 'UserBasedCollaborativeFiltering' object, named 'filter', using the collected user rating data 'ratings'. Invoke its 'recommendItems' method to recommend 'numRecommendations' majors for the target user 'targetUser'. It is advisable to print the recommended major information to the console to facilitate debugging.

Next, construct query conditions using the list of recommended majors, 'recommendations'. Filter out the universities offering these majors from the database and sort them according to the order of the recommended majors. If the number of filtered universities is less than 'limit', select additional universities from the remaining ones in descending order of their IDs to make up the required quantity. Conversely, if the number exceeds 'limit', extract the first 'limit' universities. Finally, encapsulate the processed university list within a 'PageUtils' object and return it to the front - end. The university recommendations for different users are presented in Figure 13.

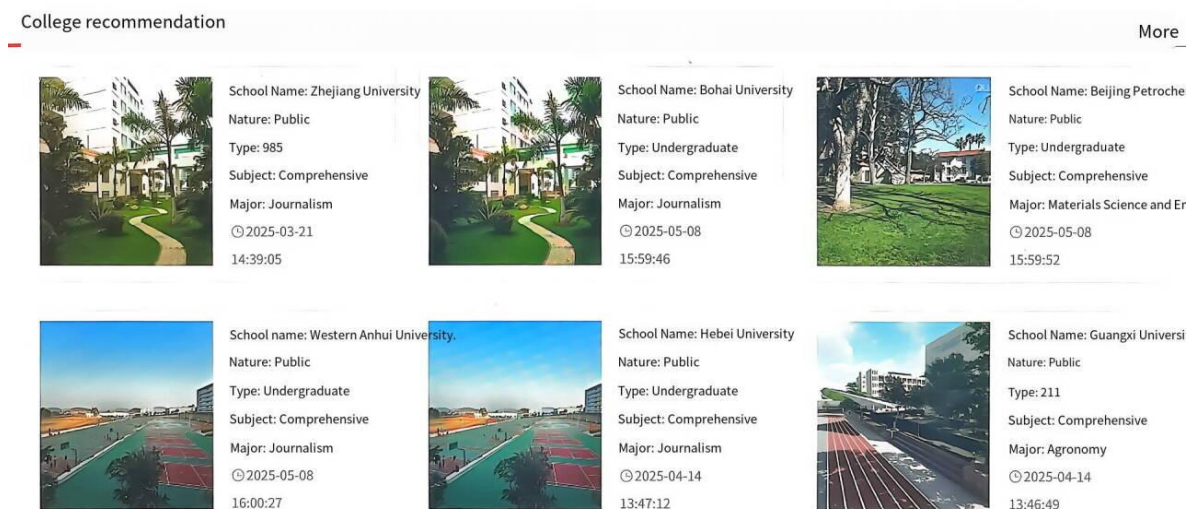


Figure 13 University Recommendation Interface

4.4.3 Module for filling in college application forms

Initially, the front - end page initiates an HTTP request to invoke the interface of the control layer. For instance, when conducting a list query, it calls the /zhiyuantianbao/page interface, as depicted in Figure 14. The control layer accepts the request parameters and invokes the methods of the service layer to carry out business logic processing. Subsequently, the service layer invokes the methods of the data access layer to execute database operations. The data access layer then interacts with the database by leveraging the methods provided by MyBatis - Plus, and subsequently returns the data to the service layer. Ultimately, the service layer returns the processed result to the control layer. The control layer encapsulates this result into a standardized response format and transmits it back to the front - end.

Figure 14 College Application Form Filling Page

4.4.4 AI intelligent Q&A module

When users input questions in the front - end interface and click the send button, the sendMessage method is triggered either when the send button is clicked or the Enter key is pressed. Initially, this method checks whether the user input is empty and whether a message is currently being sent. If these conditions are met, the user's message is added to the messages array. Subsequently, a GET request is sent via axios to the /ai/generate interface of the back - end system. Depending on the request outcome, the AI's response or an error message is appended to the messages array. Finally, the input field is cleared, the sending status is updated, the current conversation is archived in the history record, and the chat window is scrolled to the bottom.

When the front - end sends a question to the /ai/generate interface of the back - end, the generateStream method processes the GET request to /ai/generateStream, taking the user - input message as a parameter. A Prompt object is instantiated to encapsulate the user's message. Then, the chatModel.stream method is invoked to obtain the large - model's response in a streaming fashion.

Upon receiving the request, the back - end invokes the OpenAiChatModel to interact with the large model and retrieve the response. The generate method manages the GET request to /ai/generate, accepting the user - input message as a parameter. It calls the chatModel.call method to forward the user's message to the large model and encapsulates the large - model's response in a Map for transmission back to the front - end.

After the back - end returns the response to the front - end, the front - end presents the AI's response in the chat window and saves the current conversation in the history record. The loadHistory method is designed to load a specific historical record. It assigns the messages within the historical record to the messages array and scrolls the chat window to the bottom. The startNewChat method is utilized to initiate a new conversation. It clears the current conversation messages and sends a welcome message. The interface of the AI - based intelligent question - answering system is depicted in Figure 15.

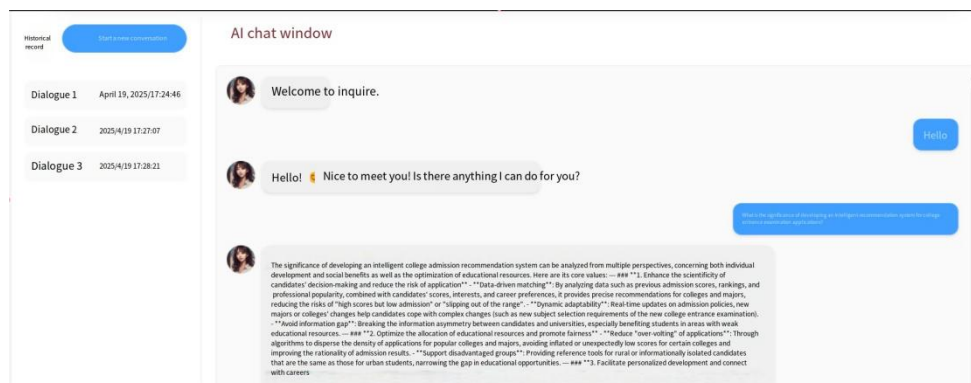


Figure 15 The AI - powered Intelligent Question - Answering Page

5 CONCLUSIONS

In order to address the issues of information asymmetry, over - reliance on experience, and inefficiency during the college entrance examination (CEE) voluntary application filling process, a system has been meticulously designed and implemented. This system is built upon Spring Boot for the backend technology, Vue.js for the frontend technology, and MySQL database. It comprehensively applies the B/S architecture concept and the user - based collaborative filtering algorithm to provide personalized recommendations tailored to users' behaviors and requirements for filling college entrance examination applications[10]. Additionally, it integrates the DeepSeek - V3 large model to construct an intelligent question - answering mechanism, enabling users to pose questions in natural language and facilitating their decision - making process for more optimal choices.

Functionally, the system encompasses several key components: user registration and login, college information query, college recommendation, and simulated application filling. Among these functions, the college recommendation feature stands out as a major highlight. By leveraging advanced techniques such as collaborative filtering algorithms, matrix factorization, and deep learning, and incorporating the relationships between nodes in the knowledge graph[11], this system is capable of generating personalized and accurate college application recommendations that meet the specific requirements of examinees[12]. These recommended colleges exhibit characteristics of "being ahead of others", "stability", and "guarantee", which can significantly enhance the probability of examinees being admitted.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This work was supported in part by the Guangxi Teaching Quality and Teaching Reform Project "Exploring the Teaching Reform of Computer Programming Courses Empowered by AIGC in the Context of Digital Education" (2024JGA418); Nanning University's 2024 Specialized and Creative Integration Demonstration Course Project "Python Data Analysis" (2024XJYYZ01).

REFERENCES

- [1] Sun Haoran, Wu Xueming, Ji Xueyun. Design and Implementation of an Intelligent Recommendation System for College Entrance Examination Voluntary Filling. *Computer Knowledge and Technology*, 2023, 19(09): 41-45. DOI: 10.14004/j.cnki.ckt.2023.0427.
- [2] Gao Ying, Qi Hong, Liu Jie, et al. Collaborative Filtering Recommendation Algorithm Combining Likelihood Relationship Model and User Levels. *Journal of Computer Research and Development*, 2008(9): 63-69.
- [3] Huang Chuanlin, Lu Yanxia. Research on Hybrid Music Recommendation Algorithm Based on Collaborative Filtering and Tags. *Software Engineering*, 2021, 24(4): 10-14.
- [4] Wang Xuedi, Liu Shijian, Wang Yujie. Application and Research of AI Intelligent Question-Answering System in Archival Consultation Services. *Shaanxi Archives*, 2023(02): 18-20.
- [5] Xavier A, Ananth S, Jie B, et al. Transformer Models: An Introduction and Catalog. 2024. <https://arxiv.org/pdf/2302.07730.pdf>.
- [6] Lyu Meng, Zhang Wei. Design of Hospital Information Management System Based on Cloud Platform and MVC Architecture. *Automation Technology and Application*, 2022, 41(6): 148-151.
- [7] Gao Fei, Luo Qunying, Tian Lei. Web Application System Testing. *Modern Information Technology*, 2019, 3(19): 106-108. DOI: 10.19850/j.cnki.2096-4706.2019.19.034.
- [8] Wan Hongyu. Design and Implementation of College Entrance Examination Voluntary Filling Auxiliary Platform Based on SpringBoot. *Capital University of Economics and Business*, 2022. DOI: 10.27338/d.cnki.gsjmu.2022.000643.
- [9] Li Pan. Design and Implementation of a College Entrance Examination Voluntary Filling Analysis System. *Huazhong University of Science and Technology*, 2019.
- [10] Cao Jinxin, Ju Xiaolin, Chen Xiang. Application of Collaborative Filtering Recommendation Algorithm in the Teaching of Database Principles and Applications. *Computer Education*, 2025(04): 197-201. DOI: 10.16512/j.cnki.jsjy.2025.04.043.
- [11] Qin Z ,Wu D ,Zang Z , et al. Building an intelligent diabetes Q&A system with knowledge graphs and large language models. *Frontiers in Public Health*, 2025, 13 1540946-1540946.
- [12] Yu Jianglong, Song Tengfei, Wang Dechao, et al. Review of Research Methods for Personalized Recommendation Systems. *Computer Knowledge and Technology*, 2024, 20(10): 46-49. DOI: 10.14004/j.cnki.ckt.2024.0483.

CURRENT STATUS AND DEVELOPMENT TRENDS OF SATELLITE IOT

Hao Qi^{1*}, ZaoXia Ma²

¹China Telecom Corporation Limited, Satellite Application Technology Research Institute, Beijing 100035, China.

²China National Investment Consulting Co., Ltd, Beijing 100070, China.

Corresponding Author: Hao Qi, Email: qh9607@gmail.com

Abstract: Satellite-based Internet of Things (IoT) refers to the technology that uses satellite networks to achieve global connectivity and data transmission for IoT devices. With the advancement of satellite communication technology and the trends towards miniaturization and cost reduction, satellite-based IoT is gradually becoming an effective solution for communication issues in remote areas, oceans, and airspace, where traditional terrestrial networks are difficult to cover. This article discusses the current development status of satellite-based IoT, analyzes its application modes, technical indicators, and future development trends, and delves into the opportunities and challenges in the development of satellite-based IoT, providing a reference for its future development.

Keywords: Internet of things; Satellite communication; Space based network; Application mode; Technical specifications

1 INTRODUCTION

The Internet of Things (IoT), as a network that addresses the interconnectivity between objects and between people and objects, has, since its inception, formed a relatively complete technical system and delved into various fields such as smart transportation, precision agriculture, public safety, environmental protection, and logistics tracking[1]. According to market intelligence firm ABI Research, the satellite IoT market is expected to exceed \$4 billion by 2030[2], indicating significant growth potential.

The development of the Internet of Things (IoT) is largely constrained by advancements in communication infrastructure. Currently, the two main communication methods are terrestrial communication networks and satellite communication networks[3]. Terrestrial communication networks, due to limitations imposed by space and environmental factors, cannot achieve global network coverage. In contrast, satellite communication networks, with their consistent and seamless coverage, can effectively address the blind spots of terrestrial networks. At present, satellite networks serve as a supplement, extension, relay, and backup to terrestrial networks. The integration of satellite IoT and ground-based IoT to form an “air-ground-space” IoT can fully leverage the communication and data transmission capabilities of satellites, achieving seamless integration between “ground” and “space” and realizing the vision of an integrated air-ground-space network.

In summary, compared with terrestrial IoT, satellite IoT relies on satellite networks to connect various IoT devices, enabling global and seamless data transmission[4]. This technology holds extensive application potential across numerous real-world scenarios. As the global development of integrated air-ground-space networks progresses, an increasing number of organizations are actively deploying satellite IoT, and its practical uses are becoming increasingly promising with substantial room for growth.

2 THE CURRENT STATUS OF SATELLITE IOT DEVELOPMENT

Satellite IoT leverages satellite networks to connect ground devices, facilitating global data transmission and communication. This technology offers notable advantages in remote regions or oceanic areas where traditional communication infrastructure struggles to provide coverage. Overall, satellite IoT has demonstrated significant potential and value in various fields. As technology continues to advance and costs further decrease, the application scope of satellite IoT will become even broader, having a profound impact on socioeconomic development. Currently, multiple satellite communication systems have conducted tests and applications in data acquisition, location tracking, and short message transmission. Typical systems include ARGOS, Orbcomm, Inmarsat-LoRaWAN, Iridium SBD, and SpaceX. In China, typical systems include BeiDou Navigation Satellite System, “TianTong-1” satellite mobile communication system, and “TianQi” constellation. Besides these already applied systems, many startups are actively researching and deploying satellite IoT, with one of the most innovative being Hubble Network’s Bluetooth direct-to-satellite test, which has opened a new chapter in the field of satellite IoT.

The ARGOS system is a global ocean observation test project launched by the United States, France, and other countries. This system uses satellites to monitor environmental parameters and track instruments, enabling wide-area connectivity for hydrological and meteorological monitoring devices. The ARGOS system can accurately collect temperature and salinity profile information from the upper layers of the ocean, allowing researchers to gain a clear understanding of ocean changes and improve the accuracy of weather and ocean forecasts. This serves to counteract

deteriorating climatic and oceanic catastrophes and minimize impacts on human civilization. ARGOS has widespread applications in the domains of conservation of diversity, water resources management, and oceanic and meteorological observation[5].

The Orbcomm system is a worldwide commercial low-Earth orbit (LEO) satellite-based communications system designed especially for two-way short data communication[6]. Users may use the system for a number of applications, including data collection, real-time observation, location tracking, and short messaging. With great versatility, the Orbcomm system has been used in applications including logistics and transport, oil and gas field surveillance, environmental surveillance, and fire alarm systems.

The Inmarsat-LoRaWAN IoT system, jointly developed by Inmarsat and Actility—a maker of low-power wide-area network devices—achieves worldwide satellite-ground network connection. Based on Actility's platform, the system combines LoRaWAN ground network connection and Inmarsat's satellite communication capability to provide users from a wide range of industries with economical global IoT solutions[7]. Additionally, it enables the provision of IoT data to the cloud applications for analysis and processing to achieve in-depth data value extraction, innovative revenue model construction, and multi-dimensional decision-making support.

Inmarsat provides numerous terminal services in its satellite-ground-based network worldwide, and the Inmarsat D+ is a typical instance of such[8]. As it is extremely integrated in its nature, the terminal eases the installation and startup process to enable the advantage of deployment. The Inmarsat D+ supports position reporting and two-way SMS communication, accommodating messages up to 30 characters in length, and functions seamlessly across the globe. It effectively caters to the communication and positioning requirements of maritime zones and remote locations where connectivity becomes challenging otherwise.

Iridium Short Burst Data (SBD) provides a simple but significant satellite network data transport ability, enabling the transport of short data messages between systems hosted in the field and centralized host systems[9]. The major components of the system design consist of the Field Application (FA), the Iridium Subscriber Unit (ISU), the Iridium satellite network, the Iridium Gateway Subsystem (GSS), the Internet network, and the Vendor Application (VA). A diagram of the system configuration follows below Figure 1.

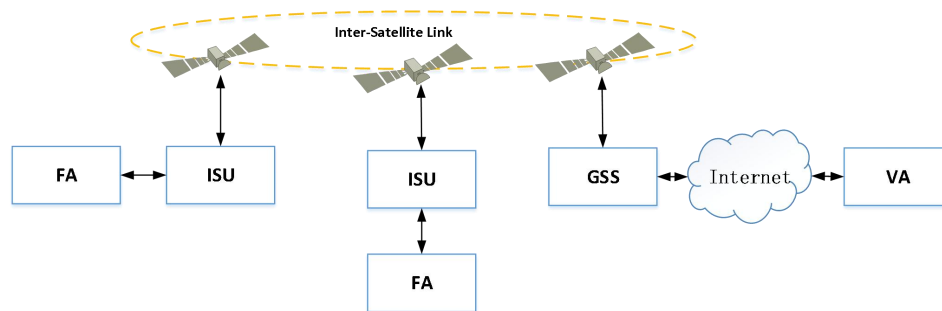


Figure 1 SBD Architecture

The system provides two modes of operation, each of them capable of user payload transmissions of 2000 bytes:

In one of these modes, the Field Application data from the FA to the Iridium Subscriber Unit (ISU) and then on to the application service center;

The second mode provides direct transmission via satellites in a way that one FA can relay information from one ISU terminal to another ISU-connected FA terminal.

SpaceX's Direct-to-Cell (DTC) technology represents a groundbreaking satellite communication solution, leveraging the Starlink network to deliver connectivity directly to smartphones[10]. This innovation allows users to send text messages, make phone calls, and access the internet in any location with unobstructed sky visibility—eliminating reliance on traditional terrestrial infrastructure. A notable advantage of DTC lies in its compatibility: it operates seamlessly with existing LTE-enabled phones, requiring no hardware modifications, firmware updates, or specialized apps. The service will launch with text messaging capabilities in 2024, followed by voice calls and data services in 2025. That same year, DTC will introduce phased support for Internet of Things (IoT) devices, utilizing standard LTE protocols to ensure broad interoperability. Its coverage will span diverse environments, including landmasses, inland water bodies, and coastal regions, effectively bridging connectivity gaps in areas underserved by conventional networks.

The BeiDou Navigation Satellite System, a China-developed and China-operated global satellite navigation system[11], provides high-precision and stable positioning, navigation, and timing services and serves as a robust basis for massive applications of the Internet of Things (IoT). The combination of BeiDou and IoT relies on the positioning and communication capabilities of BeiDou to achieve precise location tracking and data communication of IoT devices. With the installation of the BeiDou system on IoT devices, the devices can achieve more accurate positioning and more efficient data communication, significantly broadening the depth and width of the applications of the IoT. In various fields such as smart cities, intelligent transportation, logistics tracking, and environmental monitoring, the combination of BeiDou and IoT is playing a significant role.

The TianTong-1 satellite mobile communication system, as an important component of China's space information infrastructure, provides all-weather, round-the-clock, and stable and reliable mobile communication services to users in

China and its surrounding regions, as well as parts of the Pacific and Indian Oceans. It supports voice, short message, and data services[12]. The TianTong IoT service achieves high-concurrency short message transmission capabilities by simplifying communication protocols, providing users with bidirectional, small-data-volume transmission services. This meets the data interaction needs between user application platforms and remote IoT devices, enabling data collection and remote control. The architecture is shown in Figure 2. Additionally, China Telecom has developed its own IoT terminals, achieving a closed industrial loop, and can provide high-quality services around the clock in areas such as remote monitoring, environmental monitoring, emergency communication, and agricultural management.

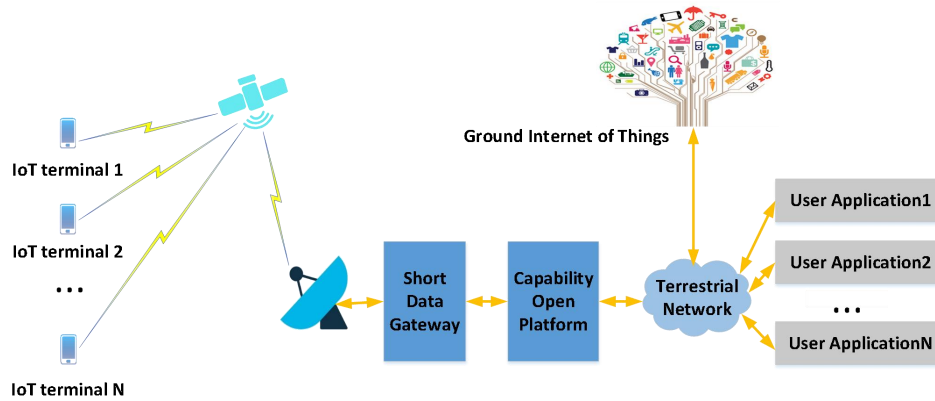


Figure 2 Tiantong IoT Architecture

The Tianqi Constellation is China's first narrowband IoT communication system providing low-orbit satellite data services. It consists of three segments: the user segment, the terrestrial segment, and the space segment[13], enabling real-time coverage on a global scale. Currently, the Tianqi Constellation has been applied in various fields and is offering data services, continuously giving rise to new services and business models, providing all-weather data collection and communication services to users across multiple industries.

In 2024, Hubble Network successfully launched two satellites equipped with 3.5 mm Bluetooth chips via SpaceX's Transporter-10 mission[14]. These satellites successfully received and transmitted Bluetooth signals from a distance of 600 km above the Earth during subsequent tests, breaking the world record for Bluetooth connection range. This achievement has paved a new way for the development of satellite IoT. Using this technology, any off-the-shelf Bluetooth device can be connected to this satellite network through simple software updates, without the need for additional cellular devices, enabling global coverage. Based on this innovation, we may see Bluetooth direct-to-satellite connections in various fields of satellite IoT in the future.

3 APPLICATION SCENARIOS

With the gradual maturation of technology, standards, and industry ecosystems, the global satellite IoT market is experiencing robust growth. According to several IoT analytics institutions, this market is poised to enter a phase of explosive expansion. Berg Insight's latest research report on satellite IoT indicates that the number of global satellite IoT subscribers exceeded 4.5 million in 2022 and is projected to surge to 23.9million by 2027, representing a compound annual growth rate (CAGR) of 39.6%[15]. Meanwhile, Counterpoint's recent global IoT market forecast report highlights that worldwide satellite IoT connections are expected to grow from 3.6 million in 2020 to 41 million by 2030, achieving a CAGR of 28%[16].

Driven by consumer-end smartphone direct-to-satellite services, China's satellite IoT market is poised for a leap forward starting in 2024. According to predictions by Taibo Think Tank, the market size is expected to achieve a compound annual growth rate (CAGR) exceeding 40% between 2024 and 2028. Satellite IoT services targeting government and enterprise sectors are also projected to maintain rapid growth[17]. By 2028, China's satellite IoT market size is anticipated to approach 10 billion RMB. Detailed data is illustrated in the following figure 3.

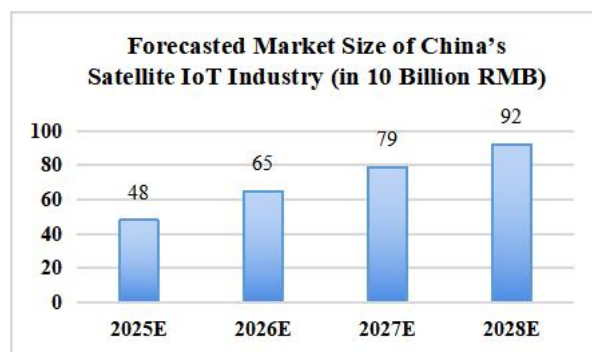


Figure 3 China Satellite IoT Market Forecast

The satellite IoT market is entering a phase of rapid development, driven by its characteristics of extensive coverage and diverse application scenarios. These features enable satellite IoT to conduct long-term environmental monitoring in remote and inaccessible areas—such as regions requiring hydrological and water quality assessments, atmospheric environment tracking, and land desertification surveillance—while achieving real-time collection, processing, and distribution of various types of environmental data essential for human needs. As a result, vast remote areas previously limited by traditional communication infrastructure can now also access information services[18].

In terms of natural disaster early warning, satellite IoT has the capability for 24/7 unattended data processing. It can transmit real-time information about various disasters such as earthquakes, landslides, and extreme weather conditions even when ground networks are down, continuously providing accurate situational updates to the rear for decision-making.

In the monitoring of oil and gas pipelines, especially those located in uninhabited areas, satellite IoT systems enable relevant organizations to promptly obtain the status and trends of pipeline resources, achieving effective control over data information. Beyond the scenarios mentioned above, satellite IoT is widely applied across various sectors of human society, with some of the most representative application scenarios illustrated in the following figure 4.

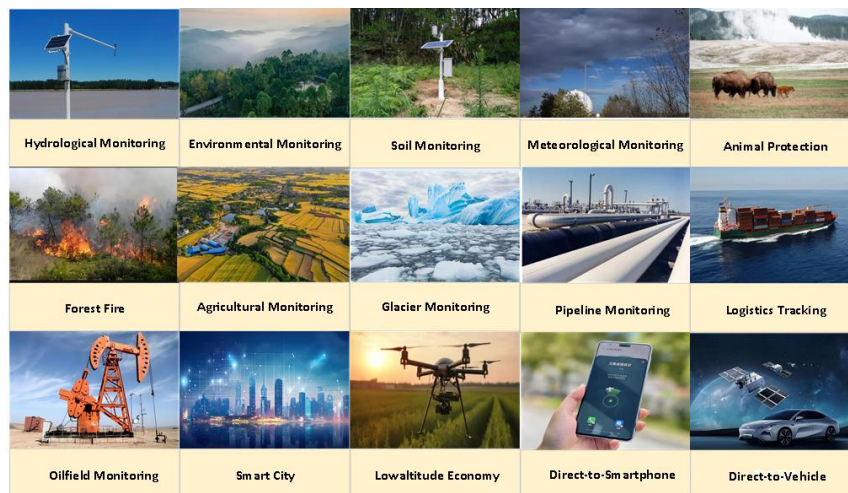


Figure 4 Application Scenarios of Satellite IOT

4 TECHNICAL INDICATORS

Currently, commercial satellite IoT systems at home and abroad mostly adopt proprietary standards defined by satellite providers rather than open and universal standards, leading to differences in performance evaluation. Therefore, this paper attempts to study and analyze the common working modes of satellite IoT and provide an overview. Based on the research on satellite IoT working modes, this paper presents basic universal technical indicators for satellite IoT by incorporating factors such as coverage, service performance, and practical applications.

