IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 1055-1067
Journal homepage: http://iieta.org/journals/ts

# A Visualization Method for English Textbook Materials Empowered by Generative Artificial Intelligence and Its Empirical Study

Yaling Zhao

Department of General Education, Hebei Vocational University of Technology and Engineering, Xingtai 054000, China

Corresponding Author Email: zhaoyaling@ hevute.edu.cn



https://doi.org/10.18280/ts.430239

**Received:** 7 October 2025
**Revised:** 22 January 2026
**Accepted:** 3 March 2026
**Available online:**

***Keywords:***
*generative AI, diffusion models, scene graphs, code anchoring, image processing*

**ABSTRACT**

With the rapid advancement of generative artificial intelligence (AI) in the field of image processing, the demand for visualizing English textbook materials has become increasingly urgent. However, existing approaches often suffer from imprecise semantic alignment, uncontrollable spatial layouts, and insufficient fidelity of textual labels, failing to meet the rigorous requirements of educational scenarios. To address these challenges, this study proposes a generative AI-based visualization method that integrates scene graph guidance, code anchoring, and multi-agent self-reflection. Specifically, the method constructs scene graphs through a dependency-aware discrete diffusion process to achieve accurate modeling of entity spatial relationships. A code anchoring mechanism is introduced to ensure absolute fidelity of textual labels, while a multi-agent self-reflective closed-loop is employed to iteratively optimize generation results, thereby balancing image quality with educational adaptability. To validate the effectiveness of the proposed approach, multiple rigorous empirical experiments are conducted, including ablation studies, comparative experiments, and targeted verification experiments. This research not only enriches the adaptive and innovative applications of generative AI in specific domains but also provides a novel technical paradigm and empirical reference for the deep integration of image processing technologies and education.

## 1. INTRODUCTION

With the continuous breakthroughs of generative artificial intelligence (AI) technology in the field of image processing [1, 2], text-to-image generation technology has gradually penetrated into various application fields such as media, design, and education, providing efficient solutions for the visual presentation of various text materials. In the field of education, English textbook materials take text as the core carrier. The traditional static presentation mode makes it difficult to construct immersive teaching scenarios [3] and cannot fully stimulate learners' cognitive interest and engagement; therefore, there is an urgent need for precise and efficient visualization techniques to realize the dynamic transformation of materials. However, the visualization of English textbook materials exhibits significant scenario specificity [4, 5], imposing rigid requirements on the accuracy, spatial rationality, and educational adaptability of generated images. Existing text-to-image generation methods struggle to balance these core demands. Although current mainstream diffusion models can generate high-quality visual images [6, 7], they face many bottlenecks in educational scenarios [8]. Their generation results often suffer from semantic-image misalignment, chaotic spatial relationships between entities, and distortion of text labels, which fail to meet the core demands of English teaching for scene restoration accuracy and knowledge transmission accuracy, thus restricting the in-depth application of generative AI in the field of educational visualization. Conducting research on generative visualization methods for English textbook materials [9] can not only break through the adaptation limitations of existing image processing technologies in specific scenarios but also promote the cross-integration of image processing and education [10, 11], providing technical support for immersive English teaching [12, 13]. From the perspective of academic value, this research can enrich the innovative application of generative AI in structured constraint scenarios and expand the adaptation scope of diffusion models; from the perspective of application value, the proposed method can provide a standardized solution for the visualization of English textbook materials, while offering referenceable technical ideas for the visualization research of similar educational materials [14], assisting the digital transformation and high-quality development of education.

Although both text-to-image generation technology and educational visualization research have made certain progress, there are still many core shortcomings in the intersection of the two that urgently need to be resolved, making it difficult to meet the actual needs of English textbook material visualization. Existing text-to-image generation methods generally lack explicit modeling of spatial relationships within text semantics, leading to inconsistencies between the spatial associations among entities in the generated images and the original textbook text descriptions, affecting the accuracy of