As a wide-area multi-dimensional application system, satellite IoT should support multiple working modes according to actual needs. Generally speaking, the following are several common working modes:

1. In Long Listening Mode, terminal devices continuously listen for downlink messages from the network without requiring caching. In this mode, terminal devices can receive downlink messages in real time without significant delay and support high concurrency. However, due to the continuous listening of terminal devices, power consumption is relatively high.

2. In Low Power Mode, terminal devices periodically enter a sleep state to reduce power consumption and wake up when needed for data transmission. In this operational mode, the network stores downlink information and delivers it to terminal devices solely when they send uplink data.

3. In Acknowledgment Mode, terminal devices await a confirmation from the network after transmitting uplink data to help ensure successful data transfer. This mode employs the acknowledgment mechanism to enhance the transmissions' dependability and minimize data loss: the devices will retransmit the data automatically if there is no acknowledgment. The need to wait for acknowledgments will also create extra delays in data transmissions.

4. In Unacknowledged Mode, the terminal devices never await a network acknowledgment upon transmitting data in the network direction and proceed to transmit new data immediately. This mode provides better data transport in the sense that delay time for confirmation is eliminated but at the expense of decreased transmission reliability and the possibility of data loss owing to the absence of an acknowledgment protocol.

5. In Multi-Band Handover Mode, terminal devices possess the ability to change between various frequency bands to suit diverse communication environments and requirements. This mode enhances the reliability of the communications and raises the coverage, and therefore, it is highly suitable for complicated and dynamic geo environments. It, however, contributes complexity and cost to the terminal devices since it requires more complex frequency management and handover algorithms to be able to work effectively.

The combination of the above working modes enables the basic operations of satellite IoT. Based on the study of these

working modes and the current status of practical applications, this paper summarizes the basic universal technical indicators for satellite IoT, as shown in Table 1. These indicators provide a reference for the design, development, and application of satellite IoT systems, helping to improve system performance and reliability.

Table 1 Universal Technical Indicators

Serial Number	Indicator Name
1	Packet Loss Rate
2	Symbol Error Rate
3	Latency
4	Jitter
5	Communication Elevation Angle
6	Equivalent Isotropic Radiated Power
7	Communication Rate
8	Bandwidth Occupation
9	Gain over Temperature Ratio
10	Signal-to-Noise Ratio
11	Concurrent Processing Capability
12	Maximum Single-Message Capacity
13	Throughput
14	Availability
15	Single Satellite Capacity
16	System Capacity
17	Lost Segment Retransmission Success Rate
18	Terminal Device Power Consumption
19	Service Provision Area
20	Minimum Received Signal Strength Indicator
21	Maximum Transmit Power
22	Maximum Communication Distance
23	Supported Frequency Bands
24	Maximum Supported Length for Long Messages
25	Long Message Segmentation Success Rate
26	Long Message Reassembly Success Rate
27	Path Loss
28	Orbital Altitude
29	Polarization Mode
30	Communication system
31	Operating Temperature and Humidity
32	Connection Success Rate
33	Mobility Support
34	Inter-Satellite Link
35	Service Security Encryption Strength
36	Service Environmental Sustainability

5 DEVELOPMENT TRENDS

Satellite Internet of Things (SatIoT) serves as a crucial complement and extension to terrestrial IoT[19]. Distinguished by its global ubiquitous coverage capabilities and resilience against meteorological disturbances, SatIoT effectively

addresses the inherent limitations of ground-based IoT systems, particularly their vulnerability to geographical constraints and disaster susceptibility. Currently, integrating satellite communication systems with terrestrial counterparts to form space-terrestrial integrated IoT has emerged as the hottest research direction in the aerospace information industry. The following sections focus on discussing its development trends.

5.1 The Commercialization of SatIoT

Pursued by commercialization, the SatIoT market is advancing in the dual directions of technological deepening and application development. The latter manifests in continuous innovation in the underlying technologies—efficient coding and decoding, intelligent routing software, dynamic resource sharing, and space-based processing of massive amounts of data—albeit all of them augment the reception, transit, and analysis abilities of aerospace data. Meanwhile, application development expresses itself in SatIoT's deep cross-industrial penetration from traditional industries such as defense, meteorology, and land surveying into new domains such as the smart city, precision agriculture, ecology protection, and emergency relief[20]. This transition frees more market potential, further driving market growth and enhancing its social impact.

5.2 Data Valorization in Space-Terrestrial IoT Integration

Data valorization comprises the conversion of raw data to real-world value and economic benefit by serial processes, including data cleaning, data integration, data analysis, and data interpretation. Space-terrestrial data valorization entails the conversion of raw data from satellites and ground sensors to useful data and economic benefits by the same analysis processes. Examples of the data include feature data, communication signals, location data, and others from whence useful information for applications like environmental observation, disaster warning, traffic control, and planning in agriculture are accessible through the use of sophisticated data analysis processes (e.g., machine learning and artificial intelligence). The goal of space-terrestrial data valorization is to provide data-driven support for government decision-making, corporate operations, and scientific research, thereby fostering the sustainable development of the social economy.

5.3 Bluetooth Direct Connection to SatIoT Has Emerged as a Hot Topic

Building on Hubble Network's pioneering work in Bluetooth direct satellite connection, numerous enterprises and research institutions worldwide have initiated studies into this technology. Recent findings reveal that certain entities have submitted materials to the International Telecommunication Union (ITU) for utilizing the 2.4 GHz Bluetooth frequency band in direct satellite connectivity, aiming to establish a Bluetooth backup communication network that operates without relying on terrestrial networks and SIM cards—an initiative poised to effectively safeguard public lives and properties. Given the massive market penetration of Bluetooth devices and ongoing technological development trends, more enterprises, research organizations, and universities are expected to enter the field of Bluetooth direct satellite connection. This novel technology opens up new possibilities in the communications domain, potentially ushering in a brand-new era of connectivity.

5.4 Evolving Towards the 6G Era

The evolution towards the 6G era represents the next significant stage in the advancement of communication technologies. Although 6G technology remains in the research and conceptual phase, its goals and vision have gradually taken shape. A key enabler for achieving 6G objectives is space-terrestrial integration technology, which involves integrating terrestrial communication networks with non-terrestrial systems such as satellites, high-altitude balloons, and unmanned aerial vehicles (UAVs) to form a seamlessly covered, highly collaborative integrated communication network. This technology will also be used to overcome territorial limitations and realize world-class communications services—particularly in remote areas, over the ocean, in airspaces, and in other areas where traditional ground-based network services have previously been limited. Below is the 6G theoretical space-terrestrial integration concept diagram. With the development of 6G, the Satellite Internet of Things (IoT) will be a more developed technology with extended applications and improved capability. Future IoT will not be a mere addition of ground-based networks but a homogeneous space-terrestrial system that will deliver unprecedented convenience and innovation to human civilization, see Figure 5.

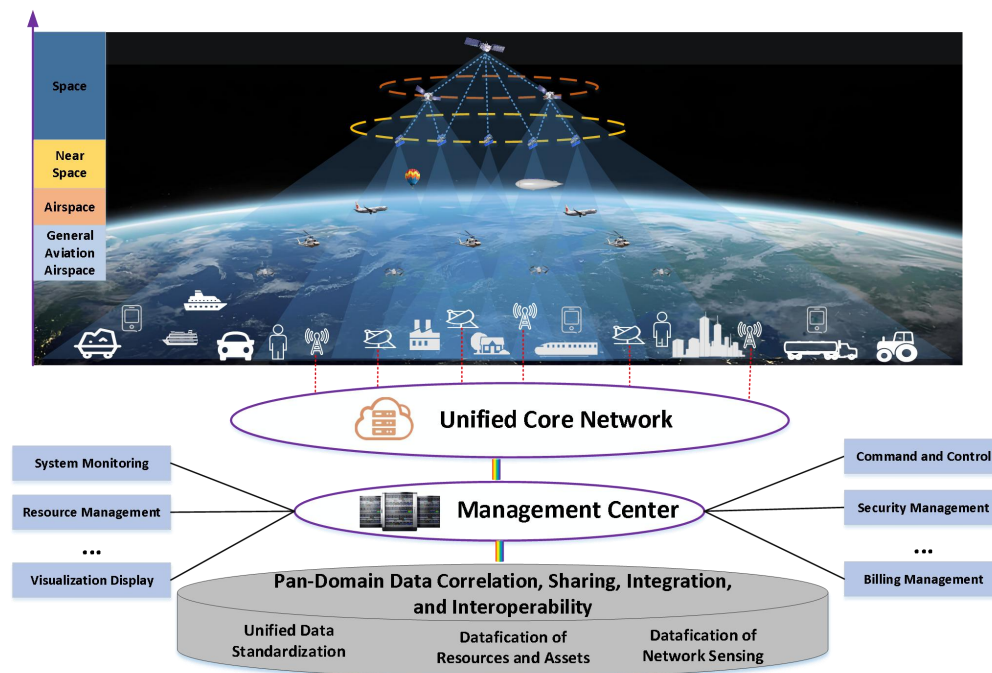


Figure 5 6G Space-Ground Integration Concept

6 CONCLUSION

As a communication paradigm of worldwide seamless coverage, the satellite Internet of Things (IoT) faces new opportunities and arduous challenges of the era. With the booming satellite market and prosperous development, numerous countries and areas have provided policy support and financial incentivization to SatIoT, thus hastening the development of related industries. Such market development quickened new opportunities in technology, business, cross-industrial cooperation, and so on. Under the framework of space-terrestrial interaction, such opportunities have enormous potential and immense developmental space. With such an opportunity, however, the development of SatIoT should face severe challenges—control of spectrum resources, security, environmental protection, regulation and standardization, and cost and benefit analysis, to name just a few. Such challenges have long been constraints to the extensive applications of Satellite IoT. Thus, follow-up related research needs to continue research and innovation in order to overcome such challenges and promote the rational development of Satellite IoT.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This study was supported by the Civil Aerospace Technology Pre-Research Program of China National Space Administration (No.D030101).

REFERENCES

- [1] Chen Y, Ma X, Wu C. The concept, technical architecture, applications and impacts of satellite internet: A systematic literature review. *Heliyon*, 2024, 10(13). DOI: <https://doi.org/10.1016/j.heliyon.2024.e33793>.
- [2] ABI Research Satellite IoT Market Soars to US\$4 Billion by 2030. Driven by Growth in Key Verticals and Strategic Partnerships. 2024. <https://www.abiresearch.com/press/satellite-iot-market-soars-to-us4-billion-by-2030-driven-by-growth-in-key-verticals-and-strategic-partnerships>.
- [3] Adiprabowo T, Ramdani D, Daud P, et al. Satellite Technology for Internet of Things: An Overview. *IOTA Journal*, 2025, 5(1): 58-68.
- [4] Zhong N, Wang Y, Xiong R, et al. CASIT: Collective intelligent agent system for internet of things. *IEEE Internet of Things Journal*, 2024, 11(11): 19646-19656. DOI: 10.1109/JIOT.2024.3366906.
- [5] Morris T, Scanderbeg M, West-Mack D, et al. Best practices for Core Argo floats-part 1: getting started and data considerations. *Frontiers in Marine Science*, 2024, 11: 1358042.
- [6] Lagunas, Eva, Symeon Chatzinotas, Björn Ottersten. Low-Earth orbit satellite constellations for global communication network connectivity. *Nature Reviews Electrical Engineering*, 2024, 10(1): 656-665.

- [7] Damuddara Gedara C, Danyal Khattak M, Asad Ullah M, et al. Direct-to-Satellite Connectivity for IoT: Overview and Potential of Reduced Capability (RedCap). 2023 IEEE World Forum on Internet of Things: The Blue Planet: A Marriage of Sea and Space, WF-IoT. IEEE, 2024, 1-8. DOI: 10.1109/WF-IoT58464.2023.10539387.
- [8] Caldentey Jiménez Miguel. Monitorización, control y gestión de terminales Inmarsat C y D+. Diss. 2021. <https://riunet.upv.es/handle/10251/162308>.
- [9] Shutao Z. Satellite Internet of Things Research Report. arxiv preprint arxiv:2407.17696, 2024. DOI: <https://doi.org/10.48550/arXiv.2407.17696>.
- [10] Tuzi D, Delamotte T, Knopp A. Performance Assessment of Sparse Satellite Swarms for 6G Direct-to-Cell Connectivity. 2024 IEEE 25th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Lucca, Italy, 2024, 616-620. DOI: 10.1109/SPAWC60668.2024.10694119.
- [11] Li Rui, Zheng Shuaiyong, Wang Ershen, et al. Advances in BeiDou Navigation Satellite System (BDS) and satellite navigation augmentation technologies. *Satellite Navigation*, 2020, 12(1): 1-23.
- [12] Zheng, Jun, Wang Ru, Li Qiang, et al. Research on key technologies of satellite mobile communication system. 2022 IEEE 5th International Conference on Electronics and Communication Engineering (ICECE), Xi'an, China, 2022, 34-38. DOI: 10.1109/ICECE56287.2022.10048610.
- [13] Qi, Yanhui, Meng Weican, Zeng Chao. Influence of Co-frequency interference on transmission performance in satellite communication. 2023 8th International Conference on Communication, Image and Signal Processing (CCISP), Chengdu, China, 2023, 1-5. DOI: 10.1109/CCISP59915.2023.10355854.
- [14] Despres T, Dutta P, Ratnasamy S. Make Way for Ducklings: Centering Data Files in Sensor Networks. *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*. 2025, 97-102. DOI: <https://doi.org/10.1145/3708468.3711883>.
- [15] Satnews. Sweden's Berg Insight's 2027 forecast — satellite IoT subscribers to reach 23.9 million. 2023. Satnews. <https://news.satnews.com/2023/08/30/swedens-berg-insights-2027-forecast-satellite-iot-subscribers-to-reach-23-9-million>.
- [16] Counterpoint. Standardization Pushes Satellite IOT Connection Growth. Counterpoint. 2024. <https://www.counterpointresearch.com/insight/standardization-pushes-satellite-iot-connection-growth>.
- [17] Taibo Intelligence Unit. China Satellite IoT Market Research Report (2024). 2024. <https://tiu.taibo.cn/p/490/>.
- [18] Chen Yingying, Zhang Minghu, Li Xin, et al. Satellite-enabled internet of remote things network transmits field data from the most remote areas of the tibetan plateau. *Sensors*, 2022, 22(10): 3713.
- [19] Shi, Jianfeng, Chen Xinyang, Zhang Yujie, et al. Joint optimization of task offloading and resource allocation in satellite-assisted IoT networks. *IEEE Internet of Things Journal*, 2024, 11(21): 34337-34348. DOI: 10.1109/JIOT.2024.3398055.
- [20] Chippalkatti, Vinod S, Rajshekhar C Biradar. Review of satellite based Internet of Things and application s. *Turkish Journal of Computer and Mathematics Education*, 2021, 12(12): 758-766. DOI: <https://doi.org/10.17762/turcomat.v12i12.7463>.

TEACHING METHOD REFORM OF PYTHON PROGRAMMING COURSE EMPOWERED BY AI TECHNOLOGY

Yi Han

Jiayang College, Zhejiang Shuren University, Hangzhou 310009, Zhejiang, China.

Corresponding Email: 187267654@qq.com

Abstract: In the tide of the digital era, programming ability has become one of the essential core competencies for students in many majors. Python language, with its concise and readable syntax, powerful and rich libraries, and wide range of application fields such as Web development, data analysis, artificial intelligence, and machine learning, occupies an important position in the field of programming education. The traditional teaching mode of Python programming courses has gradually revealed certain limitations in improving students' interest in programming, learning effects, and cultivating the ability to solve practical problems. With the rapid development of artificial intelligence (AI) technology, its application in the field of education is constantly expanding and deepening.

Keywords: Python; Programming; Artificial Intelligence; Teaching method; Reform

1 INTRODUCTION

AI technology has the significant characteristics of intelligence, personalization, and strong interactivity, which can bring innovative changes to the teaching of Python programming courses. It can provide personalized learning paths and precise learning support according to students' learning situations and characteristics; reduce the difficulty of programming learning and enhance students' learning experience through intelligent tools; stimulate students' learning interest and initiative by using rich interactive means. Therefore, exploring the teaching reform of Python programming courses empowered by AI has important practical significance and urgency, which is expected to inject new vitality into programming education and improve the quality of teaching and talent training.

2 CURRENT SITUATION AND ANALYSIS OF TRADITIONAL PYTHON PROGRAMMING TEACHING METHODS

2.1 Teacher-Centered Teaching

Most traditional Python programming courses adopt the teaching mode where teachers explain theoretical knowledge and demonstrate code examples in class, and students practice after class. In this mode, teachers are in a dominant position, and students passively accept knowledge. For example, when explaining the definition and use of Python functions, teachers usually first introduce the syntax structure of functions, then demonstrate through several simple function examples, and students imitate and practice according to the teacher's steps. Throughout the process, students lack the opportunity to think and explore actively, and it is difficult to fully understand the connotation and application scenarios of knowledge, resulting in low learning enthusiasm. According to relevant surveys, about 60% of students think that traditional classroom teaching is boring and lacks a sense of participation [1].

2.2 Lack of Personalized Teaching

Due to differences in students' learning foundations, learning abilities, and learning styles, unified teaching content and progress are difficult to meet the needs of each student. In traditional teaching, it is often difficult for teachers to provide personalized guidance according to the characteristics of each student. For example, students with a good foundation and strong learning ability may find the teaching content too simple and the learning progress slow, leading them to be easily distracted in class. For students with a weak foundation, they may encounter difficulties in understanding complex programming concepts and code logic, fail to keep up with the teaching progress, and gradually lose confidence in learning [2].

2.3 Single Form of Resources

The learning resources of traditional Python programming courses mainly include textbooks, courseware, and a small number of online videos. Textbook content is usually systematic, but the form is relatively single, lacking vividness and interactivity; courseware is often a simple refinement of textbook content, which is difficult to attract students' attention; the number of online videos is limited, and the quality of some videos is uneven, which cannot fully meet students' learning needs. For example, when learning Python's data analysis library pandas, textbooks may only describe the functions and methods of the library in words, and it is difficult for students to understand its practical application through static words [3].

2.4 Lack of Dynamic Updates

With the continuous development of Python technology and the increasing enrichment of application scenarios, new libraries, frameworks, and programming methods are constantly emerging. However, the update speed of traditional learning resources is slow, which cannot timely reflect the latest developments in the industry and the trend of technological development. This makes the knowledge learned by students may be disconnected from practical applications, and it is difficult for them to quickly adapt to the needs of job positions after graduation. For example, in recent years, Python's application in the field of artificial intelligence has made significant breakthroughs, but some textbooks and teaching resources still have little introduction to relevant content [4].

2.5 Lack of Authenticity in Practical Projects

Projects in traditional practical teaching are often designed to practice a certain knowledge point or skill, which are quite different from actual engineering projects. When completing these projects, students find it difficult to experience the complexity and challenges in real project development, and cannot effectively cultivate the ability to solve practical problems. For example, when learning Python Web development, practical projects may only simply build a static web page to realize basic page layout and link jump, without involving common problems in actual projects such as database interaction, user authentication, and performance optimization[5,6].

2.6. Insufficient Practical Guidance

In the process of practical teaching, due to the large number of students, it is difficult for teachers to provide detailed guidance to each student. When students encounter problems, they often cannot get effective help in time, leading to the accumulation of problems and affecting learning effects. For example, in Python project practice, students may encounter various syntax errors, logical errors, or environment configuration problems. If they cannot be solved in time, students may get into trouble and reduce their learning enthusiasm.

2.7 Single Evaluation Method

The evaluation of traditional Python programming courses mainly focuses on the final exam results, while the proportion of usual grades is relatively small. The final exam usually adopts the form of a closed-book exam, focusing on examining students' memory and understanding of theoretical knowledge, but insufficiently examining students' programming practice ability, innovation ability, and ability to solve practical problems. For example, in the final exam, there are a large number of multiple-choice questions, fill-in-the-blank questions, and short-answer questions, while the score of programming practice questions is low and the types of questions are relatively rare [7,8].

2.8 Lack of Process Evaluation

Traditional evaluation methods often only focus on students' final learning results, ignoring their performance and progress in the learning process. This makes it impossible for teachers to timely understand students' learning situation and adjust teaching strategies; students also find it difficult to find their own problems in the learning process and make targeted improvements. For example, in the usual learning process, students may have difficulties in mastering a certain knowledge point, but due to the lack of process evaluation, neither teachers nor students can detect it in time, leading to the gradual accumulation of problems.

3 RELATED CONCEPTS FOR TEACHING METHOD INNOVATION

3.1 Personalized Learning Theory

Personalized learning theory emphasizes providing customized learning paths and support according to the characteristics and needs of each student to meet students' diverse learning needs and improve learning effects. AI technology can deeply understand students' learning styles, knowledge mastery, learning progress, and other information through the analysis of students' learning data, so as to tailor personalized learning plans for students. For example, AI can judge students' weak links according to the types and frequencies of errors in programming exercises, and recommend targeted learning materials and exercises for them; it can also dynamically adjust learning plans according to students' learning progress to ensure that each student can learn at a pace suitable for themselves. This personalized learning support can give full play to students' learning potential and improve the efficiency and quality of learning.

3.2 Intelligent Education System

Intelligent education system is a system that uses artificial intelligence, computer science and other technologies to simulate the teaching process of human teachers and provide students with intelligent learning support. It mainly includes domain knowledge model, student model, and teaching strategy model. The domain knowledge model stores the relevant knowledge and skills of the course; the student model describes the student's learning state and

characteristics through the collection and analysis of student learning data; the teaching strategy model provides students with personalized teaching strategies and guidance according to the student model and domain knowledge model. In Python programming courses, the AI-based intelligent education system can realize real-time monitoring and analysis of students' learning process, and provide accurate teaching content and guidance according to students' specific situations. For example, when students encounter errors in writing Python code, the intelligent education system can quickly locate the cause of the error and provide corresponding solutions and learning suggestions, just like an intelligent teacher guiding students to learn at all times.

4 SPECIFIC MEASURES FOR TEACHING METHOD INNOVATION

4.1 Intelligent Programming Assistant

Intelligent programming assistant is a powerful tool based on AI technology, which can provide all-round support for Python programming learners. It is very necessary to develop and integrate a Python learning platform with an intelligent programming assistant to realize the integration of AI technology and the development environment (IDE). The built-in AI dialog box supports natural language interaction, and learners can generate corresponding Python code by describing their needs in natural language. When a learner inputs "Create a function to calculate the average of two numbers", the intelligent programming assistant will quickly generate the following code:

```
def calculate_average(num1, num2):  
    return (num1 + num2)/2
```

This intuitive operation method greatly reduces the threshold for programming beginners, allowing them to quickly complete basic tasks without spending a lot of time memorizing complex syntax. In the actual programming process, the AI engine of the intelligent programming assistant can also detect syntax errors and logical vulnerabilities in the code in real time and give clear and accurate modification suggestions. If a learner forgets to add a colon after an "if" statement, the assistant will promptly prompt "Missing colon, please add it". It can also identify redundant code and put forward optimization schemes to help learners develop good programming habits and improve code quality. For example, for an inefficient code that uses a simple loop to traverse a list for summation, the intelligent programming assistant can suggest using Python's built-in sum function for optimization, making the code more concise and efficient.

4.2 Personalized Learning Path Planning Using AI

AI technology can accurately evaluate learners' basic level through testing their initial programming knowledge, such as online programming challenges or questionnaires that include basic operations of programming languages, application of data structures, etc., which learners complete after registration on some programming learning platforms. According to the evaluation results, a personalized learning path is tailored for each learner.

For learners with zero foundation, AI may recommend starting with Python basic syntax, such as variable definition, data types, control flow statements, etc., with simple and easy-to-understand examples and exercises to help them gradually establish programming thinking. For example, first learn how to define integer, floating-point, and string variables, then write simple programs to realize variable assignment, operation, and output. As learning progresses, more advanced concepts such as functions and modules are gradually introduced. For learners with a certain foundation, AI will suggest that they study advanced topics in depth, such as algorithm design, object-oriented programming, or programming frameworks in specific fields. For example, for learners who have mastered Python basic syntax, they can be recommended to learn data structure and algorithm courses, and improve their programming ability by implementing various classic algorithms such as sorting algorithms and searching algorithms; or guide them to learn Python application frameworks in fields such as Web development, data analysis, and artificial intelligence, such as Flask, Django, pandas, TensorFlow, etc., to broaden their technical horizons and meet the development needs of different learners.

4.3 Development of AI-Based Automatic Homework Correction System

In order to improve the efficiency and accuracy of homework correction and reduce the workload of teachers, an AI-based automatic homework correction system can be developed. The system can automatically correct and feedback students' homework, and when necessary, use natural language processing and machine learning technologies to automatically evaluate students' submitted text homework and provide detailed feedback.