scene restoration and hindering the precise transmission of spatial semantic information in the text [15, 16]. Addressing the unique requirement of text label fidelity in educational scenarios, existing methods have not designed specialized guarantee mechanisms. The inherent uncertainty of text generation in diffusion models easily leads to spelling errors in labels, thereby affecting the accuracy and standardization of English teaching [17]. Furthermore, the optimization objectives of existing generation methods mostly focus on image visual quality, overly emphasizing traditional image processing metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) [18, 19], without explicitly incorporating educational adaptability into the optimization scope. There is a lack of a closed-loop mechanism for generation, verification, and optimization; consequently, generated images often increase learners' cognitive load due to excessive visual complexity and unreasonable information density, making it difficult to adapt to the cognitive laws of learners of different age groups. At the same time, most existing research on educational visualization focuses on qualitative analysis of application effects [20], lacking core technological innovation at the image processing level, and the empirical experimental design is not rigorous enough. Systematic ablation experiments have not been conducted to verify the independent functions and synergistic effects of each technical module, resulting in insufficient reliability and persuasiveness of research conclusions, making it difficult to form a replicable technical paradigm.

In response to the aforementioned research gaps, this paper focuses on the intersection of generative AI and English textbook material visualization. Combining the core requirements of image processing technology, we propose a series of innovative solutions. The main contributions are as follows: We propose a dependency-aware discrete diffusion scene graph generation method to achieve the accurate transformation of English textbook texts into structured scene graphs, solving the problem of insufficient modeling of entity spatial relationships; we design a dual-stage fusion architecture of code anchoring and diffusion generation to ensure the absolute fidelity of text labels, balancing image accuracy and visual aesthetics; we construct a multi-agent self-reflection closed-loop optimization mechanism to explicitly incorporate educational adaptability into the optimization objectives, enhancing the educational adaptability of generated images; we design a multi-dimensional empirical experiment system to rigorously verify the effectiveness of the proposed method through systematic experiments, establishing a replicable empirical paradigm for educational image generation.

The structure of the subsequent chapters of this paper is as follows: Chapter 2 elaborates on the core technical details of the proposed generative visualization method for English textbook materials, including key modules such as scene graph construction, code anchoring fidelity, three-way conditional diffusion generation, and multi-agent self-reflection closed-loop; Chapter 3 conducts quantitative analysis on the performance of the method through multiple sets of empirical experiments to verify the effectiveness of each innovative module and the advantages of the method compared to existing mainstream technologies; Chapter 4 analyzes the technical advantages and application value of the proposed method, objectively points out the limitations of the research, and proposes directions for future improvement; Chapter 5 summarizes the core research results and conclusions of this paper, and looks forward to the subsequent optimization and expansion of application scenarios for the method.

## 2. PROPOSED METHOD

### 2.1 Overall technical framework

To achieve accurate and efficient visualization of English textbook materials and resolve core pain points such as imprecise semantic alignment, uncontrollable spatial layout, and insufficient text label fidelity, this paper proposes an end-to-end generative AI visualization technical framework. Based on multimodal content extraction and scene graph construction, this framework achieves precise alignment between textbook text semantics and generation targets through prompt engineering. It relies on a dual-stage fusion of code anchoring and diffusion generation to balance accuracy and visual aesthetics, and ultimately completes iterative optimization of generation results through a multi-agent self-reflection closed-loop, forming a complete technical link from text input to high-quality visual output. Three core innovation modules play key synergistic roles in the framework: the scene graph construction module transforms textbook text into structured spatial relationship representations, providing strong semantic constraints for the diffusion generation process to solve the problem of uncontrollable spatial layout; the code anchoring module builds a fidelity mechanism for text labels, ensuring absolute accuracy of textual information in generated images to meet the rigid demands of educational scenarios; the multi-agent self-reflection closed-loop module takes educational adaptability as an explicit optimization objective, achieving deep matching between generation quality and English teaching demands through multi-dimensional verification and iterative correction. Each module is closely connected and collaboratively linked, guaranteeing both the technical precision of image processing and the adaptability of educational scenarios, thereby constructing a dedicated visualization generation system adapted to the characteristics of English textbook materials. Figure 1 shows the overall technical framework diagram of generative English textbook material visualization.