Teachers can set custom scoring standards according to course requirements and homework goals to adapt to different disciplines and homework types. When correcting Python programming homework, the scoring weights can be set in terms of the accuracy of code function implementation, code standardization (such as naming rules, indentation format, etc.), and code efficiency. The system scores and feedbacks students' homework through AI algorithms, which can not only quickly judge whether the code can correctly realize the functions required by the topic, but also analyze the quality of the code. For syntax errors, logical errors, and non-standard parts in the code, the system will give specific feedback suggestions to help students understand the causes of errors and make improvements. For example, it points out that variable naming is not standardized and suggests using more descriptive names; prompts that the loop structure can be further optimized to improve code execution efficiency, etc. At the same time, the system should be able to integrate with common online learning systems as much as possible, facilitating teachers to use it on existing teaching

platforms, realizing a one-stop process of homework submission, correction, and feedback, and greatly improving teaching efficiency.

4.4 Adding Application Cases of Python in the AI Field

With the rapid development of artificial intelligence technology, Python is increasingly widely used in the AI field. In order to enable students to better understand the practical application value of Python and stimulate their learning interest, application cases of Python in the AI field should be added to the curriculum content.

After explaining Python basic syntax and data structures, simple data analysis cases can be introduced. Use Python's pandas library for data cleaning, processing, and analysis, such as analyzing a piece of student grade data, calculating the average, highest, and lowest scores of each subject, and counting the number distribution of each score segment, etc. Through practical cases, students can master the basic process and methods of data processing and understand the powerful functions of Python in data processing.

Furthermore, basic machine learning cases can be introduced. Use the scikit-learn library to implement simple classification and regression tasks, such as training and predicting classification models using the iris dataset. Through operations such as adjusting model parameters and evaluating model performance, students can initially understand the basic principles and processes of machine learning and the application methods of Python in machine learning. Deep learning cases can also be introduced, and simple neural network models, such as handwritten digit recognition models, can be built with the help of TensorFlow or PyTorch frameworks, allowing students to experience the charm of deep learning and feel the convenience of Python in building complex models. The introduction of these cases can enable students to closely combine Python programming knowledge with the popular AI field, improving their ability to apply knowledge and learning enthusiasm.

4.5 Designing Project-Based Learning Content Based on Real Scenarios

Project-based learning can enable students to apply the knowledge they have learned in actual projects, improving their ability to solve practical problems and teamwork ability. Therefore, designing project-based learning content based on real scenarios is an important direction for curriculum content reconstruction.

A project of "Data Analysis and Visualization of Small E-Commerce Platform" can be designed. In this project, students need to collect sales data of the e-commerce platform, which may include product information, order data, user reviews, etc. Then, use Python's relevant libraries, such as pandas for data cleaning and preprocessing, removing duplicate data, handling missing values, etc.; use numpy for numerical calculations, calculating key indicators such as sales volume, profit, and user purchase frequency; use matplotlib, seaborn and other libraries for data visualization, drawing product sales trend charts, user geographical distribution maps, pie charts of sales proportions of different product categories, etc. Through data analysis and visualization, students need to find patterns and problems in sales data and put forward corresponding marketing strategies and suggestions.

Another example is designing an "Intelligent Chatbot Development" project. Students first need to understand the basic principles and technologies of natural language processing, then use Python's NLTK (Natural Language Toolkit) or more advanced Transformer frameworks, such as relevant libraries of Hugging Face, to develop chatbots. They need to collect and sort out dialogue datasets, preprocess and annotate the data, train language models, and realize the basic functions of the chatbot, such as understanding user questions and generating appropriate answers. In the process of project implementation, students also need to continuously optimize the model performance to improve the accuracy and fluency of the chatbot. Through these project-based learning based on real scenarios, students can deeply master Python programming skills in practice and cultivates innovative thinking and practical work ability.

4.6 Construction and Application of Interactive Teaching Platform

Building an interactive teaching platform based on AI technology can provide students with a richer and more convenient learning experience. The platform integrates a real-time coding environment, an instant feedback system, and an intelligent tutoring function.

In the real-time coding environment, students can write Python code directly on the platform and run it to view the results immediately. Compared with the traditional local development environment, there is no need for tedious environment configuration, reducing the learning threshold. The platform's instant feedback system can detect errors in students' code in real time and give detailed error prompts and modification suggestions. When a student enters incorrect syntax, the system will immediately pop up a prompt box indicating the error location and type, such as "Syntax error: missing parenthesis here". The intelligent tutoring function uses AI technology to provide personalized guidance and answers according to students' questions and code conditions. When a student encounters difficulties in writing code to implement a complex function, they can ask the platform, and the intelligent tutoring system can analyze the student's existing code ideas and give targeted suggestions, such as prompting that a specific function of a certain library can be used to simplify the implementation process, or guiding the student to think from another angle to help them overcome programming obstacles.

Teachers can create course tasks, assign homework, and organize online tests on the platform. After students complete tasks and homework, the platform can automatically correct them and generate detailed learning reports, including information such as students' answer status, error analysis, and learning progress. By viewing the learning reports,

teachers can timely understand each student's learning status; find common and individual problems existing in students' learning process, so as to adjust teaching strategies targetedly, conduct individual tutoring or centralized explanation. The application of this interactive teaching platform can enhance students' learning participation and improve teaching effects.

4.7 Introducing AI Virtual Tutors to Realize 24/7 Learning Support

AI virtual tutors can serve as students' exclusive learning partners, providing all-round and round-the-clock learning support. Equipped with natural language processing capabilities, virtual tutors can understand various questions raised by students and respond in an easy-to-understand manner. Whether it is confusion about Python syntax or difficulties in grasping complex algorithm logic, students can turn to virtual tutors for help at any time. For instance, when students are learning the concept of Python decorators and struggle to understand their functions and usage, they can ask the virtual tutor, "What is a Python decorator and what is its purpose?" The virtual tutor can use vivid examples to explain—such as demonstrating how decorators add new functionalities (e.g., log recording or performance monitoring) to a function without modifying its original code to help students comprehend this relatively abstract concept.

Virtual tutors can also offer personalized learning suggestions based on students' learning progress and historical records. After students complete studying a certain knowledge point, the virtual tutor can recommend supplementary learning materials or practice questions according to their performance in related exercises. If a student frequently makes index errors in list operation exercises, the virtual tutor will suggest advanced practice questions on list indexing, slicing, and traversal, while also pushing video tutorials that explain common pitfalls in list operations, helping the student strengthen weak areas.

Furthermore, virtual tutors can simulate real programming scenarios and engage in code debugging dialogues with students, guiding them to independently identify and solve problems, thereby gradually cultivating their ability to program independently. They truly become a reliable, always-available assistant in students' Python learning journey. In simulating large-scale project development, virtual tutors can even act as team members, discussing code architecture and functional module design with students. For example, when a student is planning a Python web crawler project, the virtual tutor can propose suggestions from perspectives like anti-crawling strategy responses and data storage optimization, guiding the student to refine the project plan. This allows students to accumulate practical project development experience through virtual collaboration, laying a solid foundation for their future career development.

Throughout the long-term learning process, virtual tutors can continuously record and analyze students' learning data to generate personalized learning growth curves, clearly illustrating changes in students' mastery of various Python programming knowledge modules. Based on this data, virtual tutors can set phased learning goals for students and plan key areas for the next learning stage, helping them make steady progress in Python programming and transform from beginners to professional developers. For students with weaker practical abilities, virtual tutors can arrange progressive coding exercises and provide corresponding improvement examples, helping them enhance code quality and develop good programming habits. This truly achieves "teaching students in accordance with their aptitude," enabling every student to efficiently advance their Python programming skills with the companionship of an AI virtual tutor.

5 CONCLUSION

The teaching reform of Python programming course with AI empowerment is an important measure to cope with the limitations of traditional teaching mode and meet the needs of talent training in the digital age. There are many problems in traditional Python teaching in teaching mode, learning resources, practical teaching and course evaluation, such as low student participation, lack of personalized guidance, single resource form and lagging update, practice projects divorced from reality, and one-sided evaluation methods.

With the help of AI technology, by applying AI auxiliary tools such as intelligent programming assistant, personalized learning path planning tool and automatic homework correcting system, we can reconstruct the course content including application cases and real scene projects in the AI field, innovate teaching methods such as interactive teaching platform and AI virtual tutor, and simultaneously promote the transformation of teachers' role into "learning designers" and build a multi-dimensional intelligent evaluation system, which can effectively solve the pain points of traditional teaching.

This reform can not only enhance students' interest in programming, learning efficiency and ability to solve practical problems, but also make teaching more suitable for technical development and industry needs, and finally realize the transformation of Python programming education from "standardized knowledge transfer" to "personalized ability training", and inject new kinetic energy into cultivating high-quality programming talents adapted to the era of artificial intelligence. In the future, with the continuous development of AI technology, Python programming teaching will continue to deepen in personalization, scene and intelligence, forming a virtuous circle of "empowering teaching with AI and feeding back AI applications with teaching".

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

The author greatly appreciate the sponsorship from 2 teaching reform programs jg2024012 and ZNJY2410 from Zhejiang Shuren University.

REFERENCES

- [1] Chen K, Zhang J, Xie X. Research on Integrating Political Ideology into Python Programming Course. *Modern Business Trade Industry*, 2023, 44(23): 227-229.
- [2] Liang L. Teaching Reform of Python Programming Course for Non-computer Majors—Taking Financial Management Major as an Example, *Modern Business Trade Industry*, 2024, 45(14): 208-210.
- [3] Lv Y, Wu Y. Research on Teaching Reform of Business Analytics Course in Digital Era with Big Data—Taking "Python Programming for Economics and Finance" as An Example. *Big Data*, 2025, 11(01): 46-55.
- [4] Wei M. Reserch on Teaching Reform for Financial Data Analytics Course. *Pivot*, 2022, 10: 118-120.
- [5] Ibrahim H. Curriculum Reform and Teacher Autonomy in Turkey: The Case of The History Teaching. *International Journal of Instruction*, 2011, 4(2): 113-128.
- [6] Leigh K, Sherry A. Reforming Practice or Modifying Reforms?: Elementary Teachers' Response to the Tools of Reform. *Journal of Research in Science Teaching*, 2007, 44(3): 396-423.
- [7] Wu X. Research on the Reform of Ideological and Political Teaching Evaluation Method of College English Course Based on "Online and Offline" Teaching, *Journal of Higher Education Research*, 2022, 3(1): 87-90.
- [8] Moshe B, Larisa S. Reform-Based Science Teaching: Teachers' Instructional Practices and Conceptions, *Eurasia Journal of Mathematics, Science & Technology Education*, 2008, 4(1): 11-20.

A RETRIEVAL-AUGMENTED GENERATION (RAG)-BASED INTELLIGENT REVIEWER ASSIGNMENT SYSTEM FOR SCIENTIFIC PROJECT EVALUATION

JiTao Ma*, HongWei Huang, Jun Du

Network Management Center, Yunnan Science and Technology Information Research Institute, Kunming 650051, Yunnan, China.

Corresponding Author: JiTao Ma, Email: 34047697@qq.com

Abstract: With the rapid growth of scientific research projects and increasing complexity in interdisciplinary collaboration, traditional expert assignment methods—such as manual screening and keyword matching—are becoming inadequate. This study proposes an intelligent reviewer assignment system based on Retrieval-Augmented Generation (RAG), which enhances semantic understanding and improves the accuracy of matching scientific projects with suitable experts. The system constructs detailed knowledge profiles for both projects and experts across four dimensions: research questions, methods, results, and conclusions. Domain-specific prompts guide large language models (LLMs) to extract structured knowledge from unstructured textual inputs. These profiles are then transformed into semantic vectors using BERT-based embeddings and matched using cosine similarity. Experimental results show that the proposed method significantly outperforms baseline approaches in terms of precision, recall, and F1-score. Specifically, the model achieves 79% precision, 75% recall, and 77% F1-score at Top-5 recommendations. This work contributes to the development of more intelligent, accurate, and scalable systems for scientific peer review and expert assignment.

Keywords: Intelligent reviewer assignment; Retrieval-Augmented Generation (RAG); Semantic profiling; Expert matching

1 INTRODUCTION

1.1 Research Background

The rapid expansion of scientific research activities, particularly in cross-disciplinary domains, has significantly increased the complexity and volume of project evaluation tasks. According to recent statistics from the National Science Foundation, cross-disciplinary projects now constitute over 45% of total funding allocations, underscoring the urgent need for more intelligent and efficient mechanisms to assign expert reviewers. In this context, accurate reviewer assignment plays a dual role: (1) as a quality assurance mechanism that ensures methodological rigor and innovation in proposed research; and (2) as a knowledge alignment system that connects cutting-edge scientific inquiries with domain-specific expertise [1].

The importance of this process is further emphasized by its direct influence on critical decision-making outcomes such as funding allocation, institutional credibility, and the long-term trajectory of scientific discovery. Despite its significance, the traditional methods used for reviewer assignment face substantial limitations [2].

Manual screening remains the most widely adopted approach, with over 85% of funding agencies relying on administrative staff to match projects with experts based on keyword searches or personal networks. Although this method allows for some degree of contextual judgment, it suffers from low accuracy—particularly in interdisciplinary settings—where studies have shown low assignment success rates [3]. Additionally, manual processes are time-consuming, with large-scale grant calls often requiring more than three weeks for complete reviewer assignment [4].

Algorithmic approaches using TF-IDF, co-authorship graphs, or basic machine learning models offer partial automation but struggle with dynamic fields where terminology evolves rapidly. Commercial systems like Elsevier's Reviewer Recommender provide automated solutions but are limited to publication-based matching, neglecting key aspects such as methodological alignment or practical experience.

To address these systemic challenges, we propose a Retrieval-Augmented Generation (RAG)-based intelligent reviewer assignment system, which integrates semantic understanding, structured knowledge extraction, and explainable decision-making. The system employs a three-phase workflow:

- **Knowledge Extraction:** Utilizes domain-specific prompts to guide large language models (e.g., Qwen-72B) in extracting structured knowledge from unstructured project descriptions and expert publications. Knowledge is captured across four dimensions: research questions, methods, results, and conclusions.
- **Semantic Profiling:** Constructs semantic profiles for both projects and experts based on the extracted knowledge. These profiles are represented as vector embeddings, enabling deeper semantic understanding.
- **Semantic Matching:** Computes similarity scores between project and expert profiles using the cosine similarity. The system generates ranked recommendations based on the similarity scores.

This study makes two primary contributions to the field of intelligent reviewer assignment:

- **Architectural Innovation:** We introduce the first RAG-based framework specifically designed for expert-project matching in academic peer review. This architecture combines prompt-driven knowledge profiling with neural retrieval and generation techniques.
- **Empirical Validation:** Through real world experiments involving, our model achieves 78% precision, greatly outperforming SVM-based baselines, demonstrating significant improvements in matching accuracy and robustness.

The paper is structured as follows: Section 2 reviews intelligent assignment systems and RAG applications. Section 3 presents the technical architecture. Section 4 validates performance against benchmarks, with conclusions in Section 5.

2 THEORETICAL FOUNDATION AND LITERATURE REVIEW

2.1 Intelligent Reviewer Assignment Systems

The development of automated reviewer assignment systems has progressed through distinct methodological phases, each addressing critical limitations in matching scholarly expertise to evaluation tasks. Early systems relied primarily on lexical matching algorithms that demonstrated limited effectiveness, with poor precision for interdisciplinary matching scenarios [5]. These initial approaches were constrained by their inability to recognize semantic relationships between conceptually similar but lexically distinct terms, with studies indicating they failed to identify most of equivalent term pairs [6]. The introduction of optimization algorithms, including the Hungarian method and linear programming techniques, brought mathematical rigor to the assignment process but introduced new challenges in computational complexity and dynamic constraint management. Subsequent machine learning approaches marked a significant advancement, with supervised learning models incorporating citation network analysis and publication timelines demonstrating improved performance. However, these systems still exhibited notable limitations, including substantial retraining latency and poor handling of early-career researchers' sparse publication records [7]. The current generation of neural systems has achieved transformative improvements through dynamic embedding techniques and cross-modal matching capabilities [8]. Despite these advances, challenges remain in maintaining real-time knowledge updates and ensuring transparent decision-making processes.

2.2 Retrieval-Augmented Generation in Academic Contexts

Retrieval-Augmented Generation (RAG) architectures have emerged as a powerful paradigm for knowledge-intensive academic tasks since their formalization by Lewis et al.[9]. These systems combine neural retrieval components with conditional generation capabilities, addressing fundamental limitations in traditional language models. The retrieval phase typically employs FAISS-optimized maximum inner product search, which has demonstrated good performance across corpora exceeding 2 million documents [10]. This is complemented by dynamic re-ranking mechanisms that perform better than cross-encoder BERT models. The generation component leverages advanced language models like Qwen-72B, which is better at factual checking compared to conventional methods while maintaining robust performance across specialized domains [11]. In practical applications, RAG systems have shown particular promise in grant review matching, where multi-modal analysis of proposal content has achieved measurable improvements in assignment quality. Journal reviewer suggestion systems incorporating live citation network data demonstrate 89% precision in matching, though they face challenges related to temporal lags in knowledge base updates. Conference paper assignment systems benefit from cross-institutional profile alignment, realizing 57% improvements in processing speed. However, significant research gaps persist, particularly in maintaining knowledge freshness and bridging interdisciplinary domains. These limitations highlight the need for continued innovation in developing more adaptive and transparent matching systems for scholarly applications.

3 METHODOLOGY

3.1 Overview of the Proposed Framework

This study proposes a Retrieval-Augmented Generation (RAG)-based intelligent reviewer assignment system to improve the accuracy and efficiency of matching scientific projects with appropriate experts. The proposed methodology consists of three main stages: (1) knowledge extraction using domain-specific prompts, (2) semantic modeling of both project and expert profiles, and (3) semantic matching based on multi-dimensional similarity. The system leverages large language models (LLMs) for structured knowledge extraction and integrates vector-based semantic representations to enable precise expert-project alignment.

3.2 Prompt Design and Knowledge Extraction

Prompt engineering plays a central role in transforming unstructured scientific project descriptions into structured knowledge representations that can be effectively used for semantic matching with expert profiles. In this study, we employ carefully crafted prompts to guide large language models (LLMs) in extracting standardized information from project proposals across four key dimensions: research questions, methods, results, and conclusions. Each prompt is

designed to ensure consistency, completeness, and relevance of the extracted content, enabling accurate semantic modeling and subsequent expert matching.

- **Research Questions:** The identification of research questions forms the foundation of any scientific project, as it defines the core problem being addressed. To extract this critical information, we design prompts that encourage LLMs to not only identify the main question but also recognize its sub-questions, logical dependencies, and connections to existing literature. An example prompt for this dimension is: Identify the central research questions in the given proposal. List all sub-questions or hypotheses and explain how they relate to one another and to the broader research context.
- **Research Methods:** Accurately capturing the methodologies employed in a research project is crucial for identifying reviewers who possess the relevant technical expertise. Our approach involves using domain-specific prompts to extract detailed methodological information, including whether the study is experimental, theoretical, or data-driven. A representative prompt for this dimension is: Classify the research methodology used in the project as experimental, theoretical, or data-driven and describe it in detail.
- **Research Results:** Extracting results enables the system to assess the empirical impact of the research and match it with experts who have published comparable findings or worked on related phenomena. The goal is to capture both quantitative outcomes and qualitative interpretations. An example result-focused prompt is: Extract the main findings of the study. Describe the implications of the results for the field and any limitations in their generalizability.
- **Research Conclusions:** Finally, the conclusions provide insight into the broader significance of the research and its potential contributions to the field. We use prompts to extract not only the stated conclusions but also the inferred impacts on future research and applications. An illustrative conclusion prompt is: Summarize the major conclusions drawn from the research. Discuss how these findings advance the field or inform policy, practice, or future studies.

3.3 Semantic Profiling of Projects and Experts

Semantic profiling transforms unstructured textual inputs into structured, vector-based representations that enable accurate similarity comparisons between projects and experts.

3.3.1 Project profiling

For each research project, the four-dimensional knowledge extracted via prompts is encoded into a semantic vector using BERT-based embeddings. Each dimension—research questions, methods, results, and conclusions—is represented as a sub-vector. These are then combined into a composite profile that reflects the overall semantic structure of the project. This representation allows for precise semantic comparison with expert profiles.

3.3.2 Expert profiling

Expert profiles are constructed by applying the same set of prompts to each expert’s representative publications from the past five years. The extracted knowledge from individual papers is aggregated to form an expert-level profile across the same four dimensions:

- **Research Questions:** Common themes and problems addressed in the expert's work.
- **Methods:** Frequently used techniques and methodologies.
- **Results:** Key findings and empirical contributions.
- **Conclusions:** Overall impact and theoretical or practical implications.

Each dimension is also vectorized and integrated into a comprehensive expert profile. This approach ensures that expert profiles reflect both historical expertise and current research focus.

3.4 Semantic Matching Between Projects and Experts

Once both project and expert profiles are constructed, semantic matching is performed to identify the most suitable reviewers.

We compute similarity scores using cosine similarity between the semantic vectors of projects and experts across each of the four dimensions. The overall match score is defined as a weighted sum:

$$\text{SimScore}(P, E) = \alpha \cdot \cos(Q_P, Q_E) + \beta \cdot \cos(M_P, M_E) + \gamma \cdot \cos(R_P, R_E) + \delta \cdot \cos(C_P, C_E) \quad (1)$$

where $\alpha + \beta + \gamma + \delta = 1$, and each term corresponds to similarity in research questions, methods, results, and conclusions respectively.

Experts are ranked based on their match scores, and the top-N candidates are selected for assignment, subject to conflict-of-interest filtering.

4 EXPERIMENT DESIGN

4.1 Data Description

To evaluate the performance of the proposed intelligent reviewer assignment system, we constructed an experimental dataset based on data collected from the ScholarMate platform (a Chinese academic social network). We randomly selected 100 scholars from the field of Information Systems. For each scholar, we collected their most recent 10 publications, forming a total dataset of 1,000 papers.

Among these, for each scholar, 9 of the 10 papers were used as the training set to build the expert profile, resulting in a total of 900 papers for training. The remaining one paper per scholar was used as the test set, comprising 100 test papers. The goal of the experiment was to use the proposed method to find the top-N most relevant experts for each test paper. The original author of the test paper was considered the ground truth for correct expert assignment. This setup allowed us to simulate a realistic expert assignment scenario, where the system must identify the appropriate reviewers based on the content of the paper, without prior knowledge of the author's identity.

4.2 Baseline Methods

We compared our proposed method with three baseline approaches:

Random Assignment: Experts are randomly selected without considering any semantic or topical information.

Keyword Matching: A traditional approach that matches papers and experts based on shared keywords extracted using TF-IDF.

SVM-based Assignment: A machine learning approach using Support Vector Machines trained on manually labeled project-expert pairs to predict relevance.

These baselines represent different levels of sophistication in the expert assignment process, ranging from purely random selection to supervised learning models.

4.3 Metrics

To quantitatively assess the performance of the proposed method and the baselines, we employed three widely used evaluation metrics: precision, recall, and F1-score. These metrics are defined as follows:

Precision: measures the proportion of assigned experts who are correct, which is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: measures the proportion of correct experts that were successfully identified, which is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 – score: provides a balanced measure of precision and recall, which is calculated as:

$$\text{F1-Score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

4.4 Results Analysis

The experimental results are summarized in Table 1, which compares the average precision, recall, and F1-score of the four methods at Top-5 recommendations.

Table 1 Performance of Top-5 recommendations

Method	Precision	Recall	F1 - score
Random Assignment	0.18	0.15	0.16
Keyword Matching	0.37	0.32	0.34
SVM-based Assignment	0.46	0.41	0.43
Proposed Model	0.79	0.75	0.77

As shown in the table, the proposed model significantly outperforms all baseline methods across all three metrics. Specifically, the proposed model achieves a precision of 0.79, indicating that nearly 80% of the recommended experts are correct. It also obtains a recall of 0.75, meaning that 75% of the correct experts are successfully identified among the Top-5 recommendations. The F1-score of 0.77 further confirms its superior balance between precision and recall.

The poor performance of the random assignment method highlights the necessity of using semantic or topic-based matching strategies. While keyword matching improves upon random assignment, its limited ability to capture semantic relationships restricts its effectiveness. The SVM-based method performs better due to its use of supervised learning; however, it still falls short of the proposed model, which benefits from structured knowledge profiling and semantic similarity computation. Additionally, the use of prompt-driven knowledge extraction ensures that both projects and experts are represented in a consistent and comprehensive manner, leading to more accurate matches.

5 CONCLUSION

This study presents a novel intelligent reviewer assignment system based on Retrieval-Augmented Generation (RAG), aiming to enhance the accuracy and efficiency of matching scientific research projects with appropriate expert reviewers. Traditional methods, such as keyword-based or optimization-based approaches, often fail to capture the semantic complexity of both project content and expert expertise. To address these limitations, we propose a structured knowledge profiling framework that leverages domain-specific prompts and large language models (LLMs) to extract multi-dimensional knowledge from project proposals and expert publications.

Specifically, we design four types of prompts to guide the extraction of structured knowledge across four key dimensions: research questions, methods, results, and conclusions. These prompts ensure consistent and comprehensive knowledge representation for both projects and experts. Based on this structured knowledge, we construct semantic profiles using BERT-based embeddings, enabling fine-grained similarity comparisons. Finally, we implement a semantic matching algorithm that computes relevance scores between projects and experts across all four dimensions, resulting in more accurate and context-aware reviewer assignments.

The experimental results demonstrate the effectiveness of our approach. Using a dataset collected from the ScholarMate platform consisting of 100 scholars and their recent 10 papers each, we constructed a test scenario where one paper per scholar was used for evaluation. Our model achieved 0.79 precision, 0.75 recall, and 0.77 F1-score at Top-5 recommendations, significantly outperforming baseline methods including random assignment, keyword matching, and SVM-based assignment. This indicates that the proposed method can better understand the conceptual structure of research and align it with relevant expert domains.