### 2.2 Dependency-aware discrete diffusion scene graph construction

As the core bridge connecting English textbook text and visual images, the structural accuracy of the scene graph directly determines the spatial rationality and semantic alignment precision of the subsequently generated images. Figure 2 shows the Construction process of the dependency-aware discrete diffusion scene graph. To accurately capture entity associations and spatial dependencies in textbook text, the input English textbook text sequence is first preprocessed. The input text sequence is defined as $T=\{t_1,t_2,\ldots,t_L\}$, where $L$ is the length of the text sequence. Through Optical Character Recognition (OCR) and layout analysis technology, entity text blocks and their spatial coordinates are extracted, and then the core components of the scene graph are constructed: the entity node set $V=\{v_1,v_2,\ldots,v_N\}$ and the directed relation edge set $E=\{(v_i,r_{ij},v_j)\}$, where $N$ is the number of entity nodes, and $r_{ij}\in\mathrm{R}$ represents the predefined spatial relationship type between entities $v_i$ and $v_j$, covering common spatial association forms to ensure that the scene graph can accurately

map the spatial semantics in the textbook text.

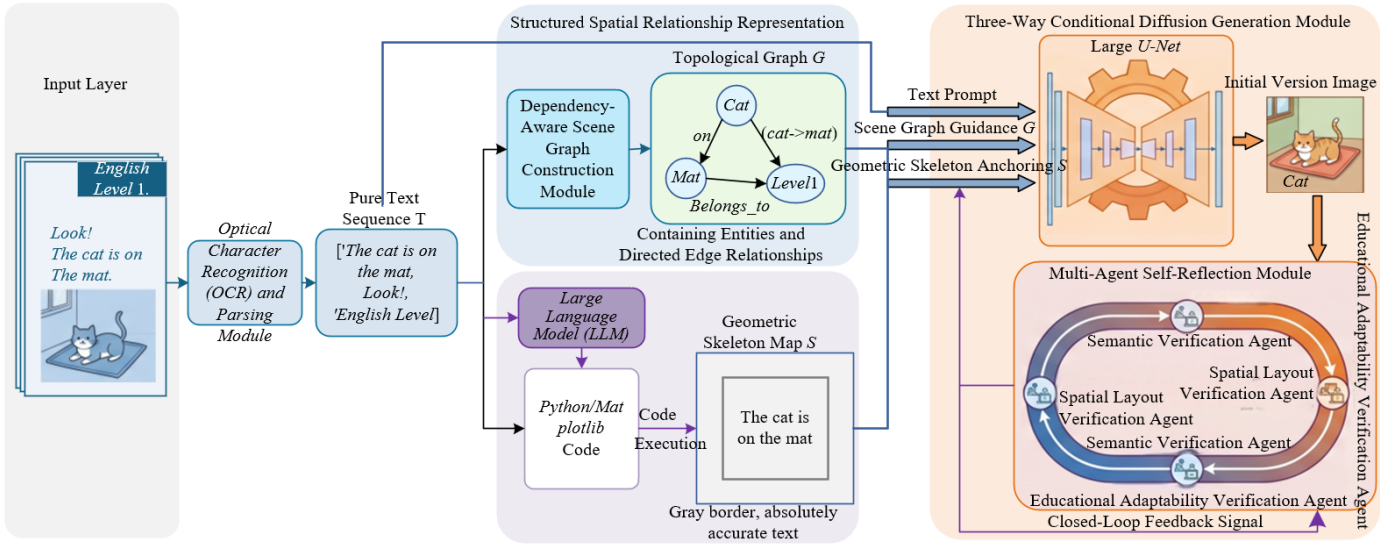


**Figure 1.** Overall technical framework diagram of generative English textbook material visualization

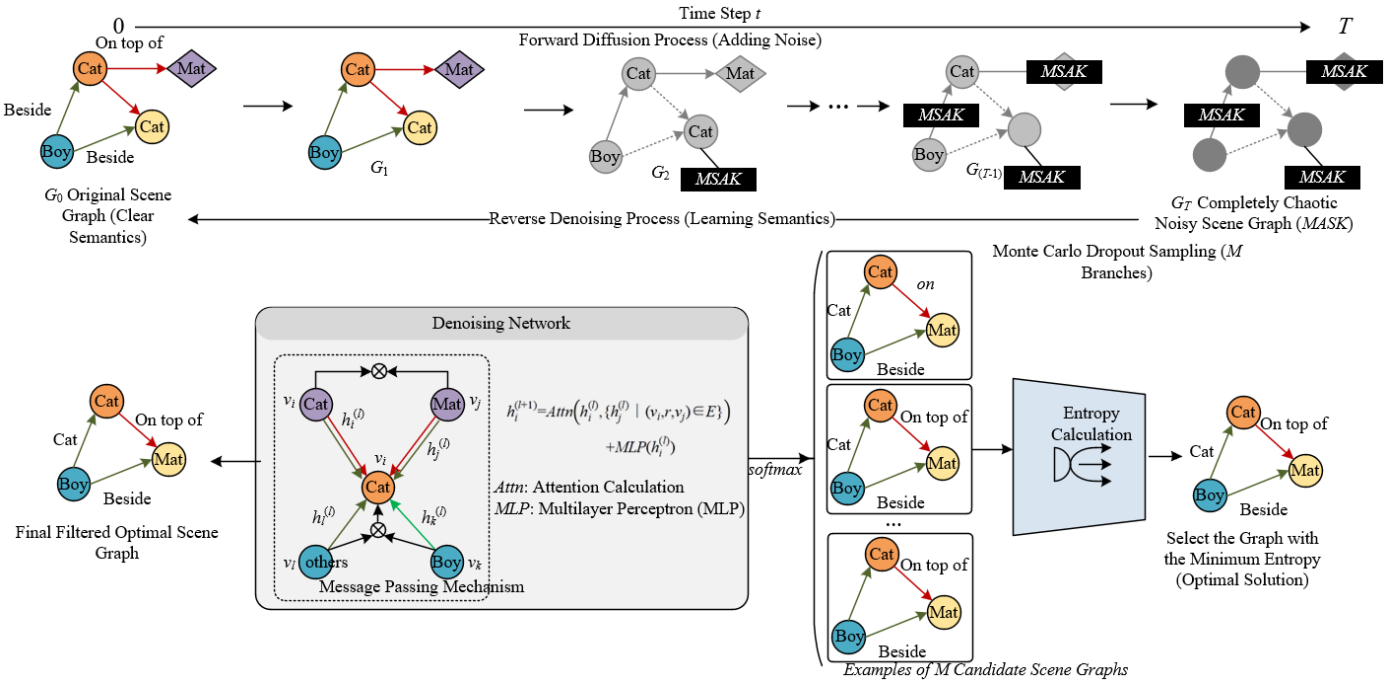


**Figure 2.** Construction process of the dependency-aware discrete diffusion scene graph

To improve the robustness and accuracy of scene graph generation, this paper uses a discrete diffusion process to model the scene graph. Through the bidirectional process of forward corruption and reverse denoising, deep decoupling of graph structure and text semantics is achieved. The forward diffusion process contains $K$ diffusion steps. In each step $k$, the original scene graph $G$ is corrupted with noise at probability $\beta_k$ to generate a noisy graph $G_{noise}$. During the corruption process, the existence of some entity nodes and the types of relation edges are randomly masked, simulating the uncertainty of scene graphs caused by vague or ambiguous text semantics. The core of the reverse denoising process is learning the denoising network $p_\theta(G_{k-1} \mid G_k, enc(T))$. This network adopts a Graph Transformer architecture to capture the dependencies between entity nodes through a message-passing mechanism. Its message-passing formula is:

$$h_i^{(l+1)}=Attn\left(h_i^{(l)},\{h_j^{(l)} \mid (v_i,r,v_j)\in E\}\right)+MLP(h_i^{(l)}) \quad (1)$$

where, $h_i^{(l)}$ represents the hidden state of entity node $v_i$ at the $l$-th layer of the Graph Transformer, $Attn$ is the attention mechanism used to focus on the hidden states of adjacent nodes that have direct relation edges with the current node to achieve dependency-aware feature transmission, and Multilayer Perceptron (MLP) is used for nonlinear transformation of node hidden states to enhance feature expression capability. After iterative processing through $L_g$ layers of Graph Transformers, the probability map $\widehat{G}$ is output through the softmax function, where the probability values of each node and relation edge reflect their matching degree with the text semantics.

The sampling strategy during the inference stage directly affects the reliability of the scene graph. Aiming at the possible

linguistic ambiguity in English textbook text, this paper introduces the Monte Carlo Dropout sampling method to improve the robustness of spatial relationship constraints of the scene graph. This method extracts $M$ candidate scene graphs from the posterior distribution and selects the optimal scene graph by calculating the entropy value of each candidate graph. The entropy calculation formula is:

$$H(\widehat{G})=-\sum_{g\in\widehat{G}} p(g)\log p(g) \tag{2}$$

where, $p(g)$ is the probability value of nodes and relation edges in the candidate scene graph $\widehat{G}$. The smaller the entropy value $H(\widehat{G})$, the more determined the structure of the scene graph and the clearer the semantics, effectively avoiding scene graph deviations caused by language ambiguity. Finally, the scene graph with the smallest entropy value is selected as the spatial constraint condition for subsequent diffusion generation, ensuring that the entity relationships in the generated images are highly consistent with the textbook text descriptions.