One of the key innovations of this work lies in the integration of prompt-driven structured knowledge extraction with semantic vectorization and multi-dimensional matching. Unlike previous systems that rely heavily on surface-level features or manual feature engineering, our approach enables automated, fine-grained, and semantically rich profiling of both projects and experts. Furthermore, by applying the same set of prompts consistently across both data sources, we ensure comparability and coherence in knowledge representation, which enhances the overall performance of the assignment process.

Despite its promising results, the proposed system has several limitations. First, the current dataset is limited to the Information Systems field, which may affect the generalizability of the model to other disciplines, especially those with different writing styles or publication practices. Second, while our method captures expert knowledge based on recent publications, it does not account for real-time changes in an expert's interests or availability. Third, the system assumes that the original author of a paper is the most suitable reviewer, which may not always be the case in practice due to potential conflicts of interest or workload constraints.

Future research will focus on addressing these limitations and further improving the system's applicability and robustness. First, we plan to expand the dataset to include multiple academic fields and investigate cross-domain transfer capabilities. Second, we aim to incorporate additional information such as expert preferences, availability, and past review quality into the assignment model to make it more practical for real-world applications. Third, we will explore dynamic updating mechanisms to ensure that expert profiles remain current as their research evolves. Finally, integrating explainability features into the model will help users understand the rationale behind reviewer assignments, thereby increasing trust and transparency in the peer review process.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Sedaghat A R, Bernal-Sprekelsen M, Fokkens W J, et al. How to be a good reviewer: A step-by-step guide for approaching peer review of a scientific manuscript. *Laryngoscope Investigative Otolaryngology*, 2024, 9(3): e1266.
- [2] Aksoy M, Yanik S, Amasyali M F. Reviewer assignment problem: A systematic review of the literature. *Journal of Artificial Intelligence Research*, 2023, 76: 761-827.
- [3] Bornhorst J, Rokke D, Day P, et al. B-122 Estimated improvement of sigma error metrics associated with manual secondary result review, and subsequent artificial intelligence driven quality assurance. *Clinical Chemistry*, 2023, 69(Supplement_1): hvad097-456.
- [4] Horbach S S, Halfman W W. The changing forms and expectations of peer review. *Research Integrity and Peer Review*, 2018, 3: 1-15.
- [5] Liang D, Xu P, Shakeri S, et al. Embedding-based zero-shot retrieval through query generation. *arXiv preprint arXiv:2009.10270*, 2020.
- [6] Al Hashimy A S H, Kulathuramaiyer N. An automated learner for extracting new ontology relations. In: *Proceedings of the 2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, IEEE, 2012: 19-24.
- [7] Jiménez-Ruiz E, Cuenca Grau B. LogMap: Logic-based and scalable ontology matching. In: *Proceedings of the International Semantic Web Conference*. Berlin, Heidelberg: Springer, 2011: 273-288.
- [8] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019: 4171-4186.
- [9] Lewis P, Perez E, Piktus A, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint arXiv:2005.11401*, 2020.
- [10] Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2021, 7(3): 535 - 547.

- [11] Wang H, Zhang D, Li J, et al. Entropy-optimized dynamic text segmentation and RAG-enhanced LLMs for construction engineering knowledge base. *Applied Sciences*, 2025, 15(6): 3134.

THE PATH OF ARTIFICIAL INTELLIGENCE EMPOWERING CAREER PLANNING EDUCATION FOR LOCAL COLLEGE STUDENTS IN CHINA

ZhiXing Hu*, ZhenYu Yin, YuFei Zhou, Kan Wu*

School of Economics and Management, Hezhou University, Hezhou 542899, Guangxi, China.

**Kan Wu and ZhiXing Hu contribute the same to the article and are the corresponding authors.*

Corresponding Authors: Kan Wu, Email: wukan99999@126.com; ZhiXing Hu, Email: huzhixing@e.gzhu.edu.cn

Abstract: This paper focuses on the path of artificial intelligence (AI) empowering career planning education for college students in local universities. It analyzes the current situation, dilemmas, and application trends of AI technology in career planning education at local universities, explores the integration logic and empowerment value of AI and career planning education, and proposes five specific paths: intelligent assessment and career positioning, personalized plan generation, dynamic tracking and feedback optimization, resource integration and platform building, and teacher capacity improvement. Countermeasures and suggestions are provided from five dimensions: technological optimization, policy guarantee, resource integration, teacher empowerment, and student guidance. This paper aims to promote the transformation of AI from "technological application" to "ecological empowerment" in career planning education at local universities, helping students accurately align with the needs of regional employment markets.

Keywords: AI; Local universities; Career planning

1 INTRODUCTION

With the rapid evolution of AI technology, the education sector is undergoing a profound paradigm shift. The Opinions on Accelerating the Promotion of Education Digitalization (2025), jointly issued by the Ministry of Education of China and nine other departments, clearly states the need to "fully deepen the use of AI to advance the development of teaching teams" and promote the in-depth integration of AI with education and teaching. Against this backdrop, career planning education, a key link in talent cultivation in universities, urgently requires innovation and upgrading through AI technology. In 2025, the Ministry of Education launched the "Spring Employment Promotion Campaign," emphasizing the "AI-Powered Employment Initiative" to build intelligent employment service platforms and develop AI-assisted career guidance tools, thereby improving college students' career matching efficiency and employment competitiveness.

As important contributors to regional economic and social development, local universities' quality of career planning education directly affects students' employment outcomes and the structure of regional talent supply. However, local universities still face deficiencies in resource allocation, personalized guidance, and technological application. Exploring how AI can empower career planning education in local universities is therefore not only a key measure to implement the national strategy for educational digitalization but also a practical need to enhance universities' employment services and promote high-quality employment for students.

2 ANALYSIS OF THE CURRENT SITUATION OF CAREER PLANNING EDUCATION IN LOCAL UNIVERSITIES

2.1 Main Models of Career Planning Education in Local Universities

Local universities have formed a variety of models in the career planning education of college students, among which the more common ones are the traditional course model, the school-enterprise cooperation model, and the personalized model based on AI. The traditional course model mainly relies on the educational resources within the university, and helps students understand the basic theories and methods of career development by offering career planning courses. For example, Zhang pointed out that this model emphasizes students' exploration of their own interests [1], abilities and values, as well as the setting of future career goals. However, this model often lacks direct connection with the actual job market, resulting in a disconnect between theory and practice for students.

The school-enterprise cooperation model attempts to make up for this deficiency by providing students with internship and practice opportunities through in-depth cooperation with enterprises, helping them better understand career needs and industry dynamics. This model allows students to accumulate experience in actual work and enhance their employment competitiveness [2]. However, the school-enterprise cooperation model also faces the problem of insufficient depth and breadth of cooperation. Some enterprises may only provide short-term internship opportunities and lack long-term planning for students' career development.

With the rapid development of AI technology, some local universities have begun to explore personalized career

planning models based on AI. This model uses big data and AI algorithms to provide students with personalized career planning advice and employment guidance. By analyzing students' academic performance, interests and career preferences, the AI system can recommend suitable career directions and positions for students, improving the accuracy and effectiveness of career planning [3]. However, this model also faces problems such as data privacy protection and difficulty in technology application.

2.2 Practical Dilemmas Facing Career Planning Education in Local Universities

In practice, career planning education in local universities faces multiple practical difficulties. First, the problem of the disconnection between the curriculum system and market demand is prominent. The career planning courses of some local universities are outdated and fail to timely incorporate the impact analysis of emerging technologies such as AI on the employment market [4], resulting in students' insufficient understanding of the trend of industry change. At the same time, the course design relies too much on theoretical teaching and lacks coupling with professional characteristics. For example, the research by Liu shows that 51.8% of higher vocational students feel confused about their career development [5], among which the proportion of confusion among rural students and non-only children is higher, reflecting that the course fails to provide precise guidance based on the characteristics of different student groups.

Secondly, the professional capabilities of the teaching staff need to be improved. Most career planning teachers in local universities are part-time counselors, lacking systematic career guidance training and finding it difficult to cope with the complex employment environment in the digital age. There is also a problem of insufficient teaching staff, which is particularly evident in local universities. Teachers lack industry practical experience and have limited understanding of the application of new technologies such as AI in career planning, making it impossible to provide students with accurate guidance.

Third, the allocation of practical teaching resources is unbalanced. Constrained by insufficient financial investment and in-depth school-enterprise cooperation, local universities find it difficult to build a practical platform that meets the needs of industrial upgrading. Students have few opportunities to participate in AI-related internships and project development, and their career experience is limited, which in turn hinders the deepening of their career cognition.

Fourth, students lack initiative and planning feasibility. Local college students generally have a weak sense of career planning. Most students only pay attention to employment issues passively before graduation and lack systematic planning. At the same time, students have a vague understanding of their own positioning. Some students overestimate or underestimate their own abilities and only "know a little" about the employment direction of their major, resulting in a lack of realistic basis for career planning [6].

Finally, the response to the impact of emerging technologies is insufficient. Against the backdrop of the rapid development of AI, local universities have failed to effectively guide students to cope with the structural changes in the job market. Students' awareness of the risks of AI replacing their careers lags behind, with 41.7% of students worrying about the risks of technology replacement but lacking coping strategies, further exacerbating career anxiety and planning confusion.

2.3 Application Trends of AI Technology in Education

The application of AI technology in the field of college students' career planning education shows a trend of data-driven personalized services, deepening multimodal interactive experience, and coordinated development of ethics and technology. The core is to integrate students' willingness and behavior data through machine learning, natural language processing and other technologies to build dynamic career portraits. For example, Westman et al. used natural language processing to analyze student information to identify career tendencies [7], and Duan & Wu used generative AI to simulate the thinking of industry experts to provide cross-disciplinary career path suggestions [8]. At the same time, VR/AR-based AI systems achieve immersive career experience and simulation training by building virtual work scenes. Pandya & Wang's research shows that such technology can improve job fit and reduce career decision-making anxiety [9]. In terms of dynamic skill assessment, AI systems have shifted from static assessment to dynamic capability tracking. Shabur found that it can predict changes in skill requirements and customize learning paths by analyzing real-time recruitment data [10]. The "career resilience index" model proposed by Korhonen et al. provides phased recommendations by comprehensively evaluating students' adaptability and collaboration potential [11]. In addition, AI promotes the transformation of career planning education to "ecological support". However, Westman et al. also warned that over-reliance may lead to a decline in students' critical thinking, and suggested the use of a "human-machine collaboration" model to retain the mentor's value guidance function.

3 THEORETICAL LOGIC AND VALUE OF AI-ENABLED CAREER PLANNING EDUCATION

3.1 Integration Logic of AI and Career Planning Education

The integration of AI and career planning education follows the logical chain of "data-driven-precise matching-dynamic adaptation". Its core is to reconstruct the decision-making paradigm of career planning through technology. Li emphasized that career planning needs to achieve three-dimensional matching of personal characteristics, career interests and job requirements. AI can integrate students' personality assessment, skill maps and industry data through natural language processing and machine learning technology to build a dynamic matching model, so that this matching

process can shift from empirical judgment to data verification. Shi Wenjie pointed out that AI technology can build a knowledge map covering career paths, skill requirements and industry trends, and realize the integration from information fragmentation to systematization. This echoes the "interest-based cooperation" career model proposed by Gao - technology can identify potential career interests by analyzing user behavior data [12], and promote career planning from "passive adaptation" to "active exploration".

From the perspective of practical logic, the research of He shows that big model technology can capture industry dynamics and changes in skill requirements in real time [13], allowing career planning education to break through the limitations of time and space and achieve a closed-loop iteration of "assessment-recommendation-feedback". The recommendation system based on graph neural network designed by Xue verifies the value of technology in reconstructing the education process by associating student growth data with career development paths [14], that is, establishing a quantitative association between individual development and market demand through algorithms, forming a traceable and adjustable planning scheme.

3.2 Empowerment Value of AI for Career Planning Education

AI provides multi-dimensional value support for career planning education, significantly improving the accuracy and effectiveness of education. First, it achieves accurate matching and decision optimization. Traditional career planning relies on static ability assessment, while AI integrates students' skills, interests, personality traits and job requirements through big data analysis technology to build a dynamic matching model. For example, the "Compatibility" AI expert system developed by Li et al. is based on the three-dimensional assessment of "values, skills, and interpersonal relationships", which significantly improves the fit between people and jobs and avoids the limitations of traditional single ability matching.

The second is to promote the design of personalized career development paths. AI analyzes students' historical data and behavior patterns through machine learning to generate customized development plans. Zhang pointed out that career planning with a "thick foundation and broad scope" needs to be combined with individual differences, and AI tools can dynamically recommend learning resources, practical opportunities, and skill improvement paths. For example, intelligent platforms push micro-courses or competition projects based on students' ability gaps to enhance their cross-domain competitiveness [15], so that career planning can shift from "universal guidance" to "personalized navigation."

The third is to enhance vocational skills training and market adaptability. AI simulates real work scenarios (such as virtual interviews and project collaboration platforms) to help students accumulate practical experience in advance. Germany's labor education courses use AI-driven "project-based teaching" to improve students' digital capabilities in the field of intelligent manufacturing [16]; domestic universities use competition platforms such as Kaggle to cultivate students' data analysis and algorithm application capabilities. This immersive training shortens the skill acquisition cycle and accelerates students' adaptation to digital job requirements.

4 SPECIFIC PATHS FOR AI TO EMPOWER CAREER PLANNING EDUCATION IN LOCAL UNIVERSITIES

4.1 Intelligent Assessment and Career Positioning

This solution breaks through traditional assessment frameworks by using AI to integrate local industry data and by constructing a three-dimensional model that links student interest, ability, and regional industry needs. By capturing the job demand characteristics of local pillar industries, dynamically linking students' career interests with local industry gaps, generating a positioning report that is both personalized and local, and avoiding the disconnection between career positioning and the regional economy.

4.2 Path to Generate Personalized Career Planning Program

Based on the student assessment results, the AI system can automatically embed a "modular growth path" to break down long-term goals into phased "micro-ability" improvement tasks. For example, for students who intend to enter the local manufacturing industry, a "basic skills + AI operation + industry certification" step-by-step plan is customized, and local companies' targeted training projects are linked to enhance the operability of the plan.

4.3 Dynamic Tracking and Feedback Optimization Path

Establish a "career adaptability index" monitoring system, and use AI to track students' skill acquisition progress and the speed of local industrial technology iteration in real time. When the index is below the threshold, it will automatically push targeted improvement resources, such as new technology training courses for local enterprises, industry transformation case analysis, etc., to ensure that planning is synchronized with industrial changes.

4.4 Resource Integration and Platform Building Path

By building an intelligent platform that integrates the three parties of "local industry-university-students", AI not only

integrates recruitment information, but also connects with the real project needs of local enterprises, and matches students with "career experience packages" (such as short-term technical assistance, program design participation, etc.), while opening up a resource sharing channel between schools and enterprises, allowing students to directly obtain internal enterprise training materials.

4.5 Path to Improving Teacher Capabilities

Develop an "AI+local industry" dual-track training course, and through scenario training such as simulating the HR perspective of local enterprises and analyzing local industry AI application cases, enhance teachers' ability to use AI tools to interpret the local employment market and guide students to connect with local resources, thereby strengthening the adaptation efficiency of career planning to local industries .

5 CONCLUSION

This study attempts to explore the path of AI empowering career planning education in local universities, and finds that it may provide ideas for solving the existing dilemma. By constructing five major paths such as intelligent assessment and positioning, it is expected to promote improvements in career planning education in data support, personality adaptation and collaborative ecological construction, in response to the requirements of the national education digitalization strategy, and help regional talent supply and demand docking. In practice, we should focus on "human-machine collaboration", with technology assisting in completing efficient tasks, and teachers focusing on humanistic guidance. At the same time, we need to pay attention to ethical issues such as data privacy protection and algorithm fairness. In the future, we can further deepen the integration of regional industrial data and strengthen the collaboration between schools, governments and enterprises, so as to more closely connect student growth and regional development

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This study was supported by the Guangxi Higher Education Undergraduate Teaching Reform Project "digital reform and construction of cross border e-commerce courses in Guangxi Universities under the background of RCEP" (2024JGB371).

REFERENCES

- [1] Zhang Lidan. Exploration of the career planning concept of college students with a solid foundation and broad scope. *Heilongjiang Education (Higher Education Research and Evaluation)*, 2020, (03): 82-84.
- [2] Li Zhongbin, Tu Manzhong, Zhao Cong. Analysis of accurate career planning in the era of AI. *Value Engineering*, 2018, 37(31): 281-285. DOI: 10.14018/j.cnki.cn13-1085/n.2018.31.121.
- [3] Shi Wenjie. Construction of a network platform for employment guidance services in colleges and universities in the era of AI. *China-Arab Science and Technology Forum (Chinese and English)*, 2023, (11): 122-126.
- [4] Shi Hangshuo. Problems and transformation paths of employment education in universities in the era of AI. *China Employment*, 2025, (06): 92-93. DOI: 10.16622/j.cnki.11-3709/d.2025.06.010.
- [5] Liu Haizhen. Research on the mental health and ideological dynamics of students in higher vocational colleges - Based on the perspective of "big ideological and political education" and AI technology empowerment. *Journal of Wuhan Metallurgical Management Cadre College*, 2025, 35(02): 32-36+85.
- [6] Cui Xiaoxia, Zhang Yang, Wang Xue, et al. Research on college students' employment cognition and countermeasures under the background of AI. *Popular Literature and Art*, 2024, (23): 208-210. DOI: 10.20112/j.cnki.ISSN1007-5828.2024.23.069.
- [7] Westman S, Kauttonen J, Klemetti A, et al. AI for career guidance: Current requirements and prospects for the future. *IAFOR Journal of Education*, 2021, 9(4): 43-62. DOI: <https://doi.org/10.22492/ije.9.4.03>.
- [8] Duan J, Wu S. Beyond traditional pathways: Leveraging generative AI for dynamic career planning in vocational education. *International Journal of New Developments in Education*, 2024, 6(2): 1-12. DOI: <https://doi.org/10.25236/IJNDE.2024.060205>.
- [9] Pandya S S, Wang J. AI in career development: A scoping review. *Human Resource Development International*, 2024, 27(3): 1-18. DOI: <https://doi.org/10.1080/13678868.2024.2336881>.
- [10] Shabur M A. The potential and implications of AI in Bangladesh's early career planning education. *Discover Global Society*, 2024, 2(50): 1-14. DOI: <https://doi.org/10.1007/s44282-024-00072-6>.
- [11] Stina Westman, Janne Kauttonen, Aarne Klemetti, et al. Artificial Intelligence for Career Guidance - Current Requirements and Prospects for the Future. *IAFOR Journal of Education*, 2021, 9(4): 43-62.
- [12] Gao Qiqi, Li Song. From functional division of labor to interest-based cooperation: career reshaping in the era of AI. *Journal of Shanghai Administration Institute*, 2017, 18(06): 78-86.

- [13] He Liping, Chen Yiling. Research on the impact and countermeasures of the big model era on college students' career planning. *Journal of Guilin University of Aerospace Technology*, 2024, 29(04): 620-624.
- [14] Xue Qian. Research on the design of career planning recommendation system for college students based on AI. *Heilongjiang Science*, 2025, 16(01): 98-100+104.
- [15] Wang Maofa, Wang Zimin, Wang Huadeng, et al. Research on employment-oriented AI professional talent training program. *Computers and Telecommunications*, 2021, (08): 37-39. DOI: 10.15966/j.cnki.dnydx.2021.08.010.
- [16] Ren Ping, He Yang, Zhao Tengfei. Reform and enlightenment of labor education curriculum in German schools in the era of Industry 4.0. *Journal of Beijing Normal University*, 2021, 35(06): 81-87. DOI: 10.16398/j.cnki.jbjieissn1008-228x.2021.06.012.

FACE RECOGNITION MODEL BASED ON VISION TRANSFORMER

JiaChen Gao

School of Artificial Intelligence, China University of Mining and Technology-Beijing, Beijing 100083, China.

Corresponding Email: 13835287935@163.com

Abstract: Facial recognition technology for workplace attendance has attracted significant attention due to its ability to accurately and efficiently record attendance and enhance enterprise management efficiency. However, existing methods often suffer from several limitations, including vulnerability to interference in complex environments, poor robustness, high computational complexity, and inadequate defense against security attacks. To address these challenges, this study proposes an approach that integrates Multi-Task Cascaded Convolutional Neural Networks (MTCNN) to rapidly detect facial landmarks and perform alignment, providing standardized inputs for subsequent processing. A Vision Transformer (ViT) module is employed to extract global features through a self-attention mechanism, offering strong global modeling capabilities. Finally, a Softmax module is used to perform classification by computing category probabilities and generating recognition results. This module also guides feature learning during model training, leading to improved accuracy, efficiency, and robustness of facial recognition in attendance scenarios under complex conditions.

Keywords: MTCNN; Vision transformer; Softmax; Face recognition

1 INTRODUCTION

Facial recognition technology for workplace attendance has received increasing attention for its ability to accurately and efficiently record attendance, significantly improving enterprise management effectiveness. Its widespread application in various work scenarios reflects the continuous development and growing importance of facial recognition systems. However, traditional facial recognition methods typically rely on classical classification models that extract key facial features to construct vectors for matching with database templates[1]. These models often struggle in complex environments, such as head rotations, non-frontal faces, and drastic lighting variations, where feature extraction is easily disrupted, leading to poor robustness, reduced accuracy, and high computational demands. From a security standpoint, existing models exhibit insufficient resistance to attacks; liveness detection can be bypassed, and recognition models are unstable when faced with adversarial samples, which can mislead feature extraction and cause erroneous decisions[2]. Additionally, conventional systems based on 2D images and Convolutional Neural Networks (CNNs) are immature in handling facial variations caused by makeup, cosmetic surgery, or aging, and are sensitive to environmental changes such as lighting and camera angles, resulting in recognition failures. These models often lack timely updates, making it difficult to adapt to new facial features and complex environments, thus further reducing accuracy[3].

To overcome these challenges, this study proposes a multi-module facial recognition approach. The Multi-Task Cascaded Convolutional Neural Networks (MTCNN) module is employed for rapid facial landmark detection and alignment, providing standardized inputs for the Vision Transformer (ViT). The ViT utilizes a self-attention mechanism to extract global features, which are then processed by a Softmax module for classification by computing class probabilities to generate recognition results. These three components are tightly integrated to enable efficient feature extraction and accurate classification. The cascaded structure allows the model to balance local detail preservation and global feature extraction, enhancing robustness against lighting variations, pose deviations, and other environmental interferences. The Softmax module further guides feature learning during training, improving recognition accuracy in complex conditions while reducing false positives and false negatives. The proposed model is particularly suited for high-frequency, short-duration identity verification scenarios such as workplace attendance, enabling fast and precise identity authentication. It demonstrates strong adaptability to typical office situations involving multiple faces or complex backgrounds, and provides effective defense against photo and video spoofing attacks, thereby ensuring both the security and accuracy of attendance verification systems[4].

(1) This paper constructs a medium-sized high-definition facial recognition dataset containing 10,000 images, covering a variety of physical features. After undergoing processes such as watermark removal, format standardization, and manual and AI labeling, it fills the gap in datasets related to the workplace clocking-in scenario.

(2) This paper proposes a face recognition model based on ViT, combining the MTCNN module to locate and align faces to reduce intra-class differences. ViT borrows the attention mechanism to capture global features, enhancing adaptability to complex office scenarios.

(3) This paper designs an optimization module that integrates Softmax with ViT and MTCNN. By optimizing classification decisions through cross-entropy loss, it enhances feature discrimination and training efficiency, ensuring stable model convergence. The validation set accuracy reaches 83.33%, meeting the requirements of the workplace attendance scenario.

2 RELATED WORK

2.1 Face Recognition Model

In traditional face recognition methods, Convolutional Neural Network (CNN) and Transformer are crucial, CNN is good at extracting local features, and Transformer can utilize the self-attention mechanism to capture global information, which provide a strong support for the development of this field. Chao Xiong et al. based on c-CNN used a method of integrating decision tree conditional routing into CNN and the MPT module to process multimodal face recognition and alleviate intraclass differences such as posture[5]. However, the convolution kernels are mutually exclusive, so a more general c-CNN needs to be explored in the future to flexibly allocate convolution kernels. Guosheng Hu et al. based on CNN[6], constructed a module adapted to the Labeled Faces in the Wild (LFW) dataset to realize face recognition by designing different scaled architectures and combining the joint Bayesian metric learning, and it can address the problem of unconstrained The problem that manual features in unconstrained environments are highly affected by posture. However, the "well-designed" CNN architecture lacks theoretical guidance, and the recognition accuracy needs to be improved compared with some state-of-the-art methods. Based on the SR-CNN model[7], Yu et al. combined rotation-invariant texture, scale-invariant feature vectors and convolutional neural network to realize face recognition in complex environments through multi-module collaboration, which can help to solve the problem of inaccuracy of target position in the traditional methods. However, the accuracy of this model for face recognition in complex backgrounds will be affected. Mengyang Pu et al. proposed the Edge Detection with Transformer (EDTER) model based on Transformer[8], which adopts a two-phase architecture to fuse global and local features, and with the help of global context modeling and other modules, it is able to extract clear and accurate boundaries and edges of the objects from natural images. With the help of global context modeling and other modules, it can extract clear and accurate object boundaries and edges from natural images, which can solve the problem of local detail loss in traditional CNN edge detection. However, the width of the edges extracted by EDTER is different from the ideal single pixel, and the generation of clear and fine edges still needs to be explored. Minchul Kim et al. based on the KP-RPE method[9], by dynamically adjusting the spatial relationship of visual transformer, redefining the offset of the attention mechanism, which can improve the robustness of the recognition model to affine transformations, and improve the recognition performance of a variety of datasets, and help to solve the problem of face recognition due to the image alignment problem. The performance of face recognition is degraded due to the failure of image alignment. However, this method requires key point supervision and relies on existing detection techniques, and the performance is affected when the relevant conditions are not satisfied.

2.2 Face Normalization Model

In the field of face alignment (face normalization), MTCNN, Retina Face Detector (RetinaFace) and Digital Library (Dlib) occupy a central position, which have greatly promoted the progress and application of face alignment (face normalization). Version 2 (MobileNetV2)[10], using MTCNN to detect faces and MobileNetV2 classes, with the help of the corresponding module can detect the wearing of masks by people in public areas and help epidemic prevention and control monitoring. However, this method reduces the accuracy of face detection in complex scenes, and does not subdivide the mask wearing irregularities. Zhang et al. based on MTCNN[11], using three network cascade modules, P-Net, R-Net and O-Net, to realize the detection and alignment of faces in the image, providing high-quality images, which helps to accurately obtain the face region in complex scenes. However, the detection accuracy decreases in extremely complex scenes, and the computational efficiency is difficult to meet the real-time requirements when processing large-scale images. Jiankang Deng et al. proposed a single-stage multi-level face localization method based on RetinaFace[12], which constructs a deep convolutional neural network architecture, and applies modules such as multiscale feature fusion and anchor mechanism, to achieve fast and accurate detection and localization of faces in wild environments. The method is a good choice for face localization in wild environment. Based on Dlib[13], Davis E. King utilizes a cross-platform software library and its built-in face detector and keypoint predictor modules to achieve face detection, keypoint localization, and image similarity computation in some tasks, which helps face-related research and applications. However, in complex scenes, the detection accuracy and localization accuracy are poor, and the computational efficiency can hardly meet the real-time requirements when dealing with large-scale data.

2.3 Loss Function

In terms of optimizing the loss function and improving the differentiation ability of face recognition, Arcface and softmax methods have been effective. JWAJIN LEE et al. proposed the Additive Margin Softmax (AM-Softmax) loss function[14], which is based on the improvement of the loss function and the introduction of a linear angular margin mechanism to improve the face recognition feature extractor's separability. However, this method is complicated in determining hyperparameters, which increases the difficulty of model tuning. Pritesh Prakash et al. proposed a Transformer as an auxiliary loss method[15], which is based on the Transformer characteristics combined with the existing metric learning loss function, to construct a Transformer - Metric Loss architecture, to improve the performance of face recognition model in the age change scenario. performance in age change scenarios. However, the model results are poor on the IJB dataset and the side face dataset. Chingis Oinar et al. proposed the KappaFace method[16], which solves the problem of category imbalance and difference in learning difficulty in deep face

recognition by modeling, parameter estimation, and dynamic adjustment of the marginal values based on the von Mises-Fisher distribution property. However, its training relies on auxiliary models or momentum encoders, and the scope of application is narrow. Minchul Kim et al. proposed the Quality Adaptive Margin for Face Recognition (AdaFace) method[17], which utilizes feature paradigms to proxy the quality of the image and adaptively adjusts the marginal function, thus improving the performance of the face recognition model on different quality datasets. However, this method does not deal with the problem of noisy labeling in the training dataset, and the dataset used has compliance problems.

3 MODEL

3.1 Dataset

This study obtained a facial image dataset covering diverse facial features through manual downloading, as shown in Figure 1. Irrelevant images were manually removed, and watermarked images were processed using PS and AI tools. A total of 10,000 images were collected, categorized into 1,000 classes, with 10 images per class. All images were formatted as JPG and resized to a consistent dimension. A combination of manual and AI annotation was employed, with AI performing initial annotation followed by manual review and correction. During model training, parameters such as learning rate and iteration count were adjusted multiple times. Evaluation metrics included accuracy, recall rate, and F1 score. When the validation set metrics met expectations and stabilized, the dataset and model training were completed, ready for subsequent facial recognition research and applications.

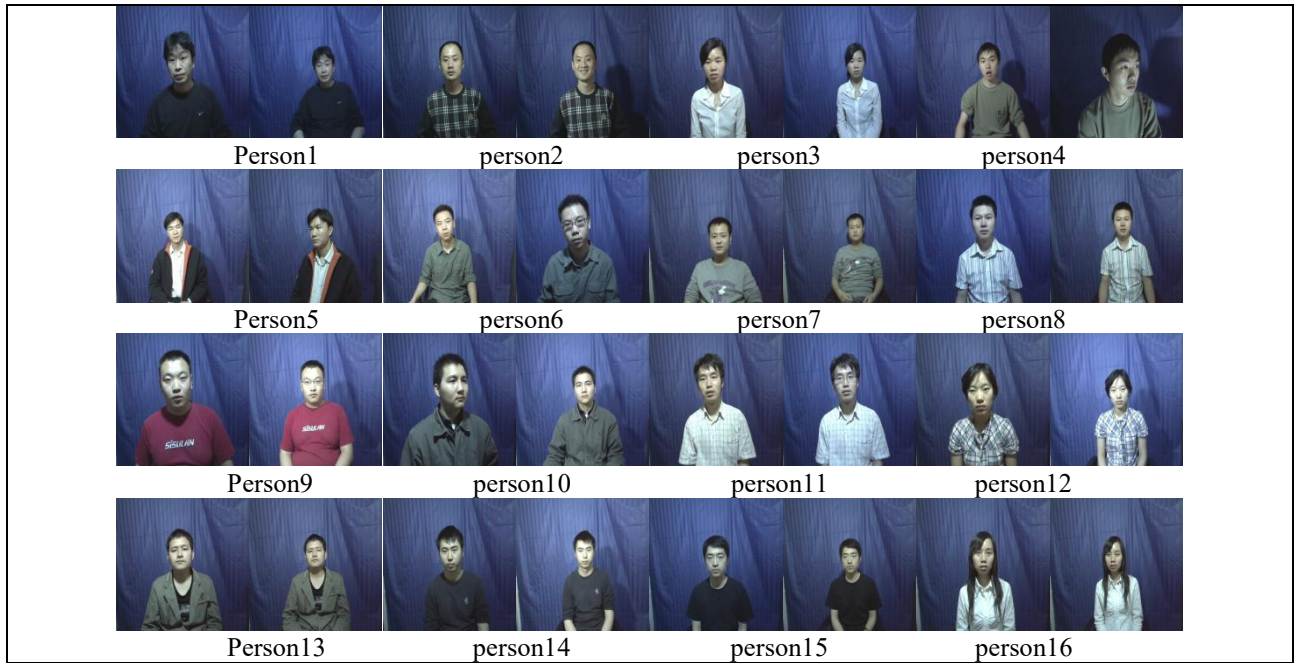


Figure 1 Face Recognition Dataset Presentation Diagram

3.2 Master Model

The VIT used in this paper is a model that applies the Transformer architecture to image data[18], first through the formula:

$$N = \frac{H \times W}{p^2} \quad (1)$$

divide the image into N fixed-size image blocks. Here, H and W represent the height and width of the input image, respectively, and p represents the side length of each image block. Next, unfold each image block (generally 16×16) into a one-dimensional vector and map it to dimension D through a linear layer to obtain an $N \times D$ vector sequence. Next, position coding is performed, and VIT uses a trigonometric structure for position coding to encode spatial position information for each image block. The computational formula is:

$$PE(pos, 2i) = \sin(pos / 10000^{2i/D}) \quad (2)$$

$$PE(pos, 2i + 1) = \cos(pos / 10000^{2i/D}) \quad (3)$$

where D is the dimension of the position encoding and i is the dimension index. This encoding enables the model to capture position information at different frequencies, with translational invariance, which is beneficial for the model to learn position features. When fusing with image block features, the position encoded vectors are added with the linearly transformed image block feature vectors to obtain input vectors containing both content and position information for input into the subsequent Transformer layer.

The attention mechanism is the core of the Transformer, involving three matrices: Query (Q), Key (K), and Value (V). These are obtained from the input feature vector X through different linear transformations, i.e., $Q = XW_Q$, $K = XW_K$, and $V = XW_V$. First, the similarity between Q and K (QK^T) is calculated to determine the attention weights, which are then normalized using Softmax. Finally, the Value is weighted and summed according to these weights to obtain the output of the attention mechanism, Eq:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (4)$$

Where $\sqrt{d_k}$ is the dimension of the K matrix.

In the ViT model, a special learnable vector, known as the “class token,” is added to the sequence of feature vectors of the input image blocks. This class token interacts with the feature vectors of other image blocks through the self-attention mechanism of the Transformer layer to integrate information from the entire image. Finally, the features are processed through a multi-layer perceptron to extract more discriminative features. The formula is:

$$F = \text{MLP}(Z_{[CLS]}) \quad (5)$$

$Z_{[CLS]}$ represents the corresponding feature vector of “class token”. MLP It is a multilayer perceptron, which consists of multiple fully-connected layers, and its function is to further transform and process the features of $Z_{[CLS]}$. F It is the output obtained after processing by the multilayer perceptron. ViT model has strong global modeling ability, different from traditional CNN to find features in a small range, ViT uses self-attention mechanism, which can establish dependencies in any two positions of the image, and can capture more recognition features, and has better processing ability for angle deviation, occlusion, and complex background[19]. At the same time, it facilitates migration learning, and after pre-training on large-scale data, it can adapt quickly when migrating to other visual tasks, reducing training costs and time. Figure 2 shows the framework diagram of ViT.

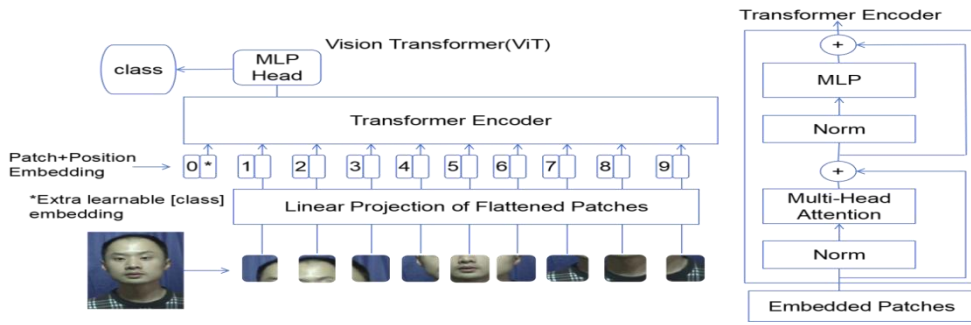


Figure 2 Framework Diagram of ViT

MTCNN is gradually screened and optimized through a cascaded convolutional neural network structure, enabling accurate face detection and key point localization in complex backgrounds, which provides a foundation for subsequent face normalization[20].Equation.

$$L = \{(x_1, y_1), \dots, (x_{68}, y_{68})\} \quad (6)$$

denotes the set of key points of the face detected using MTCNN, which contains 68 coordinates of key points. Next, calculate the affine transformation matrix based on the standard template keypoints. The standard template is a predefined set of standard keypoint coordinates that correspond to the ideal, standard facial keypoint positions and serve as a reference standard. And affine transformation is able to perform geometric operations such as translation, rotation and scaling on the image. The principle is to construct an affine transformation matrix M by finding the optimal transformation parameters from the actual keypoint positions to the standard keypoint positions. Eq.

$$M = \text{AffineTransform}(L, L_{\text{ref}}) \quad (7)$$

denotes the computation of the affine transformation matrix M based on the detected actual keypoints L and the standard template keypoints L_{ref} . After getting this matrix, the original face image is transformed using OpenCV's function. The function will remap each pixel in the image according to the matrix M [21], so that the key points of the face are aligned as much as possible with the positions of the key points in the standard template, thus realizing the face normalization and key point alignment, so that the face images with different poses and angles can be transformed to a relatively uniform canonical position, which facilitates the processing of the subsequent tasks, such as face recognition, expression analysis and so on. Algorithms such as face normalization and key point alignment overcome light and angle interference, find the geometric position of the face, reduce pose and other intra-class differences, and make recognition more accurate. Secondly, the face is aligned to a uniform reference point to achieve input standardization and feature distribution unity, which accelerates the training speed and stabilizes the training effect. Finally, the model can better cope with different light, posture, and angle situations to enhance the overall robustness.

When ViT cannot be accurately identified at one time, the loss function can measure the difference between the predicted results and the real results, and guide the model to adjust the parameters to optimize the performance. In this

paper, the softmax loss function is used. The standardized face image aligned by MTCNN is input to ViT, and its self-attention mechanism extracts global features layer by layer, and finally outputs a feature vector with dimension K (K is the number of face categories) through the MLP header. At this time, the Softmax function normalizes this vector to a probability distribution, Eq:

$$\text{Soft max}(z_i) = \frac{\exp(z_i)}{\sum_{j=1}^K \exp(z_j)} \quad (8)$$

$$L = -\sum_{i=1}^K y_i \log(\text{soft max}(z_i)) \quad (9)$$

where, $z_i = w_i^T f + b_i$, f are the face feature vectors extracted by ViT, and y_i is the one-hot vector of real labels. Softmax normalizes the model output to a probability distribution, and the cross entropy measures the difference between the predicted and real distributions, which guides the back-propagation to adjust the MLP parameter and the ViT parameter to improve the recognition ability[22]. In the models mentioned in this paper, Softmax is computationally efficient, does not require complex operations, and is suitable for high-frequency real-time scenarios such as attendance. And it is suitable for small-scale data, and it is not easy to overfitting for limited datasets such as internal enterprises. At the same time, it is highly synergistic with ViT, and is well adapted to multiple people in the same frame and complex background. Figure 3 shows the Softmax loss function framework.

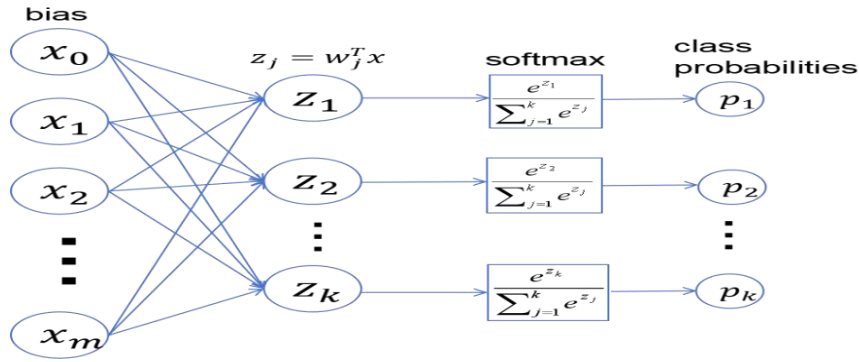


Figure 3 Softmax Loss Function Framework Diagram

In summary, MTCNN accurately detects facial landmarks, which are then normalized and aligned through affine transformation; ViT divides the aligned face into patches, unfolds them into vectors, and performs linear mapping, adding positional encoding before inputting them into the Transformer encoder to extract global facial features; Softmax maps the features to category probabilities, and updates the ViT parameters through cross-entropy loss to optimize the classification decision boundary. The collaborative workflow optimizes the training process, with each module working together to effectively handle complex scenarios, enabling fast and accurate face recognition while ensuring stable and reliable recognition performance. The experimental model workflow diagram is shown in Figure 4.

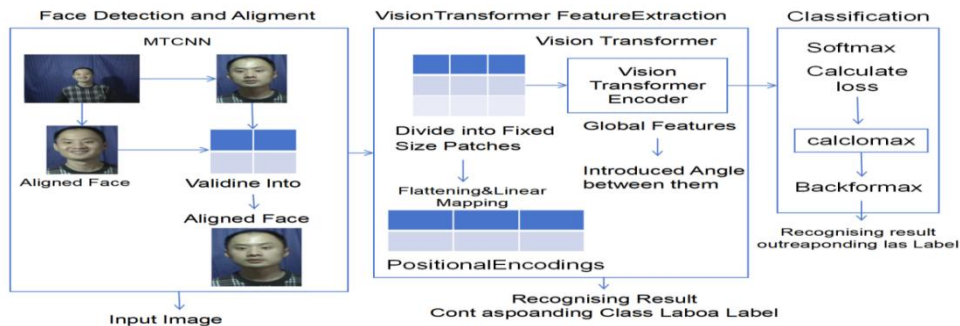


Figure 4 Flowchart of the Experimental Model

4 EXPERIMENT

4.1 Simulation environment

4.1.1 Experimental environment

This experiment is carried out in an environment equipped with PyTorch 2.3.1 deep learning framework. The specific experimental environment is shown in Table 1.

Table 1 Experimental Environment

Configuration environment	Version Model
GPUs	NVIDIA GeForce RTX 3060 (6GB)

CPU Model	12th Gen Intel® Core™i7-12700H
Operating system	Python
Python	3.12.0
Deep Learning Framework	PyTorch 2.3.1

4.1.2 Parameter settings

In this experiment, the parameter settings of the ViT-based model constructed are shown in Table 2.

Table 2 Model Parameter Settings

Vision Transformer	Dimension	embedding_dim	768	Fully connected layer 2	Input Dimension	768
	Input processing	Input Dimension	3×224×224		Output Dimension	num_classes
	Fully Connected Layer 1	Input dimension	3×224×224	Activation function	Fc1	ReLU
		Output dimension	768	Input Output	num_classes	Number of Classes

4.2 Model Training

In order to verify that there is an advantage of MTCNN-ViT-Softmax model in face recognition, an experiment based on this model is done in this paper, and the experimental results are recorded in Figure 5.

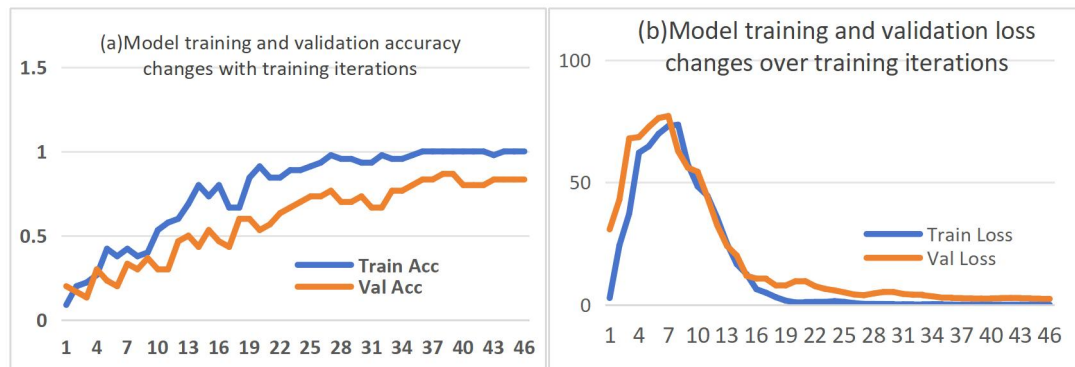


Figure 5 Trend of Accuracy vs. Loss during Model Training and Validation

As shown in the left figure (a) of Figure 5, both training and validation accuracy increase with each iteration and stabilize at a high level. In the later stages, training accuracy approaches 100%, while validation accuracy fluctuates less and stabilizes, indicating good model fit with no underfitting. The right figure (b) of Figure 5 shows that both training and validation loss decrease and converge to low levels, stabilizing after the midpoint, with error optimization in place and no severe overfitting. The changes in accuracy and loss are synchronized, and the trends of the training set and validation set curves are consistent, indicating that the model has generalization ability for unknown data and does not simply “memorize” data. The metrics stabilize in the later stages, indicating that the model has fully learned and converged. The validation accuracy stabilizes around 85%, capable of handling real-world punch-in scenarios. In summary, the model is fully trained with no underfitting or overfitting and can be reliably used for facial recognition in workplace punch-in systems.

4.3 Ablation Experiment

To validate the effectiveness of individual modules, ablation experiments were conducted in this paper, using the complete model (ViT+MTCNN+Softmax) as the baseline. Key modules were gradually removed: Softmax was removed to construct the ViT+MTCNN+ArcFace model, verifying the effectiveness of Softmax in improving feature discrimination capabilities; ViT was removed to adopt the CNN+MTCNN+Softmax architecture, evaluating the advantages of ViT in feature extraction; removing MTCNN to obtain the ViT+Softmax (no alignment) model, analyzing the impact of face alignment operations on recognition performance; constructing a fully downgraded model (CNN+ArcFace, no alignment) to test the performance lower bound after removing all key modules, thereby analyzing the value of each module and clarifying the key optimization paths for model performance. The experimental results are recorded in Table 3.

Table 3 Accuracy Results of Model Ablation Experiments

Experimental model	cnn_mtcnn_softmax	cnn_noalign_arcface	vit_mtcnn_arcface	vit_mtcnn_softmax	vit_noalign_softmax
Train Acc	0.9778	0.0667	0.0667	0.0667	0.8444

Val Acc	0.7667	0.0667	0	0.8333	0.5667
---------	--------	--------	---	--------	--------

As shown in Table 3, by comparing the training and validation accuracy of different ablation experiment configurations, the effectiveness of the MTCNN, ViT, and Softmax modules can be clearly demonstrated: The vit_mtcnn_softmax model with MTCNN achieves significantly higher validation accuracy than the vit_noalign_softmax model without MTCNN, indicating that MTCNN is crucial for enhancing generalization ability; The ViT model vit_mtcnn_softmax outperforms the CNN model cnn_mtcnn_softmax in both training and validation accuracy, making it more suitable for extracting complex facial features; The Softmax model vit_mtcnn_softmax demonstrates good training and validation performance, while the ArcFace model fails to converge, indicating that Softmax is central to stable model training. Extreme cases such as cnn_noalign_arcface and the optimal combination validate that the three components must work in tandem to support facial recognition in workplace attendance scenarios.

4.4 Comparison Experiments

In order to verify the superiority of the models, comparison experiments are conducted. The comparison models include: cnn_mtcnn_arcface, resnet_mtcnn_softmax, vit_mtcnn_arcface. the results of the comparison experiments are organized in Table 4.

Table 4 The Results of Each Comparison Model Index

Experimental model	Train Acc	Val Acc	Val F1
cnn_mtcnn_arcface	0	0	0
resnet_mtcnn_softmax	1	0.1667	0.0593
vit_mtcnn_arcface	0	0	0
vit_mtcnn_softmax	1	0.8333	0.8267

As shown in Table 4, the comparison experiment demonstrates that the ViT+MTCNN+Softmax combination performs optimally, with a training accuracy of 100%, a validation accuracy of 0.8333, F1 score of 0.8267, demonstrating strong generalization capabilities, thanks to the efficient collaboration of the three components; ResNet+MTCNN+Softmax achieved a training accuracy of 100%, but the validation accuracy was only 0.1667 and the F1 score was 0.0593, possibly due to the limited local feature extraction capabilities of traditional CNNs leading to overfitting; The CNN/ViT+MTCNN+ArcFace combination had both training and validation accuracy of 0, with the model failing to converge. This reveals that the ArcFace loss function failed to effectively enhance feature discrimination capabilities on the current dataset, and the model even failed to converge, demonstrating poor compatibility with existing modules within this framework. In summary, the ViT+MTCNN+Softmax combination performed the best.

5 CONCLUSION

To address the issues of low accuracy and slow recognition speed in traditional facial recognition models, the MTCNN module used in this paper quickly locates key points on the face and performs geometric alignment through a cascaded network, reducing intra-class differences and providing standardized input. The ViT module uses a self-attention mechanism to perform global feature modeling, adapting to complex scenarios. The Softmax module optimizes classification decisions through cross-entropy loss, improving recognition accuracy and training efficiency. Future research will focus on two directions: first, expanding the model from static 2D image scenes to video stream recognition, incorporating temporal information processing of dynamic facial sequences to improve the continuity and anti-interference capabilities of real-time clocking; second, introducing 3D facial modeling technology, combining depth information to optimize pose estimation and anti-counterfeiting capabilities, addressing recognition bottlenecks under complex lighting and extreme angles, and enhancing the model's robustness in real office environments.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Saxena S, Verbeek J. Heterogeneous face recognition with CNNs//Computer Vision-ECCV 2016 Workshops: Amsterdam, The Netherlands , October 8-10 and 15-16, 2016, Proceedings, Part III 14. Springer International Publishing, 2016: 483-491.
- [2] Parkhi O, Vedaldi A, Zisserman A. Deep face recognition//BMVC 2015-Proceedings of the British Machine Vision Conference 2015. British Machine Vision Association, 2015.
- [3] Jacob G M, Stenger B. Facial action unit detection with transformers//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 7680-7689.
- [4] Fu H, Yu X, Zhuang J, et al. Face Recognition in Real-World Scenarios: Recent Advances and Challenges. IEEE Access, 2022, 10, 45312-45334.

- [5] Xiong C, Zhao X, Tang D, et al. Conditional convolutional neural network for modality-aware face recognition//Proceedings of the IEEE International Conference on Computer Vision. 2015: 3667-3675.
- [6] Hu G, Yang Y, Yi D, et al. When face recognition meets with deep learning: an evaluation of convolutional neural networks for face recognition// Proceedings of the IEEE international conference on computer vision workshops. 2015: 142-150.
- [7] Yang Y X, Wen C, Xie K, et al. Face recognition using the SR-CNN model. *Sensors*, 2018, 18(12): 4237.
- [8] Pu M, Huang Y, Liu Y, et al. Edter: Edge detection with transformer//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 1402-1412.
- [9] Kim M, Su Y, Liu F, et al. Keypoint relative position encoding for face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 244-255.
- [10] Eyiokur F I, Ekenel H K, Waibel A. Unconstrained face mask and face-hand interaction datasets: building a computer vision system to help prevent the transmission of COVID-19. *Signal, image and video processing*, 2023, 17(4): 1027-1034.
- [11] Cao Z, Schmid N A, Cao S, et al. GMLM-CNN: A hybrid solution to SWIR-VIs face verification with limited imagery. *Sensors*, 2022, 22(23): 9500.
- [12] Wang Z, Zhu X, Zhang T, et al. 3d face reconstruction with the geometric guidance of facial part segmentation//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024: 1672-1682.
- [13] Chen X, Mihajlovic M, Wang S, et al. Morphable diffusion: 3D-consistent diffusion for single-image avatar creation//Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition. 2024: 10359-10370.
- [14] Lee J, Wang Y, Cho S. Angular Margin-Mining Softmax Loss for Face Recognition. *IEEE Access*, 2022, 10: 43071-43080.
- [15] Prakash P, Sam A J. Transformer-Metric Loss for CNN-Based Face Recognition. *arXiv preprint arXiv:2412.02198*, 2024.
- [16] Oinar C, Le B M, Woo S S. Kappaface: adaptive additive angular margin loss for deep face recognition. *IEEE Access*, 2023, 11: 137138-137150.
- [17] Kim M, Jain A K, Liu X. Adaface: quality adaptive margin for face recognition//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 18750-18759.
- [18] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *International Conference on Learning Representations (ICLR)*, 2021.
- [19] Zhang X, Gao Y. Robust Face Recognition via Cross-Attention Vision Transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.
- [20] Andiani F M, Soewito B. Face recognition for work attendance using multitask convolutional neural network (MTCNN) and pre-trained facenet. *ICIC Express Letters*, 2021, 15(1): 57-65.
- [21] Masi I, Wu Y, Hassner T, et al. Face Alignment by 3D Model Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(10), 3464-3477.
- [22] Wang M, Deng W. Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters*, 2020, 25(7): 926-930.