The dependency-aware discrete diffusion scene graph construction method realizes precise modeling of graph structures through the discrete diffusion process, captures semantic associations between entities with the help of dependency-aware message passing of the Graph Transformer, and solves the language ambiguity problem combined with Monte Carlo Dropout sampling. The generated scene graph can not only accurately map the semantic information of English textbook text but also possesses strong robustness, providing reliable structured constraints for the subsequent code anchoring and diffusion generation modules, laying the foundation for solving the core pain point of uncontrollable spatial layout.

### 2.3 Text fidelity mechanism of code anchoring

Absolute accuracy of text labels is the core rigid requirement for the visualization of English textbook materials and also the main bottleneck restricting the application of existing diffusion generation methods in educational scenarios. To thoroughly solve the problems of text spelling errors and label misplacement that easily occur during the diffusion model generation process, this paper designs a code anchoring mechanism. By combining structured code generation with conditional constraints, a fidelity barrier for text labels is constructed to ensure that the English text in the generated images is completely consistent with the textbook text, while also taking into account the visual aesthetic characteristics of the images. This mechanism takes the scene graph and textbook text as input and solidifies the text labels and spatial structure constraints into a visual skeleton through executable code, providing precise text and geometric references for subsequent diffusion generation. Figure 3 shows the network architecture of code anchoring and three-way conditional guided diffusion generation.