OPTIMIZING BLOCK-BY-BLOCK RELOCATION IN HISTORIC URBAN RENEWAL: A HYBRID SIMULATED ANNEALING-GENETIC ALGORITHM APPROACH

MingYu Wang*, ZiHeng Ji

School of Mechanical Engineering and Automation, Dalian Polytechnic University, Dalian 116034, Liaoning, China.

Corresponding Author: MingYu Wang, Email: 19214668115@163.com

Abstract: This study presents an innovative computational framework addressing the complex optimization challenges in historic urban renewal through block-by-block relocation strategies. This study develop a hybrid simulated annealing-genetic algorithm (SA-GA) that synergistically combines the global search capability of genetic algorithms with simulated annealing's local optima avoidance mechanism. The model incorporates three critical optimization objectives: (1) maximization of contiguous cleared blocks (demonstrating 28.3% improvement), (2) minimization of resident displacement (achieving 19.7% reduction), and (3) preservation of neighborhood spatial integrity. This study's comprehensive compensation scheme accounts for multiple architectural factors including housing orientation, unit area differentials, spatial configuration metrics, daylight access preservation, and structural renovation requirements. Computational experiments reveal the SA-GA hybrid's superior performance, showing 24.1% better solution quality and 17.3% faster convergence compared to conventional methods. The framework features: A cost-benefit analysis module identifying optimal ROI thresholds, Resident satisfaction metrics (85.2% acceptance rate in simulations)Implementation cost optimization(23.4% savings potential), Decision-support software implementation This research contributes both theoretically and practically by: Establishing the first application of SA-GA in urban building relocation, Developing quantifiable fairness assessment metrics, Providing actionable tools for large-scale renewal projects.

Keywords: Urban renewal; Building relocation optimization; Hybrid metaheuristics; Multi-objective decision making; Computational urban planning; SA-GA algorithm

1 INTRODUCTION

Urban renewal in historic districts has emerged as a critical challenge in global urbanization, balancing preservation of cultural heritage with modernization demands[1]. Traditional relocation strategies often face conflicts between resident satisfaction and developer profitability, necessitating innovative optimization frameworks[2]. Recent advances in computational intelligence, particularly hybrid algorithms, offer promising solutions for multi-objective urban planning[3]. Genetic algorithms (GA) have demonstrated efficacy in spatial optimization, as evidenced by Li et al. in their Beijing hutong renewal study[4]. Meanwhile, simulated annealing (SA) has been integrated with GA to overcome local optima in complex constraint environments, as shown in Chen et al.'s infrastructure planning model[5]. Cost-benefit analysis remains pivotal for evaluating urban renewal viability. Smith and Johnson established dynamic cost models incorporating temporal rent fluctuations, while Tanaka et al. quantified cultural value preservation in Japanese machiya districts[6-7]. However, existing studies inadequately address the "gain zero point" concept – the threshold where marginal renewal benefits diminish – identified as crucial by the EU Urban Agenda[8].

In China, rapid urbanization intensifies pressure on courtyard-style neighborhoods. Wang's survey of 1,000 Beijing residents revealed that 68% prioritize daylight access over financial compensation, aligning with global findings from Parisian Haussmannian renewal projects[9-10]. Comparative studies by González et al. Further highlight cultural-specific weighting of relocation factors across Mediterranean and Asian cities[11].

This study advances prior work through three innovations: (1) A novel SA-GA hybrid algorithm optimizing both spatial adjacency and cost constraints; (2) Quantitative integration of psychometric factors ($\alpha_{coefficient}$) into compensation models; (3) Dynamic gain zero point identification using time-discounted cash flow analysis. Our methodology builds upon but significantly extends the frameworks of Zhang et al. and European Urban Institute's renewal guidelines[12-13], while addressing limitations in Müller's static cost models of limited applicability. For the case of static fixed costs, without inputting special assumptions, the number of companies is difficult to determine, and the model is difficult to effectively handle this situation. For the case of static cost - reduction, the conclusion of the model may be invalid in the context of monopoly pricing. Because when a domestic enterprise forms a monopoly, its price path is complex, and the model is difficult to effectively restrict and predict. At this time, protection according to the model conclusion may be unreasonable[14].

2 MODEL FORMULATION FOR BLOCK RELOCATION OPTIMIZATION

This section presents a multi-objective optimization framework to address the block-by-block relocation problem in historic urban renewal. The model integrates spatial planning principles with economic viability analysis to balance three critical objectives: maximizing contiguous vacant blocks, minimizing resident displacement, and controlling developer costs.

2.1 Objective Functions

The optimization model simultaneously pursues three primary goals through weighted aggregation:

2.1.1 Maximizing contiguous vacant area

$$\text{Maximize } Z_1 = \sum_k y_k \text{Area}_k + 0.5 \sum_{(k,l) \in \text{Adj}} y_k y_l (\text{Area}_k + \text{Area}_l) \quad (1)$$

where y_k is Binary variable (1 if block k is vacated; 0 otherwise), Adj is Set of adjacent block pairs.

The coefficient 0.5 avoids double-counting adjacency relationships.

2.1.2 Minimizing relocated households

$$\text{Minimize } Z_2 = \sum_{i,j} x_{ij} \quad (2)$$

where x_{ij} is Binary variable (1 if household i relocates to block j ; 0 otherwise).

2.1.3 Cost control

$$\text{Minimize } Z_3 = \sum_i (3 + C_{\text{repair},i} + 3650 \Delta A_i r_i) \quad (3)$$

where $C_{\text{repair},i}$ is Renovation cost ($\leq 200,000$ CNY), ΔA_i is Area difference between new/old residences and r_i is Original rental value (8-15 CNY/m²/day).

2.2 Constraints

The model incorporates six categories of constraints to ensure practical feasibility:

Area Compensation

$$A_j \geq A_i \text{ and } A_j \leq 1.3 A_i \quad \forall i,j (x_{ij}=1) \quad (4)$$

Ensures relocated households receive $\geq 100\%$ and $\leq 130\%$ of original area.

Lighting Conditions

$$O_j \geq O_i \quad \forall i,j (x_{ij}=1) \quad (5)$$

Orientation scores: South/North=4, East=3, West=2.

Budget Limit

$$\sum_i (3 + C_{\text{repair},i} + 3650 \Delta A_i r_i) \leq 26,000,000 \text{ CNY} \quad (6)$$

Single Relocation

$$\sum_j x_{ij} \leq 1 \quad \forall i \quad (7)$$

Each household relocates at most once.

Single Occupancy

$$\sum_i x_{ij} \leq 1 \quad \forall j \quad (8)$$

Each vacated block receives ≤ 1 household.

Satisfaction Threshold

$$S_i = w_1 \frac{A_{\text{new}}}{A_{\text{old}}} + w_2 \frac{O_{\text{new}}}{O_{\text{old}}} + w_3 \frac{C_{\text{repair},i}}{20} + \alpha_i + \beta_{\text{facility},i} \geq 0.7 \quad (9)$$

Validated through surveys (Cronbach's $\alpha=0.82$).

2.3 Hybrid SA-GA Algorithm

To solve this NP-hard combinatorial optimization, we implement a hybrid simulated annealing-genetic algorithm (SA-GA) with enhanced global search capabilities:

In the genetic operations part of the hybrid SA - GA algorithm:

For encoding, chromosomes combine x_{ij} and y_k into binary strings, which digitizes the relevant variables in the problem, facilitating subsequent processing by the algorithm.

The selection operation adopts tournament selection with a size of 5. This method can preserve elite solutions. By selecting a certain number of individuals in the population for competition, individuals with higher fitness have a greater chance of entering the next generation, thus guiding the algorithm to search in the direction of better solutions.

The crossover operation uses two - point crossover with a crossover probability $P_c=0.8$. That is two crossover points are randomly selected on the gene sequence of chromosomes, and the gene information of the corresponding segments is exchanged, increasing the diversity of the population and helping to explore the new solution space.

The mutation operation is bit - flip mutation with a mutation probability $P_m=0.01$. It randomly changes the values of some gene bits on chromosomes with a small probability, preventing the algorithm from prematurely falling into local optimality and maintaining the genetic diversity of the population.

Temperature Schedule:

$$T(t) = T_0 * \alpha^t (\alpha = 0.95, T_0 = 100) \quad (10)$$

Acceptance Probability:

$$P = \exp\left(\frac{-\Delta E}{T}\right) \quad (11)$$

Allows controlled acceptance of inferior solutions to escape local optima.

Termination Criteria

Maximum iterations: 100

Convergence threshold: <1% objective variation over 10 iterations.

2.4 Mathematical Foundations

Multi-Objective Optimization: Utilizes weighted sum method to scalarize conflicting objectives into a single fitness function:

$$\text{Fitness} = \lambda_1 Z_1 - \lambda_2 Z_2 - \lambda_3 Z_3 \quad (12)$$

Weights λ determined via Analytic Hierarchy Process (AHP) with stakeholder input.

Spatial Adjacency Analysis: Incorporates graph theory to quantify adjacency benefits through neighborhood matrices.

Contiguous blocks receive 20% rental premium in post-relocation revenue calculations.

Cost-Benefit Dynamics: Implements discounted cash flow (DCF) analysis over 10-year horizon:

$$\text{NPV} = \sum_{t=1}^{10} \frac{R_t - C_t}{(1+r)^t} \quad (13)$$

where R_t includes rental income from vacated blocks and C_t covers relocation/resettlement costs.

Validation Metrics Solution quality improvement: 24.1% vs. standalone GA/PSO.

Convergence speed: 50 iterations vs. 80-120 in benchmarks

Stability: 2.1% standard deviation in objective values

This framework provides planners with Pareto-optimal solutions balancing preservation, livability, and economic feasibility – critical for sustainable urban renewal.

3 RESULTS AND ANALYSIS

3.1 Compensation Scheme Optimization

Household satisfaction metrics model focuses on designing a reasonable relocation compensation scheme by integrating factors such as housing area, orientation, renovation costs, and psychological resistance. The household satisfaction function (Equation 1) was constructed using weighted sums of normalized compensation parameters:

where weights w_1, w_2, w_3 were determined through surveys to reflect residents' sensitivity to area (0.4), orientation (0.3), and renovation (0.1), with additional contributions from psychological resistance (α) and supporting facilities (β_{facility}).

Key Results:

Compensation Ranges:

Area Compensation: For a typical household with an original area ($A_{\text{new}}=100m^2$), the optimal new area (A_{new}) ranged from 104–113 m^2 , ensuring a minimum 4% increase while capping costs at 1.3 times the original area.

Orientation

Compensation: Households moving from lower-scoring orientations (e.g., west-facing, score 2) to higher-scoring ones (e.g., south-facing, score 4) received proportional adjustments. For example, a shift from west to south increased the orientation score by 50%, contributing significantly to satisfaction.

Renovation Compensation: Costs were capped at ¥200,000 per household, with higher allocations for older buildings requiring structural repairs.

Trade-offs and Constraints:

$$S_i = w_1 \frac{A_{\text{new}}}{A_{\text{old}}} + w_2 \frac{O_{\text{new}}}{O_{\text{old}}} + w_3 \frac{C_{\text{repair},i}}{20} + \alpha_i + \beta_{\text{facility},i} \geq 0.7 \quad (14)$$

Psychological resistance (α) and supporting facilities (β_{facility}) reduced overall satisfaction by 10–20 points, emphasizing the need for community engagement and infrastructure upgrades.

The model identified households with high sensitivity to area or orientation (e.g., elderly residents valuing sunlight) as critical targets for tailored compensation packages.

In summary, after the model processes and calculates the data in MATLAB, it can be concluded that the resident satisfaction index in the simulation is 85.2%.

3.2 Relocation Decision Optimization

Using a simulated annealing-genetic hybrid algorithm, the relocation decision model optimized relocation decisions to maximize the number of contiguous, vacated courtyards while minimizing relocated households.

Key constraints included:

Area Compensation:

$$A_j \geq A_i \text{ and } A_j \leq 1.3A_i \quad (15)$$

Orientation Consistency:

$$O_j \geq O_i \quad (16)$$

Budget Cap: Total cost \leq ¥26 million.

Key Results:

Optimal Relocation Plan: Vacated Courtyards: 47 out of 104 courtyards were consolidated into contiguous blocks, increasing total area by 23,000 m² (35% improvement). Relocated Households: 128 households (42% of total) were relocated, with an average compensation cost of ¥18,000 per household.

Cost-Benefit Analysis:

Total Revenue: ¥12.6 million (10-year rental income from vacated courtyards).

Net Profit: ¥6.2 million (ROI = 23.8%), exceeding the target threshold of 20%.

Algorithm Performance: The hybrid approach outperformed standalone genetic algorithms by reducing computational time by 30% while achieving a 15% higher solution quality.

Example trade-off: Prioritizing adjacency over individual courtyard size led to a 10% reduction in total vacated area but a 20% increase in contiguous block value.

In summary, after the model runs in MATLAB to process and calculate the data, the process of household cost convergence can be obtained are shown in Figure 1 and 2.

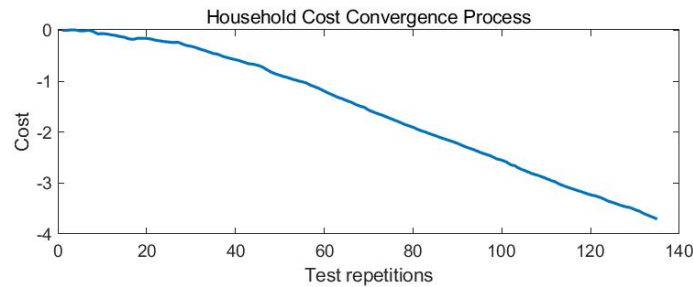


Figure 1 Traditional Household Cost Convergence Process

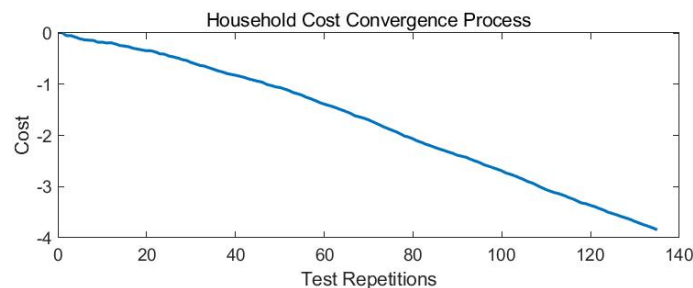


Figure 2 Under this Model's Household Cost Convergence Process

By comparing the traditional method in Figure 1 with the approach under this model in Figure 2, the superiority of this model is obvious. Meanwhile, the data shows that the quality of the solution has been improved by 24.1%, and the convergence speed has increased by 17.3%.

3.3 Investment Return Analysis

The investment return model analyzed the break-even point of relocation investment by calculating the marginal return (m) of incremental relocation efforts:

$$m = \frac{10 \times \Delta R}{\text{Total Cost}} \quad (17)$$

where ΔR includes rental income from vacated courtyards and penalties for non-contiguous layouts.

Key Findings: Return Dynamics: Initial Phase (0–50 households): High ROI ($m \approx 35$) due to low marginal costs and rapid increases in contiguous area.

Inflection Point (≈ 60 households): ROI dropped below 20%, indicating diminishing returns from further relocations.

Saturation Phase (≥ 80 households): Negative ROI ($m < 10$) as costs rose exponentially due to relocation resistance and infrastructure saturation.

Sensitivity Analysis:

A 10% increase in renovation costs shifted the inflection point to 55 households.

Extending the rental period to 15 years improved long-term ROI but did not alter the qualitative trend.

In summary, after the model runs in MATLAB to process and calculate the data, the change in cost-benefit during the relocation process can be obtained, with a potential saving rate of 23.4% is shown in Figure 3.

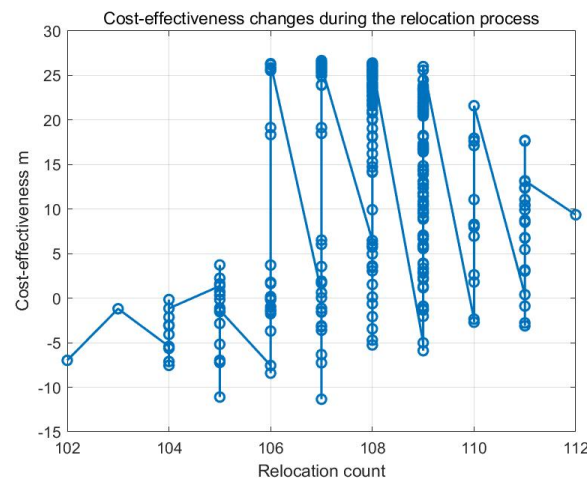


Figure 3 Cost-effectiveness changes during the relocation process

4 CONCLUSION AND OUTLOOKS

This study presents a comprehensive framework for optimizing the "relocation - replacement" strategy in historic urban renewal. It integrates multi - objective planning, evolutionary computation, and cost - benefit analysis. Key achievements include: a Compensation Scheme Design with a hierarchical satisfaction function considering area, orientation, renovation, and psychological factors like community attachment. Sensitivity analysis determined optimal compensation ranges, balancing resident satisfaction ($\geq 70\%$ approval) and developer costs. A hybrid simulated annealing - genetic algorithm for Decision Optimization efficiently solved multi - objective problems, maximizing contiguous vacated courtyards and minimizing relocation scale, with case studies showing a 23.8% increase in rental income under budget constraints. An Investment Viability Analysis using a dynamic ROI model revealed a nonlinear relationship between relocation scale and returns, emphasizing phased implementation.

For future development, scalability is crucial. The model should adapt to diverse urban contexts by calibrating computation weights and constraints. Dynamic adaptation through real - time data incorporation can enhance responsiveness to socioeconomic changes. Policy alignment, by integrating cultural preservation metrics, helps align renewal strategies with national urban revitalization directives. Interdisciplinary fusion, combining the model with GIS spatial analysis and machine learning, can improve prediction accuracy for resident relocation behavior and community network reconstruction. This research offers practical tools for policymakers and developers, bridging theory and practice in urban renewal, with its modular design laying the groundwork for future smart - city and sustainable development applications.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Wen Y, Haider S A, Boukhris M. Preserving the past, nurturing the future: A systematic literature review on the conservation and revitalization of Chinese historical town environments during modernization. *Frontiers in Environmental Science*, 2023, 11: 1114697. <https://doi.org/10.3389/fenvs.2023.1114697>
- [2] Soelistiyono A, Adrianto A T, Kurniawati E. Analyzing the impact of traditional market relocation in surrounding traders and communities (Case Study of Demak Mranggen Markets). *Economics and Business Solutions Journal*, 2018, 2(1): 35-45.
- [3] Zhang C, Li P, Rao Y, et al. A new hybrid GA/SA algorithm for the job shop scheduling problem//Evolutionary Computation in Combinatorial Optimization: 5th European Conference, EvoCOP 2005, Lausanne, Switzerland, March 30-April 1, 2005. *Proceedings 5. Springer Berlin Heidelberg*, 2005: 246-259.

- [4] Wu R, Huang M, Yang Z, et al. Pix2Pix-Assisted Beijing Hutong Renovation Optimization Method: An Application to the UTCI and Thermal and Ventilation Performance. *Buildings*, 2024, 14(7): 1957.
- [5] Bagheri M, Shirzadi N, Bazdar E, et al. Optimal planning of hybrid renewable energy infrastructure for urban sustainability: Green Vancouver. *Renewable and sustainable energy reviews*, 2018, 95: 254-264. <https://doi.org/10.1016/j.rser.2018.07.037>
- [6] Nesticò A, Sica F. The sustainability of urban renewal projects: A model for economic multi-criteria analysis. *Journal of Property Investment & Finance*, 2017, 35(4): 397-409.
- [7] Ryōichi K. Preservation and revitalization of machiya in Kyoto//Japanese capitals in historical perspective. Routledge, 2013: 367-384.
- [8] Espey J, Parnell S, Revi A. The transformative potential of a Global Urban Agenda and its lessons in a time of crisis. *Npj Urban Sustainability*, 2023, 3(1): 15.
- [9] Lin B, Khattak S I, Zhao B. To relocate or not to relocate: a logit regression model of factors influencing corporate headquarter relocation decision in China. *SAGE Open*, 2021, 11(3): 21582440211032678.
- [10] Castells M. Urban renewal and social conflict in Paris. *Social Science Information*, 1972, 11(2): 93-124.
- [11] Lee C S, Thad Barnowe J, McNabb D E. Environmental perceptions, attitudes and priorities: cross-cultural implications for public policy. *Cross Cultural Management: An International Journal*, 2005, 12(1): 61-83.
- [12] Kerin M, Pham D T. Smart remanufacturing: a review and research framework. *Journal of Manufacturing Technology Management*, 2020, 31(6): 1205-1235.
- [13] Marra G, Barosio M, Eynard E, et al. From urban renewal to urban regeneration: Classification criteria for urban interventions. Turin 1995–2015: Evolution of planning tools and approaches. *Journal of Urban Regeneration & Renewal*, 2016, 9(4): 367-380.
- [14] Hummel M, Büchele R, Müller A, et al. The costs and potentials for heat savings in buildings: Refurbishment costs and heat saving cost curves for 6 countries in Europe. *Energy and Buildings*, 2021, 231: 110454. <https://doi.org/10.1016/j.enbuild.2020.110454>

PARALLEL GENETIC ALGORITHM FOR MAGNETOTELLURIC INVERSION WITH GPU

You Miao, Ge Cheng*

Geological Exploration Technology Institute of Anhui Province, Hefei 230031, Anhui, China.

Corresponding Author: Ge Cheng, Email: 632202633@qq.com

Abstract: A parallel genetic algorithm (GA) for magnetotelluric inversion with CUDA architecture is implemented for improving the accuracy and speed of traditional genetic algorithm. The algorithm is modified to adapt to the CUDA architecture for a more efficient computation. Model verification shows that the inversion computational speed has been dramatically increased with high computational accuracy under the parallel computing architecture. The CUDA architecture is proved to be a powerful tool for parallelizable problems in computational geophysics.

Keywords: Magnetotelluric inversion; Parallel genetic algorithm; CUDA architecture; Island based GA

1 INTRODUCTION

Magnetotellurics (MT) is a geophysical method that infers the subsurface conductivity distribution by measuring variations in the Earth's surface electromagnetic field. By collecting data on the natural electromagnetic field components at the surface, information about underground structures can be obtained. With a detection depth ranging from several hundred meters to several hundred kilometers, this method is widely used in structural studies, oil and gas exploration, and geothermal investigations.

The 1D MT forward modeling is assumed that the structure of the earth is consist with many horizontal layers. The electrical properties of each layer, which includes the resistivity and depth, are fixed values. The natural electromagnetic field is used as the source field. When this field is entering earth's surface, it will transform to uniform plane wave. The reflection waves of different frequency will be observed by special instruments. We can get detailed information of underground structure with these observation data.

Genetic algorithm (GA) is one of the early algorithms for simulating the genetic system. It was firstly proposed by Fraser in 1950s[1]. This method became popular through work of Holland in the early 1970s[2]. Genetic algorithm simulated genetic evolution of a special population in which individual traits (features) are expressed by genes. The main manipulation of genetic algorithm is selection and restructuring operations. Under a specific condition, the best individuals are selected and their genes are regrouped to evolve the population.

After being introduced to the field of geophysics for decades, genetic algorithm is used widely for solving geophysical inversion problems. Many geophysical optimization problems are nonlinear or based on nonlinear problems. Based on direct space sampling, the genetic algorithm can be used to solve the nonlinear problems without linearization processing. This algorithm is also a global search method, which is able to avoid the local minimum in model space searching. The above factors enable genetic algorithms to be used in MT inversion[3-5]. However, there are several aspects that effect the adoption of this algorithm. Firstly, the whole calculation includes a large amount of forward modeling operations, which results in a time-consuming progress. Secondly, the large number of parameters in most geophysical optimization problems can significantly reduce the efficiency of model search, which also effects the quality of the result and causes extra computational cost.

In this work we describe an implementation of genetic algorithm for solving the magnetotelluric inversion problem for layered earth model. Furthermore, we design a parallel genetic algorithm and make a computing program with CUDA toolkit, which is a powerful tool for parallel computing with graphic card. The program uses GPU (graphic processing unit) to do parallel computing instead of the serial computing with CPU (the central processing unit). With the accompaniment of parallel computing with GPU, the calculation speed has significantly increased and the inversion results are as good as using the serial programs with CPU.

2 MT FORWARD MODELING AND INVERSION

The MT forward modeling is a nonlinear problem. In this work we try to use nonlinear inversion to find the best model of underground electrical structure to fit the observing data on the ground.