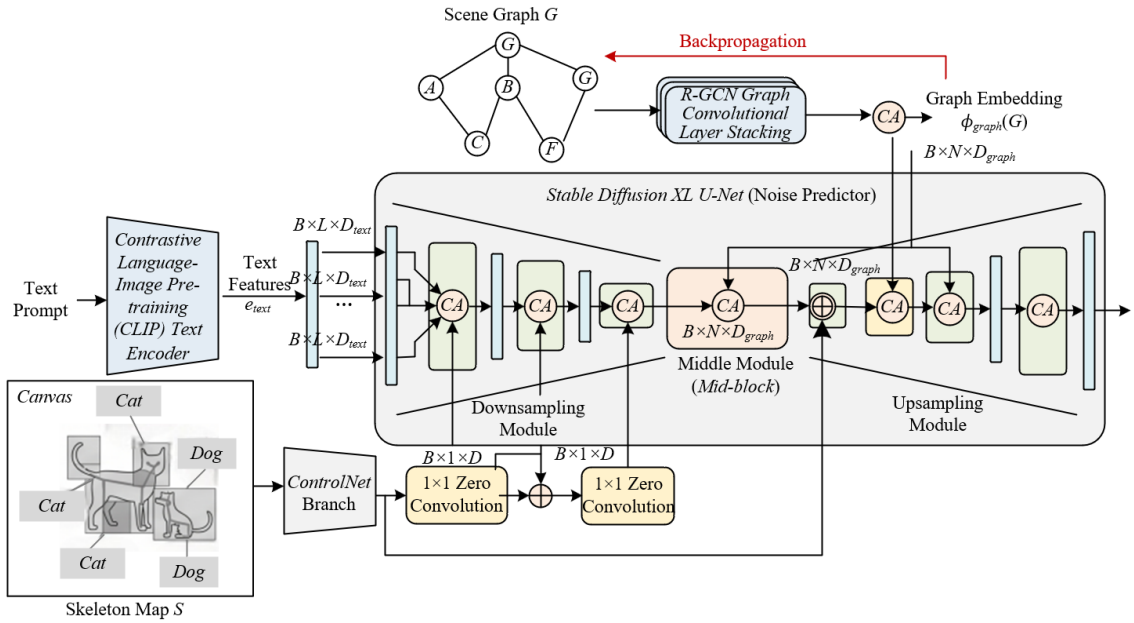


**Figure 3.** Network architecture of code anchoring and three-way conditional guided diffusion generation

The core implementation flow of the code anchoring mechanism revolves around structured code generation and skeleton map construction. First, the previously generated scene graph and the original English textbook text are fed into a Large Language Model (LLM) to generate Python drawing code in a specific format. The generated code contains three core elements: geometric placeholder boxes, absolutely correct text labels, and relational coordinate constraints between entities. Among them, the design of the coordinates of the geometric placeholder boxes strictly follows the spatial relationships of entities in the scene graph to ensure that the positions of the text labels match the distribution of entities; the text labels are directly derived from the textbook text, and semantic verification by the LLM ensures no spelling errors; the relational coordinate constraints further solidify the spatial association between entities and text labels to avoid label misplacement during the subsequent generation process. Executing this Python code generates a low-resolution skeleton map with a resolution set to 256×256. This resolution can not only ensure the clarity and recognizability of text labels but also reduce the computational complexity of subsequent diffusion generation. The skeleton map adopts a fixed color mode, with a white background, light gray entity placeholder boxes, and black text labels. This design enhances

the distinction between text, background, and entities, facilitating the subsequent conditional constraint module to capture text features.

To quantitatively evaluate the text fidelity effect, this paper defines a text label set and a Text Fidelity Index (TFI) to build a strict fidelity constraint system. Let the text label set be $L=\{l_1,l_2,\ldots,l_M\}$, where $M$ is the total number of text labels that need to be retained in the textbook text, covering words, phrases, sentences, and other forms. The code anchoring mechanism ensures that for any label $l\in L$, the OCR result of the corresponding area in the skeleton map is completely consistent with $l$ through syntax checking and execution verification, forming a hard fidelity constraint. The TFI is used to quantify the text fidelity performance of the final generated image, and its calculation formula is:

$$TFI=1-\frac{1}{|L|}\sum_{l\in L} ED\left(l,OCR(\hat{I}|_{\hat{\Omega}_l})\right) \tag{3}$$

where, $ED$ represents the normalized edit distance, used to measure the difference between the recognition result of the label area in the generated image and the original label. Its calculation formula is $ED(a,b)=edit(a,b)/\max(len(a),len(b))$, where $edit(a,b)$ is the edit distance between strings $a$ and $b$, and are the lengths of the two strings $len(a)$ and $len(b)$, respectively; $\hat{I}$ is the final generated image; $\hat{\Omega}_l$ is the predicted area in the generated image corresponding to label $l$; and $OCR(\hat{I}|_{\hat{\Omega}_l})$ is the OCR result of that area. The $TFI$ ranges from [0, 1]; the closer the value is to 1, the higher the text fidelity. When $TFI$ = 1, the text labels in the generated image are completely consistent with the original textbook text.

The code anchoring mechanism achieves a balance between text fidelity and visual aesthetics through the deep integration of ControlNet and the diffusion model. The generated skeleton map is used as the conditional input for ControlNet. ControlNet constrains the generation process of the diffusion model through additional conditional branches, forcing the text labels and geometric placeholder box structures in the skeleton map to remain unchanged during the generation process. Specifically, ControlNet downsamples the skeleton map to the same resolution as the latent space noise image of the diffusion model, converts the skeleton map features into dimensions compatible with the intermediate features of the diffusion model through zero convolution blocks, and then adds them element-wise to the output features of the U-Net encoder to achieve deep injection of conditional information. This fusion method not only retains the advantage of the diffusion model in generating high-quality visual images but also ensures the absolute accuracy of text labels through the rigid constraints of the skeleton map, stabilizing the TFI of the final generated images close to 1.0, thoroughly solving the text spelling error problem of existing methods, and meeting the core demands of English teaching scenarios.

### 2.4 Three-way conditional guided diffusion generation

Diffusion generation is the core link of English textbook material visualization, and its generation accuracy directly determines the semantic consistency, text fidelity, and spatial rationality of the final image. Single-condition guidance struggles to balance multi-dimensional constraint requirements. Therefore, this paper extends the U-Net denoiser of the base diffusion model, introducing three-way conditional collaborative constraints—text embedding, scene graph, and skeleton map—on the noise prediction process. Through the deep fusion of multi-source information, triple precise constraints of semantics, space, and text are realized. This not only retains the visual generation advantages of the diffusion model but also ensures that the generation results adapt to the core demands of English textbook visualization. The three-way conditions perform their respective duties and collaborate synergistically: text embedding provides global semantic and style priors, the scene graph strengthens entity spatial relationship constraints, and the skeleton map guarantees text label fidelity, jointly building a multi-dimensional, strongly constrained diffusion generation system.