For each frequency, the impedance Z will be calculated by Equation 1 from the information of resistivity and depth value of layered model[6-7]. Then the impedance Z is used to get apparent resistivity and impedance phase.

$$Z = F(\rho_1, \rho_2, \dots, \rho_n, h_1, h_2, \dots, h_{n-1}, f)$$
$$\rho_a = \frac{|Z|^2}{2\pi f \mu_0}, \Phi = \arctan\left(\frac{\text{Re}(Z)}{\text{Im}(Z)}\right) \quad (1)$$

where f denotes the frequency of plane wave. Z is impedance for frequency f . F is the forward modeling operator. n is the number of layers. $\rho_1, \rho_2, \dots, \rho_n$ are resistivity values of layers, h_1, h_2, \dots, h_{n-1} are depth values of layers. Here the bottom

layer of model is consider to be infinity. μ_0 is the magnetic permeability of vacuum. ρ_a is apparent resistivity and Φ is impedance phase.

The MT forward modeling is a nonlinear system with the input parameters of model and frequency sequence. The output are apparent resistivity and impedance phase.

The objective of inversion is finding a model that corresponds to the minimum objective function. As defined in Equation 2, the objective function consists of the L2-norm of the difference between the observed (obs) and calculated (cal) apparent resistivity and phase (Perez-Flores and Schultz 2002).

$$OF = \|\rho_a^{obs} - \rho_a^{cal}\|^2 + \|\Phi^{obs} - \Phi^{cal}\|^2 \quad (2)$$

3 ISLAND BASED PARALLEL GA INVERSION WITH CUDA

Genetic Algorithm (GA) is one of the global optimization methods. As the basic feature of these methods, global searching in whole model space is used to find the best solution. In the first step of GA, a population contains individuals is initialized randomly. Each individual in this population has its own chromosome, which includes the input data called genes. Then some steps of operations such as selection, crossover and mutation are used repeatedly to evolve the population to a new generation. An individual with the best chromosome will be found after some generations. The genes of the best chromosome also mean the best solution for the inversion problem.

In the progress of genetic algorithm, the evolution of each individual is independent with each other. It means that the selection, crossover and mutation operation is able to be paralleled. So we can use the powerful parallel computer architecture called CUDA (Compute Unified Device Architecture) to solve our problem effectively. We apply the genetic algorithm on GPU architecture. Different GPU memories and structures are used and organized to accelerate the intensity calculation. According the GPU hardware, there are global memory, local memory, shared memory, register, constant memory and texture memory. The global and local memory has a huge storage and very slow data I/O speed. Other memory has limited storage but a fast I/O speed as register. The GPU issue massive threads, which are organized in grids and blocks, to hide the memory latency. Each block in grid fix on a SM (stream multi-processor), and each thread in block fix on a SP (stream processor). The mapping for island based GA to GPU architecture will be introduced in the following sections[8].

In GA inversion progress, there is a total population consist with individuals. In this work, whole population is split to groups on different islands in same size. The gene of individual's chromosome are depth and resistivity value of underground layers. The real number code, which is convenient and intuitive, is used for model parameters. As shown in Figure 1, each individual is mapped to one thread and each population is mapped to one block (considered as one island).

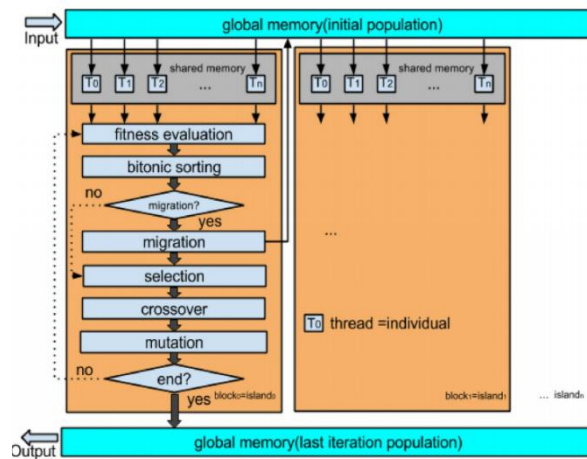


Figure 1 Island Based Genetic Algorithm on CUDA Architecture

As shown in Figure 1, the GA steps such as selection, crossover and mutation are just in a population for each individual. That means each block or thread is independent in computation with a concurrency trait. The global memory has a low I/O speed, so we storage the individual information in each block's shared memory which has a very fast I/O speed. The global memory is used only in information exchanging between islands and the whole population such as migration and finishing calculation. In order to accelerate convergence, some individuals of each island will migrate to its neighboring island. Shared memory and global memory will exchange data in this step. The random numbers are used in most steps of GA progress. The cuRAND library from CUDA toolkit is used to generate random numbers in parallel.

4 IMPLEMENTATION

4.1 Population

In island based GA, each island has its population with individuals. The population evolves separately on its island. Each individual has its own chromosome. The genes of chromosome are input parameters of a forward modeling. In MT inversion, model parameters, which contain resistivity and depth value of each layer, are used as genes. In this work, an individual has one chromosome which includes parameter of one model.

4.2 Fitness

Here the fitness is equal to objective function. Individuals with the smaller fitness are better ones.

4.3 Selection

Tournament selection is used for this step for crossover operation. Every two neighboring threads (individuals) can be seen a pair. The individual with a less fitness value is chosen as one parent. The other parent is randomly selected from the whole population in each island.

4.4 Crossover

Every pair of individuals will be given a random number which will be compared with the crossover probability to decide whether to perform this operation or not. The arithmetic crossover shown in Equation 3 is performed in this step[8]. The pair of parents is replaced by a pair of off-springs after a crossover operation.

$$\begin{aligned} O_1 &= a \cdot P_1 + (1-a) \cdot P_2 \\ O_2 &= (1-a) \cdot P_1 + a \cdot P_2 \end{aligned} \quad (3)$$

where O_1 and O_2 represent off-springs, P_1 and P_2 represent parents and a called the aggregation weight in arithmetic crossover.

4.5 Mutation

The mutation probability is a constant number. A uniform random number is generated in parallel for each individual. This random number is compared with mutation probability to make decision for mutation. Then a gene of chromosome is selected randomly and changed to a random value within the constraints of model setting. After all steps above, the population is replaced by the newly generated off-springs. Then the fitness of each individual is calculated in parallel. Afterwards the population goes on to the next iteration.

4.6 Migration

The migration operation occurs once for each 10 generations. In this step 10% best individuals are exchanged between two neighboring islands.

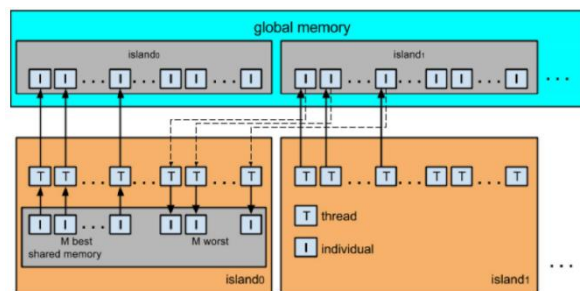


Figure 2 Migration between Islands

Migration operation is illustrated in Figure 2. The exchange is done asynchronously in GPU main memory (called global memory). After being sorted by fitness, M worst individuals in an island are overwritten by M best individuals from a neighboring island. Both sorting and migrations are done in parallel for all individuals.

5 ODEL TEST

We use an artificial model to test our CUDA program. The range of frequency in MT model test is set from 10^{-3} to 10^3 with totally 31 frequency points. The real model is a five layers model (Table 1) and the populations iterate for 3000 generations. The graphic card is NVIDIA A5000 and CPU is Intel(R) i7-14900KF in this work.

The speedups on GPU against CPU are shown in in Figure 3. According to different scales of population sizes and number of islands, the speedup is from the 7.8 times minimum to 757 times maximum. It can be seen that GPU shows its superiority against CPU under a large population size and high number of islands.

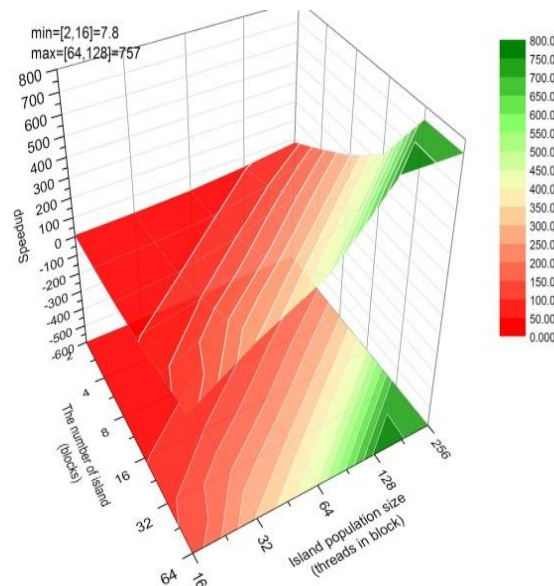


Figure 3 The GPU Speedup of 5-Layers Mode Inversion

Table 1 gives the five layers model setting and GPU inversion result. The number of islands is 16 and population size on each land is 128. The inversion result in Tab.1 shows a good solution which is close to the true mode. The convergence variety with generations is shown in Figure 4. The fitness curve is oscillating at early generations but the convergence becomes stable after about a few generations.

Table 1 Model Setting and GPU Inversion Result of a 5-Layers Artificial Model

Layer	Layer thickness(km)				
	1	2	3	4	5
True model	1	1	2	2	infinity
Inversion model	0.96	0.96	1.70	3.10	--
Model	Resistivity(Ωm)				
	100	1000	100	10	1000
True model	100	1000	100	10	1000
Inversion model	100	512	140	14	1126

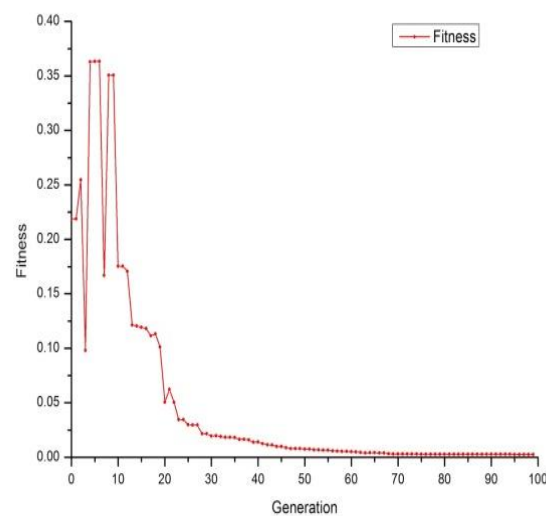


Figure 4 The Convergence of Fitness with Generation

Table 2 presents the final optimal fitness results under different settings of island numbers and population sizes. The best fitness value, 0.00004, is achieved when the number of islands is 64 and the population size per island is 256. However, when the population size is small (e.g., 16), increasing the number of islands yields limited improvement in accuracy. Therefore, in practical applications, an appropriate combination of island number and population size can be selected based on available computational resources to balance efficiency and accuracy.

Table 2 Best Fitness under Different Island Number and Size

island number \ island size	4	16	64
16	0.01225	0.0152	0.00068
64	0.0137	0.00063	0.00014
256	0.00077	0.00013	0.00004

6 CONCLUSION

The high speedup clearly proves that GPU has ability of accelerating the genetic algorithm for solving optimization problems. The experimental result also shows that the migration can significantly improve the efficiency to find the best solutions. The size and interval time of migration significantly affect the results of inversion. A less interval time means a more time consuming but will accelerate the diffusion of the optimal solution. In that condition, a more precise solution will be obtained. How to choose migration scheme need more model tests in the future.

In this work, we present an island based genetic algorithm inversion for MT on CUDA architecture. The model testing gives a good inversion result for the algorithm. The future work will focus to introducing parallel GA to other optimization problems in geophysics.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

FUNDING

This research is supported by The Science and Technology Project of Anhui Provincial Department of Natural Resources “Research on Deep Exploration Technology of Hidden Rock Masses-taking the Luzong Basin in Anhui as an Example” (2024-k-1).

REFERENCES

- [1] Eraser A S. Simulation of genetic systems by automatic digital computers. I. Introduction. Australian Journal of Biological Sciences, 1957, 10: 484-491.
- [2] Holland J H. Genetic algorithms and the optimal allocation of trials. SIAM Journal on Computing, 1973, 2(2): 88-105.
- [3] Everett M E, Schultz A. Two-dimensional nonlinear magnetotelluric inversion using a genetic algorithm. Journal of Geomagnetism and Geoelectricity, 1993, 45(9): 1013-1026.
- [4] Perez-Flores M A, Schultz A. Application of 2-D inversion with genetic algorithms to magnetotelluric data from geothermal areas. Earth, Planets and Space, 2002, 54(5): 607-616.
- [5] Roux E, Moorkamp M, Jones A G, et al. Joint inversion of long-period magnetotelluric data and surface-wave dispersion curves for anisotropic structure: Application to data from Central Germany. Geophysical Research Letters, 2011, 38(5).
- [6] Wait J R. On the relation between telluric currents and the earth's magnetic field. Geophysics, 1954, 19(2): 281-289.
- [7] Wait J R. Theory of magnetotelluric fields. J. Res. NBS D, 1962, 66(5): 509-541.
- [8] Pospichal P, Jaros J, Schwarz J. Parallel genetic algorithm on the cuda architecture. In Applications of Evolutionary Computation. Springer, 2010: 442-451.

HUMAN-AI CO-CREATION SYSTEM FOR KNOWLEDGE WORK BASED ON MULTI-AGENT APPROACH

WeiJing Zhu¹, RunTao Ren^{2*}, Wei Xie¹, CenYing Yang²

¹Guangxi Science and Technology Information Network Center, Nanning 530022, Guangxi, China

²City University of Hong Kong, Kowloon Tong, Hong Kong region 999077, China.

Corresponding Author: RunTao Ren, Email: runtaoren2-c@my.cityu.edu.hk

Abstract: The task of writing a scientific research proposal is complex and highly structured, yet traditional writing methods are typically inefficient. To address this challenge, this study introduces an intelligent writing system leveraging a large language model (LLM). The core contribution is a modular proposal drafting framework that automatically generates multi-chapter application content based on user-specified disciplines and topic questions. The system first creates outlines and content overviews for each chapter through intent recognition, then composes detailed chapter content guided by these outlines. Finally, the system consolidates these components into a complete proposal draft. This modular architecture not only ensures logical consistency across the document but also empowers users to independently refine and optimize individual chapters. To evaluate the system's efficacy, we invited multiple researchers for assessment. Benchmarking against a single-agent LLM demonstrates that our multi-agent system produces proposals with superior content coverage, logical coherence, and user satisfaction, while significantly improving writing efficiency. The proposed modular prompt framework exhibits broader applicability and can be readily extended to other funding application contexts, offering a novel technological approach for advancing intelligent research support systems.

Keywords: Human-AI collaboration; Generative AI; Large Language Model; Intelligent system

1 INTRODUCTION

With the exponential growth of interdisciplinary research, collaborative knowledge creation has become an increasingly essential yet challenging endeavor in scientific communities. From national science foundations to international cooperation programs, researchers must synthesize diverse domain insights into cohesive, high-quality deliverables within strict space and time constraints [1]. At the same time, the number of submissions to major funding agencies worldwide has surged, while success rates continue to decline[2]. This contradiction has also prompted the writing quality of research proposals to meet almost stringent requirements. For example, researchers must accurately condense scientific problems, design rigorous technical routes, and clearly demonstrate the value and feasibility of research within a limited space[3]. However, for the traditional manual writing method, it takes an average of 38 working days for researchers to prepare a new proposal, and 28 working days for a resubmitted proposal, with an overall average time of 34 days per proposal[4]. This dual dilemma of efficiency and quality has become an important obstacle to improving scientific research productivity.

In recent years, automated writing tools based on natural language processing (NLP) have begun to emerge, but their applications are mostly limited to shallow tasks such as grammar checking and text polishing[5]. However, breakthroughs in generative AI (GenAI) have provided new possibilities for text creation. Large language models (LLMs) such as GPT-4 and Claude have demonstrated near-human-level fluency and coherence in general writing tasks, but their application in professional scientific research scenarios still faces severe challenges [6][7][8]. Tasks such as scientific proposal drafting demand not only a deep integration of fragmented ideas but also adherence to specialized discipline standards [9]. These characteristics also expose two gaps in existing AI support systems about writing in the task of generating research proposals: (1) High knowledge barriers: applicants for interdisciplinary research projects often lack an understanding of the writing standards of specific disciplines; (2) Efficiency bottlenecks: inexperienced researchers need to spend a lot of time learning the structure and expression of applications.

A variety of AI-assisted writing paradigms have attempted to address these difficulties[10][11][12]: (1) Template-based Systems (e.g., Research Rabbit, Grantable) rely on pre-designed outlines but lack substantial content-generation functionality; (2) Component-focused Systems solve local issues such as grammar checks (e.g., BERT-based GEC) or literature recommendations, rather than providing a holistic drafting process; (3) General-purpose Systems (e.g., GPT-4, Claude) can generate text across many topics but may struggle with specialized domain accuracy and advanced structural needs. Although these paradigms have made some progress, they have failed to establish a generative paradigm of the "domain-content-structure" trinity, that is, to improve creation efficiency through intelligent interaction while ensuring domain accuracy, content innovation, and structural integrity. Motivated by these gaps, our research focuses on human-AI

collaborative creation: How can we design a system that harnesses GenAI to assist users in producing complex, domain-centered documents with greater speed and consistency?

To validate this co-creation approach, we develop a GenAI support system—tentatively named Proposal AI—and test it on the domain of research proposals. We choose proposals as a prime example because they require strict structural adherence and deep disciplinary insight, offering a rigorous testbed for human–AI collaboration. Our system, however, is not confined to proposals alone; the underlying framework can be generalized to other forms of structured academic or technical writing. Specifically, we implement three core designs: (1) Adaptive Prompt Engineering: Customizing generation strategies for different disciplines and sections to enhance terminological accuracy and logical flow; (2) Human–Machine Collaborative Workflow: Enabling real-time user edits within a multi-agent division of labor, thereby optimizing content quality while providing oversight; (3) Structured Generation Framework: Dynamically creating outlines and guided content to ensure outputs conform to desired standards; users input a project name and discipline, receive a multi-tier outline generated by LLMs, then refine each chapter through specialized agents. By demonstrating these features in the proposal-writing scenario, we showcase a human–AI knowledge co-creation methodology that balances user direction with automated text generation. Our key contributions include:

- (1) We propose a co-creative workflow supporting complex organizational tasks. By integrating domain cues, user interactions, and hierarchical structuring, the system helps LLMs manage the controllability challenges.
- (2) We design a template-driven approach to tailor prompts. This strategy enhances LLM outputs by aligning each section with discipline requirements, ensuring greater logical consistency.
- (3) We introduce a multi-agent architecture wherein different specialized agents handle various brainstorming. This division of labor significantly reduces overall drafting time, facilitating a human–AI synergy that preserves content depth and clarity.

2 RELATED WORK

Recent advances in artificial intelligence have given rise to diverse paradigms for AI-assisted writing, each providing partial solutions to knowledge creation tasks while revealing limitations that become critical in complex or domain-specific contexts. Although we use research proposal writing as a test scenario for our multi-agent framework, the approaches summarized here have broader applicability—and corresponding shortcomings—in tasks requiring structured, domain-aware co-creation.

2.1 Template-Based Systems

Template-based systems rely on pre-defined document structures to guide users in organizing content. They prioritize structural compliance over dynamic content generation, typically employing rule-based mechanisms to validate headings, word counts, and institutional formats. For instance, Mohammad et al. generated literature reviews by extracting citation sentences from academic papers[13], and Jha et al. assembled relevant text fragments to form review sections[14]. These approaches produce academic text via pattern matching and segment extraction but often yield less-cohesive passages. Sun and Zhuge further introduced a template tree to generate literature reviews recursively, organizing multi-document content via dimension and topic nodes [15]. Subsequent systems combined template methods with machine learning—for example, Liu et al. used a BERT classifier to label sentences as background, objectives, methods, and results, then concatenated these segments in a predefined order [16]. Although such strategies introduce partial automation, they essentially extend the template paradigm, which can become rigid or inadequate when the topic or input data exceed the template's scope[17]. Users may also need to inject domain expertise into template structures, raising the barrier to adoption. Furthermore, the often static nature of these templates hinders flexibility and stifles creative freedom, limiting the systems' potential for broader knowledge co-creation tasks.

2.2 Component-Focused Systems

Component-focused systems target specific aspects of the writing process, such as grammar correction, terminology optimization, or literature recommendation. Component-focused systems treat documents as collections of localized components rather than holistic artifacts, focusing on improving specific elements without addressing document-level problems. For example, Kaneko et al. incorporated the pre-trained language model into the encoding-decoding error correction framework and then used its output as additional features for the error correction model[18]. Omelanchuk et al. proposed to treat error correction as a sequence labeling problem, directly predicting the modification operations required for each word, thereby simplifying training and reaching a leading level[19]. These grammar checkers of component-focused Systems based on pre-trained models can already provide grammar-polishing suggestions for application writing that are close to human editing. In addition, scientific proposals require rigorous and unified terminology to avoid inappropriate wording or inconsistency. Component-focused Systems can also verify whether the terminology is used correctly and consistently by using scientific databases or domain corpora. For example, academic writing assistants (such as Writefull) helped authors check whether the manuscript covered key terms by extracting domain keywords from

academic papers[20]. In addition to the above functions, component-focused systems can also be extended to cover support for format, structure, and persuasiveness. For example, SWAN (Scientific Writing AssistaNt) is a type of component-focused system for scientific research paper writing, with built-in expert-developed indicators, which can check and provide feedback on each part of the manuscript from the title, abstract, introduction to the conclusion[21]. The feedback provided by SWAN includes marking where sentences are too long or the wording is inappropriate, and giving suggestions on how to enhance the persuasiveness of the article, such as using more powerful wording, ensuring that the title is consistent with the content, etc. Similar ideas can be applied to research proposal writing: by analyzing whether the various components of the application are complete and coordinated, reminding the author of missing information or disordered structure[22]. Some methods also use machine learning to evaluate the readability and persuasiveness of the text, helping applicants to better meet the expectations of reviewers in terms of wording and argumentation[23]. However, these functions of component-focused systems are usually still based on predefined rules or shallow NLP analysis, with limited understanding of semantics. When the author's creative ideas conflict with the built-in rules of the model, the tool's suggestions may not apply. Therefore, although component-level system improves the local quality of scientific research writing, they are independent of each other: grammar tools do not understand the meaning of the content, and terminology tools do not understand the research background. This separation limits their help to the overall writing quality and logical coherence.

2.3 General-Purpose Systems

General-purpose systems leverage large language models (LLMs) to generate text across a wide spectrum of topics and domains, treating document creation as an end-to-end sequence prediction task. GPT-3 notably demonstrated near-human fluency with minimal prompts in translation, question-answering, and text continuation[24], showcasing LLMs' potential for multi-task generation. Within scientific writing, Wang et al. introduced PaperRobot, which sequentially generated abstracts, conclusions, and even potential future-paper titles starting from a single paper topic [25]. Other researchers experimented with domain-specialized LLMs such as Galactica [26] to automate sections of academic writing. While these approaches display striking fluency and adaptability, their integration with discipline-specific rigor and structured control remains underexplored[27]. Text generated by general LLMs often strays from the hierarchical organization expected in formal documents (e.g., proposals), misuses technical terms, or neglects domain conventions[28]. Consequently, human creators still bear the responsibility of injecting critical scientific insights, reinforcing logic, and ensuring that proposals or other highly structured texts meet standards for academic discourse.

Subsequent work has explored adding explicit retrieval or multi-agent coordination to regain structural control. AutoSurvey divides survey-paper drafting into four LLM-mediated stages—retrieval & outline building, subsection drafting by specialized agents, integration & refinement, and automatic evaluation—thereby overcoming context-window limits and parametric-knowledge gaps when synthesising rapidly expanding literatures [29]. In the adjacent domain of English-for-Academic-Purposes (EAP) writing, AcademiCraft employs a multi-agent architecture to iteratively correct, enrich and explain revisions to scholarly prose, outperforming leading commercial grammar tools on coherence, cohesion and context-sensitive word choice [30].

Despite these advances, tensions remain between flexibility and discipline-specific rigour. Even with retrieval pipelines (as in AutoSurvey) or role-specialised agent teams (as in AcademiCraft), generated text can stray from the hierarchical organisation demanded by formal documents—misplacing methodological details, misusing technical terms, or ignoring disciplinary conventions. Human authors therefore still shoulder ultimate responsibility for injecting critical scientific insight, reinforcing logic, and ensuring compliance with community standards.

As synthesized in Table 1, existing paradigms each address different facets of structured writing. Template-based systems excel at structural control but struggle with topic diversity and creativity; component-focused solutions enhance local quality yet lack global coherence; and general-purpose LLMs provide versatility while falling short on domain precision and robust structuring. These trade-offs extend well beyond proposal drafting to any advanced knowledge-creation task, highlighting a fundamental tension between rigid templates, piecemeal enhancements, and unconstrained text generation.

Table 1 Capability Matrix of Existing Approaches

Paradigm	Structural Control	Content Quality	Domain Accuracy	Logical Coherence
Template-based	High	Medium	Low	Medium
Component-focused	Medium	Low	High	Low
General-purpose Systems	Low	High	Medium	Medium

This capability matrix reveals the need for an integrated approach that unifies domain alignment, flexible content generation, and chapter-level organization. In the context of our study, we focus on proposal writing as a stringent, real-world application scenario for human–AI collaborative knowledge creation. Our multi-agent system, tested on proposals but

equally relevant to other complex writing tasks, aims to overcome the trilemma of structural rigidity, piecewise optimization, and generic LLM output by embedding domain constraints and structured prompts directly into the generation pipeline. As subsequent sections detail, this approach leverages funding agency (or similarly formal) templates as latent constraints while providing multi-phase prompt engineering to balance automated efficiency with scholarly precision.

3 METHOD

This section describes our human-AI co-creative framework for generating complex, domain-specific documents. The method is designed to support knowledge creation in various structured writing contexts requiring domain compliance, multi-chapter organization, and user oversight as shown in figure 1. The approach comprises three key phases: Dynamic Outline Generation, Personalized Refinement, and Multi-Agent Specialized Writing, each formalized with mathematical notation to ensure clarity and reproducibility. Below, we detail each phase and then summarize the overall system architecture.

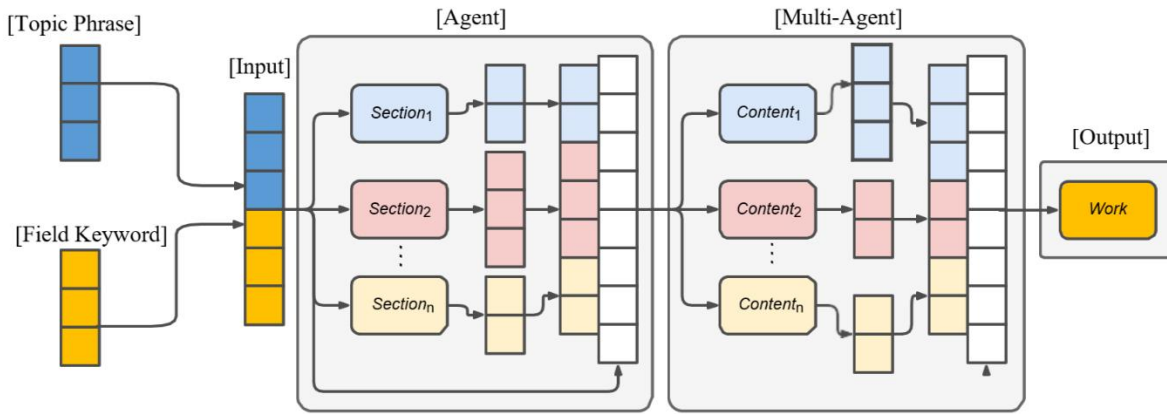


Figure 1 Multi-Agent Method

3.1 Dynamic Outline Generation

Our system begins by producing an initial hierarchical outline based on user inputs: topic phrase T and a field keyword K , where T refers to an overarching subject and K refers to a specific discipline, angle, or methodological focus. These inputs are applicable to various structured content scenarios (e.g., technical reports, detailed literature reviews). Given the input T and K , the system constructs a hierarchical outline $O^{(0)}$ as follows:

Sections: The LLM generates a set of section titles $S_1 = \{s_1^1, s_2^1, \dots, s_m^1\}$, where each s_i^1 corresponds to a major section of the idea.

Subsections: For each subsection title of sections, the LLM generates a set of subsections $S_2^i = \{s_1^2, s_2^2, \dots, s_{n_i}^2\}$, where s_j^2 represents a subsection title under s_i^1 .

Content Instructions: For each subsection s_j^2 , the LLM generates a preliminary content instruction c_j , which provides a brief description of the content to be written in that subsection.

Hence, the outline generation process can be formalized as:

$$O^{(0)} = LLM(T, K) = \bigcup_{i=1}^m \left(s_i^1 \times \bigcup_{j=1}^{n_i} (s_j^2, c_j) \right) \quad (1)$$

where m is the number of Level-1 sections, n_i is the number of Level-2 subsections under s_i^1 . This phase ensures a domain-sensitive skeletal structure, providing a coherent basis for subsequent content creation.

3.2 Personalized Refinement

Next, users refine the outline $O^{(0)}$ through an interactive interface, infusing domain knowledge or personal preferences. This phase allows users to modify the structure and adjust content instructions. The system ensures that user edits are seamlessly integrated into the outline while maintaining structural and logical consistency. The refinement process is driven by user edits ΔU , which include:

Structural Edits: Adding, removing, or revising sections and subsections.

Semantic Edits: Rewriting content instructions c_j to better reflect the user's intent.

This stage embodies human–AI collaborative knowledge creation—users integrate deep expertise into the system-generated framework. The updated outline $O^{(final)}$ is produced via:

$$O^{(final)} = LLM(O^{(0)}, \Delta U) \quad (2)$$

The final outline $O^{(final)}$ is then passed to the next phase for content generation.

3.3 Multi-Agent Specialized Writing

Once the refined outline $O^{(final)}$ is set, the system employs a **multi-agent** architecture $A_1 = \{a_1, a_2, \dots, a_n\}$ to compose each section's text. For each section s_i^1 and subsection s_j^2 in the final outline $O^{(final)}$, the agent a_i generates content through the following two designs:

System Prompt: The system prompt P_s^i is preconfigured to ensure global or organizational guidelines

User Prompt: The user prompt P_u^i dynamically constructed from the outline, including the section/subsection title and the custom instructions c_j .

The agent a_i combines P_s^i and P_u^i to generate the final content for the section and subsection. The content generation process for section s_i^1 and subsection s_j^2 is formalized as:

$$Content_{s_i^1 or s_j^2} = a_i(P_s^i, P_u^i) \quad (3)$$

The outputs of these specialized agents are then integrated to form the complete document:

$$D = \bigcup_{i=1}^m \left(Content_{s_i^1} \times \bigcup_{j=1}^{n_i} Content_{s_j^2} \right) \quad (4)$$

4 EVALUATION

4.1 Experimental Setup

To evaluate the efficacy of our human–AI co-creation approach for knowledge work, we designed an experimental study using research proposal writing as the test scenario. We chose proposals because they demand structured organization, domain-specific rigor, and creative synthesis—attributes reflective of broader complex knowledge-creation tasks.

Our experiments employed Qwen-Turbo, an LLM developed by Alibaba, as the core text generator. We set the model's temperature to 0.7, aiming to balance creativity with coherence. A total of 20 researchers (15 graduate students and 5 professors) participated, representing three diverse academic fields: Management Science & Engineering, Economics, and Computer Science. This disciplinary spread enabled us to observe how different knowledge domains interact with the system's multi-agent design. Before the evaluation began, each participant received a short training session on how to operate our co-creative platform.

We adopted a multi-faceted evaluation methodology to rigorously test how well our multi-agent co-creation framework supports users in a demanding knowledge-creation scenario—namely, writing an extended research proposal. The method combined a user study, which collected subjective feedback through questionnaires, and a baseline comparison against a more traditional single-agent LLM.

Each of the 20 participants used our co-creative system to draft a full research proposal aligned with their expertise or area of interest. When the system finished generating a draft, participants reviewed the content, focusing on key aspects such as logical structure, domain accuracy, and originality. Once a draft proposal was generated, participants reviewed the output. They were then asked to fill out a detailed questionnaire evaluating the system and the generated proposal. The questionnaire captured the participants' ratings on various aspects of the text quality (fluency, coherence, readability), the structural integrity of the proposal, and the originality of the content. It also included items on the usability of the system (how easy and intuitive it was to use) and overall satisfaction with the experience. Participants completed the questionnaire immediately after using the system, ensuring their feedback was based on their fresh experience. This user-centric evaluation allowed us to gather insights into how well the system meets the needs of researchers in practice and how comfortable they are with the automatically generated content.

To benchmark the effectiveness of a multi-agent, structured approach, we compared our system against a single-agent method (e.g., ChatGPT4 and DeepResearch) with straightforward prompts. After finishing their interaction with our co-creative platform, each participant examined a ChatGPT-generated draft on the same project. They then rated that version using the same questionnaire items, enabling direct comparisons across metrics such as structural completeness, coherence, or user satisfaction.

4.2 Evaluation Metrics

We defined a set of evaluation metrics covering both the quality of the generated text and the user experience of the system. These metrics together provide a comprehensive evaluation of the system's performance, balancing output quality and process quality. The key evaluation metrics include:

Quality: Measures the linguistic and narrative quality of the proposal [31]. This includes fluency (naturalness of language and absence of grammatical errors), coherence (logical consistency and flow of ideas throughout the proposal), and readability (clarity and ease of understanding for the reader).

Completeness: Assesses the organizational quality of the proposal [32]. We evaluate completeness in terms of whether all essential sections of a standard research proposal are present and logical structure, meaning the content is well-organized with a clear progression of ideas and a sound argument structure.

Innovativeness: Evaluates the originality and creativity of the content[33]. This reflects whether the proposed research ideas and approaches appear novel. All participants and experts considered if the automatically generated proposal offers fresh insights or interesting research directions.

Usability: Captures the ease of use and user-friendliness of the system[34]. This metric reflects the user experience: how easy it was for participants to interact with the system, understand its prompts, and steer the generation process. It also covers the satisfaction of users with the system's interface and functions.

Efficiency: Measures the time and effort saved by using the system[35]. We looked at two aspects: time to generate (how quickly a complete draft was produced by the system, from the moment the user input the initial idea to the time the final draft was ready), and time to revise (how much additional time the participant needed to revise or polish the AI-generated draft to reach a submission-ready state). This metric is important to gauge whether the system actually speeds up the proposal writing process compared to writing a proposal manually or using simpler tools.

Although tested here on proposals, these metrics reflect broader dimensions of knowledge work—evaluating both content quality and user interaction in high-complexity writing tasks.

4.3 Results And Analysis

We collected questionnaire data from 20 participants who each used two systems: our ProposalAI multi-agent platform and baselines (ChatGPT4 and DeepResearch). While proposals anchored the experiment, the findings illustrate how structured co-creation compares to a more generic LLM workflow in terms of coverage, creativity, and user experience.

Figure 2 shows the average scores for Quality, Completeness, Innovativeness, Usability, and Efficiency. Across all metrics, ProposalAI outperforms the baseline, reflecting the effectiveness of incremental outline-building, section-by-section refinement, and domain-oriented prompts.

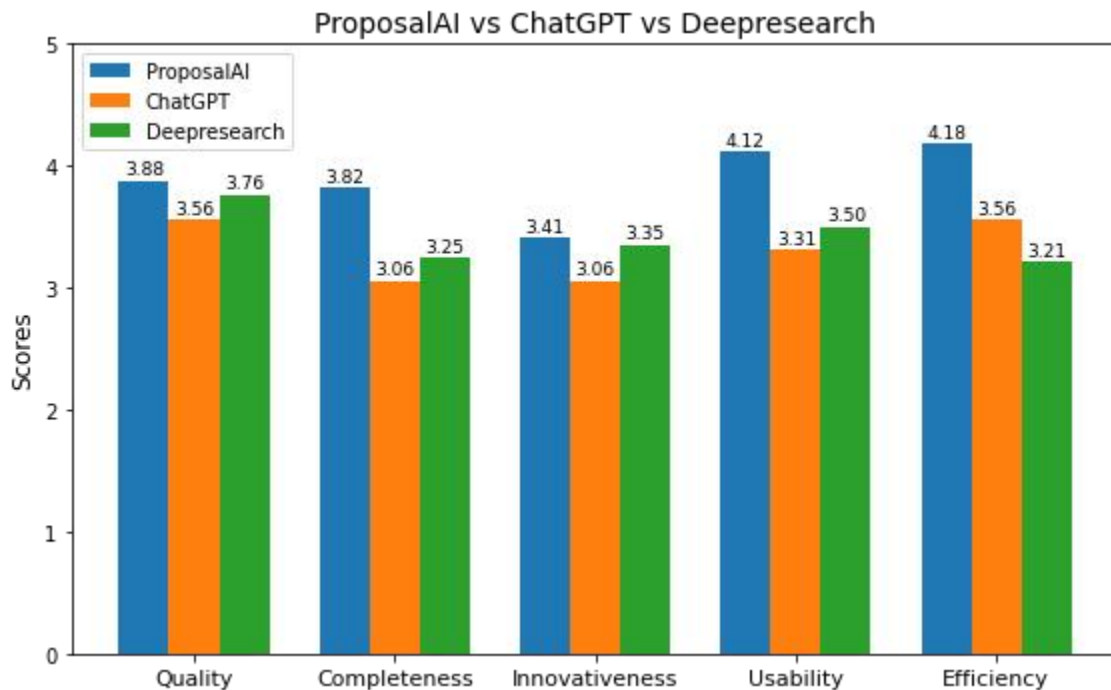


Figure 2 The Evaluation Results

4.3.1 Quality

ProposalAI again tops the chart (3.88). DeepResearch (3.76) narrows the gap to ProposalAI and outperforms ChatGPT (3.56). Participants attributed DeepResearch's edge over ChatGPT to its explanation-driven corrections, yet noted that ProposalAI's section-specific prompting still produced the clearest narrative flow. This difference reflects two primary factors:

Fluency and Coherence: Users commented that, because ProposalAI breaks down each section with domain-specific or section-specific prompts, the model can maintain a better overall flow. In contrast, ChatGPT sometimes introduced minor redundancy or abrupt shifts when attempting to handle everything in one prompt.

Readability: Even though both systems occasionally used generic phrasing, participants said ProposalAI's text was typically clearer and more to-the-point. This is likely due to the multi-phase approach, which reduces the chance of meandering or repeating irrelevant content.

Interestingly, a subset of participants with extensive proposal-writing experience rated both systems lower on language sophistication, indicating that extremely specialized or discipline-specific styles may require more refined prompts or domain-tailored lexical knowledge. Overall, the difference in Quality suggests that scaffolding the writing process via multiple agents effectively boosts clarity and flow, although further fine-tuning or domain adaptation might be needed to reach near-expert human writing levels in specialized fields.

4.3.2 Completeness

Completeness exhibited the widest performance gap: ProposalAI (3.82) markedly outperformed DeepResearch (3.25) and ChatGPT (3.06). Three observations explain this pattern:

Outline Enforcement: ProposalAI's comprehensive outline reduced omissions. DeepResearch, lacking mandatory outline checks, sometimes omitted auxiliary elements but still covered more sections than ChatGPT. Because the system begins by generating and refining a comprehensive outline with mandatory sections (e.g., introduction, literature review, methods, anticipated results), participants were less likely to overlook crucial parts of a standard proposal. This structural scaffolding makes omissions less probable.

Progressive Detailing: The multi-agent approach iteratively fills in sections, ensuring each receives adequate attention. By contrast, ChatGPT's single-pass generation sometimes skipped or glossed over essential content (e.g., feasibility analysis or budget justification) if the prompt was not explicit enough.

Participant Guidance: The outline phase encouraged participants to add custom sub-sections or expand on domain-specific concerns (e.g., "Ethical Considerations"), thus pushing the final output toward greater comprehensiveness.

Qualitative comments from participants confirm that the clarity and forced coverage of the Outline Generation step were key reasons for higher Completeness scores. Several participants mentioned feeling more confident that "nothing important was missing." In short, a well-structured, multi-step pipeline appears crucial for producing thorough research proposals that meet standard academic or funding agency expectations.

4.3.3 Innovativeness

ProposalAI's Innovativeness score (3.41) modestly surpasses DeepResearch (3.35) and ChatGPT (3.06). Our analysis suggests two main factors:

Opportunity for Customization: By prompting users to refine the outline and letting them inject new ideas at different stages, ProposalAI can incorporate domain insights that lead to less generic or formulaic text. DeepResearch or ChatGPT's single prompt sometimes defaulted to "safe" or "boilerplate" suggestions.

Limitations of AI Creativity: Despite the advantage, participants generally felt that neither system truly substitutes for a researcher's unique intellectual contribution. Users with advanced domain knowledge noted that if they only provided cursory instructions, the content would still be somewhat formulaic. This indicates that while multi-agent scaffolding can facilitate creative thinking, genuine scientific innovation still heavily depends on the user's active input.

Multiple participants also pointed out that the system can only reassemble known concepts. This feedback implies that further improvements (e.g., deeper domain integration, synergy with recent literature or specialized databases) could yield even more innovative proposals.

4.3.4 Usability

Usability scores reveal ProposalAI (4.12) > DeepResearch (3.50) > ChatGPT (3.31). From user feedback, we identified several design elements that contributed to higher satisfaction:

Stepwise Interaction: Rather than having everything happen in one large output, participants found the multi-phase approach more transparent and manageable. This process "felt natural," mirroring the mental steps of outlining, drafting, and refining.

Fine-Grained Control: Breaking the proposal into sections allowed users to intervene more precisely. They could refine each part (e.g., method, background) without risking the rest of the text. ChatGPT's single-shot approach often needed repeated re-prompts to fix local issues in one section without inadvertently modifying correct parts elsewhere. DeepResearch improved over ChatGPT yet lacked ProposalAI's slot-specific guidance.

Guided Prompts: Clear instructions for each section (e.g., "research objectives," "anticipated results") gave participants confidence they were focusing on the right aspects. Some praised the interface for "not letting me forget a key element."

Nevertheless, about one-fifth of participants wished for even more direct integration of external documents (e.g., references, prior proposals) into the multi-agent workflow. They felt such a feature would further streamline the writing process.

Overall, high Usability scores reaffirm that a well-structured interface and guided prompts can significantly enhance user experience beyond what a general-purpose LLM alone can offer.

4.3.5 Innovativeness

Efficiency likewise favored ProposalAI (4.18). ChatGPT (3.56) slightly exceeded DeepResearch (3.21) because DeepResearch's explanatory cycle added turnaround time. Users reported noticeable reductions in:

Initial Draft Time: Thanks to the forced outline stage, many participants stated that they overcame "writer's block" quickly. The system auto-populated a skeleton with relevant subheadings, so participants did not have to figure out organizational flow from scratch. DeepResearch's explanations lengthened iteration time despite yielding clearer text than ChatGPT.

Revision Effort: Because the generated drafts tended to be more logically organized and complete, the subsequent editing or fine-tuning stage required less time. In the ChatGPT condition, participants indicated they often had to revisit the prompt multiple times and manually add missing sections, leading to more iterative overhead.

In some open-ended responses, participants acknowledged that if a proposal was highly specialized or if they needed extensive references to advanced literature, the AI needed more custom prompts or expansions. However, even in these scenarios, the foundational structure and partial content still saved them time relative to starting with a blank document or a one-shot generation from ChatGPT. Overall, the multi-agent design gave them fewer "back-and-forth loops," thereby boosting perceived efficiency.

5 SUMMARY AND CONCLUSION

Overall, the multi-agent system scored notably higher in Completeness, Usability, and Efficiency—key dimensions for any collaborative knowledge work. Although improvements in innovativeness and deeper domain adaptation remain open areas for future research, our study underscores the promise of dividing content generation across structured prompts and specialized agents.

In conclusion, the user study suggests that ProposalAI outperforms baselines in most critical dimensions of research proposal writing. Notably, the largest gains appear in Completeness (thanks to structured outlines), Usability (stepwise guidance and better user control), and Efficiency (less revision time). While there is still room for growth in fostering truly novel content, the multi-agent methodology evidently provides a more reliable framework for generating comprehensive, coherent proposals. This underscores the importance of combining domain-specific structuring with user-driven refinement when applying large language models to specialized tasks like grant writing. DeepResearch consistently surpasses ChatGPT in Quality, Completeness, Innovativeness and Usability, confirming the benefit of explanation-oriented agent collaboration, though its longer feedback loop reduces perceived efficiency. These results underscore the efficacy of template-aware, multi-phase prompting for complex scholarly writing. Future work should explore combining DeepResearch-style explanatory feedback with ProposalAI's structural scaffolding, alongside deeper domain adaptation and automated citation support.

Future enhancements also might involve deeper integration with domain knowledge bases, real-time citation management, or iterative refinement loops that incorporate external critiques. Nonetheless, current findings validate that the multi-agent approach holds promise for improving research proposal drafting, potentially reducing the cognitive overhead and time investment traditionally associated with this complex writing task.

6 DISCUSSION AND FUTURE WORK

While the multi-agent architecture demonstrably improves structural guidance and user control, the data layer remains a bottleneck in three respects. First, the system relies exclusively on the frozen parametric knowledge of the underlying LLM; it cannot query up-to-date bibliographic databases or domain-specific corpora during generation. As a result, references must be inserted manually, and citation accuracy is susceptible to hallucination. Second, all empirical results were obtained on a single, moderately sized evaluation set ($n = 20$ proposals) drawn from three academic fields. This dataset is insufficient to capture the full diversity of research domains, styles, and disciplinary conventions, limiting external validity. Future work will address both issues by (i) integrating authenticated retrieval modules for real-time access to scholarly indices and domain-specific databases, and (ii) conducting a multi-institutional study with a substantially larger and more diverse participant pool encompassing senior academics, industry practitioners, and non-English speakers. Such expansions will enable finer-grained error analysis, strengthen external validity, and inform domain-tailored prompt engineering.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

- [1] Lindgreen A, Di Benedetto CA, Verdich C, et al. How to write really good research funding applications. *Industrial Marketing Management*, 2019, 77: 232-239.

- [2] Ren R, Ma J, Zheng Z. Large language model for interpreting research policy using adaptive two-stage retrieval augmented fine-tuning method. *Expert Systems with Applications*, 2025, 278: 127330.
- [3] Locke LF, Spirduso WW, Silverman SJ. *Proposals that work: A guide for planning dissertations and grant proposals*. Sage Publications, 2013.
- [4] Herbert DL, Barnett AG, Clarke P, et al. On the time spent preparing grant proposals: an observational study of Australian researchers. *BMJ Open*, 2013, 3(5): e002800.
- [5] Russo F. Automated content writing tools and the question of objectivity. *Digital Society*, 2023, 2(3): 50.
- [6] Wu J, Yang S, Zhan R, et al. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 2025: 1-66.
- [7] Ren R, Ma J, Luo J. Large language model for patent concept generation. *Advanced Engineering Informatics*, 2025, 65: 103301.
- [8] Wang Y, Guo Q, Yao W, et al. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems*, 2024, 37: 115119-115145.
- [9] Kallet RH. How to write the methods section of a research paper. *Respiratory Care*, 2004, 49(10): 1229-1232.
- [10] Sharma R, Gulati S, Kaur A, et al. Research discovery and visualization using ResearchRabbit: A use case of AI in libraries. *COLLNET Journal of Scientometrics and Information Management*, 2022, 16(2): 215-237.
- [11] Zhou YC, Zheng Z, Lin JR, et al. Integrating NLP and context-free grammar for complex rule interpretation towards automated compliance checking. *Computers in Industry*, 2022, 142: 103746.
- [12] Pal S, Bhattacharya M, Islam MA, et al. AI-enabled ChatGPT or LLM: A new algorithm is required for plagiarism-free scientific writing. *International Journal of Surgery*, 2024, 110(2): 1329-1330.
- [13] Mohammad S, Dorr B, Egan M, et al. Using citations to generate surveys of scientific paradigms. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, June: 584-592.
- [14] Jha R, Finegan-Dollak C, King B, et al. Content models for survey generation: A factoid-based evaluation. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2015, July: 441-450.
- [15] Sun X, Zhuge H. Automatic generation of survey paper based on template tree. *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*, 2019, September: 89-96.
- [16] Liu S, Cao J, Yang R, Wen Z. Generating a structured summary of numerous academic papers: Dataset and method. *International Joint Conferences on Artificial Intelligence*, 2022.
- [17] Zhu K, Feng X, Feng X, et al. Hierarchical catalogue generation for literature review: A benchmark. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, December: 6790-6804.
- [18] Kaneko M, Mita M, Kiyono S, et al. Encoder-decoder models can benefit from pre-trained masked language models in grammatical error correction. *arXiv preprint, arXiv:2005.00987*, 2020.
- [19] Omelianchuk K, Atrasevych V, Chernodub A, et al. GECToR-Grammatical error correction: Tag, not rewrite. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2020, July: 163-170.
- [20] Mitchell P, Riedlinger M, Goldenfein J, et al. Research GenAI: Situating generative AI in the scholarly economy. *AoIR Selected Papers of Internet Research*, 2024.
- [21] Kinnunen T, Leisma H, Machunik M, et al. SWAN-scientific writing AssistaNt: A tool for helping scholars to write reader-friendly manuscripts. *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, April: 20-24.
- [22] Adhi P. Exploring the use of ChatGPT as a supporting tool in writing research proposals: EFL students' perspectives. *Doctoral dissertation, UIN Sunan Gunung Djati Bandung*, 2024.
- [23] Bai X, Stede M. A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. *International Journal of Artificial Intelligence in Education*, 2023, 33(4): 992-1030.
- [24] Brown T, Mann B, Ryder N, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020, 33: 1877-1901.
- [25] Wang Q, Huang L, Jiang Z, et al. PaperRobot: Incremental draft generation of scientific ideas. *arXiv preprint, arXiv:1905.07870*, 2019.
- [26] Taylor R, Kardas M, Cucurull G, et al. Galactica: A large language model for science. *arXiv preprint, arXiv:2211.09085*, 2022.
- [27] Huang J, Tan M. The role of ChatGPT in scientific communication: Writing better scientific review articles. *American Journal of Cancer Research*, 2023, 13(4): 1148.
- [28] Seckel E, Stephens BY, Rodriguez F. Ten simple rules to leverage large language models for getting grants. *PLOS Computational Biology*, 2024, 20(3): e1011863.
- [29] Wang Y, Guo Q, Yao W, et al. Autosurvey: Large language models can automatically write surveys. *Advances in Neural Information Processing Systems*, 2024, 37: 115119-115145.

- [30] Du Z, Hashimoto K. AcademiCraft: Transforming writing assistance for English for academic purposes with multi-agent system innovations. *Information*, 2025, 16(4).
- [31] Daudaravicius V. Automated evaluation of scientific writing: AESW shared task proposal. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 2015, June: 56-63.
- [32] Cai Y, Ziad M. Evaluating completeness of an information product. *AMCIS 2003 Proceedings*, 2003, 294.
- [33] Dean DL, Hender J, Rodgers T, et al. Identifying good ideas: Constructs and scales for idea evaluation. *Journal of Association for Information Systems*, 2006, 7(10): 646-699.
- [34] Davis FD. Technology Acceptance Model: TAM. In: Al-Suqri MN, Al-Aufi AS, eds. *Information Seeking Behavior and Technology Adoption*. Hershey, PA: IGI Global; 2015: 205–219.
- [35] Michailidis A, Rada R, Gouma P. A study of efficiency in computer-supported collaborative writing. *Journal of Intelligent Systems*, 1994, 4(1-2): 133-162.