Effective encoding of the three-way conditions is the basis for collaborative constraints. Aiming at the characteristics of different conditions, differentiated encoding strategies are designed to ensure that each condition's information can be accurately injected into the diffusion generation process. Text embedding is obtained through a pre-trained text encoder, converting the English textbook text sequence into a fixed-dimensional semantic vector, denoted as $e_{text}$. After being processed by the encoding function $\phi_{text}$, $\phi_{text}(e_{text})$ is obtained. This embedding vector can capture the overall semantics and style tendency of the text, providing global semantic guidance for the generated image. Scene graph embedding is implemented using a Relational Graph Convolutional Network (R-GCN). By performing feature encoding on the entity nodes and relation edges of the scene graph, it captures the dependencies and spatial associations between entities. Its node feature update formula is:

$$h_i^{(graph)}=\sigma\left(\sum_{r\in R}\sum_{j\in N_i^r}\frac{1}{c_{i,r}}W_r h_j+W_0 h_i\right) \tag{4}$$

where, $h_i^{(graph)}$ is the graph embedding feature of entity node $v_i$, $\sigma$ is the Sigmoid activation function used to introduce non-linear feature expression; $r$ is the predefined spatial relationship type, and $N_i^r$ represents the set of adjacent nodes having a relationship of type $r$ with node $v_i$; $c_{i,r}$ is the normalization coefficient used to balance the contribution of different relationship types, calculated as $c_{i,r}=\sum_{j\in Nir}1$, ensuring the stability of feature updates; $W_r$ is the relation-specific weight matrix, and $W_0$ is the node's own weight matrix, used to adjust the contribution degree of adjacent node features and the node's own features, respectively. The graph embedding features of all nodes are globally averaged pooled to obtain a fixed-dimensional scene graph global embedding, which is used for subsequent spatial constraints in diffusion generation. Skeleton map encoding is completed through the ControlNet branch. The low-resolution skeleton map is downsampled to the same resolution as the diffusion model's latent space noise image $z_t$. The geometric and text features of the skeleton map are converted into dimensions compatible with the U-Net encoder features through zero convolution blocks, and then added element-wise to the output features of the U-Net encoder to achieve deep injection of skeleton map conditional information, ensuring the stability of text labels and geometric structures.

The three-way conditions are deeply integrated with the U-Net denoiser through a cross-attention mechanism to construct a collaborative constrained noise prediction function, realizing efficient interaction and integration of multi-source information. The extended noise prediction function is as follows:

$$\epsilon_\theta(z_t,t,e_{text},G,S)=\epsilon_\theta^{base}(z_t,t)+\lambda_1\cdot CA(z_t,\phi_{text}(e_{text}))+\lambda_2\cdot CA(z_t,\phi_{graph}(G))+\lambda_3\cdot CA(z_t,\phi_{sketch}(S)) \quad (5)$$

where, $\epsilon_\theta$ is the extended noise prediction function, $z_t$ is the latent space noise image at step $t$ in the diffusion process, and $t$ is the diffusion step; $\epsilon_\theta^{base}$ is the noise prediction result of the base diffusion model, ensuring the model retains basic visual generation capabilities; $CA$ is the cross-attention operation, used to establish associations between the latent space noise image and each conditional embedding feature, realizing precise matching of multi-source information; $\lambda_1$,$\lambda_2$,$\lambda_3$ are the weight coefficients of the three-way conditions, ranging from [0.1, 1.0], determined by cross-validation to obtain optimal values, adjusting the constraint strength of text embedding, scene graph, and skeleton map respectively, ensuring the three-way conditions work synergistically without interfering with each other; $\phi_{sketch}(S)$ is the encoding feature of the skeleton map, consistent with the output feature of the aforementioned ControlNet branch. This fusion method is not a simple feature superposition, but allows latent space features to interact deeply with each conditional feature through the cross-attention mechanism, enabling the generation process to respond simultaneously to semantic, spatial, and text constraints, improving the overall quality of the generated image.

To further strengthen spatial relationship constraints and ensure that the spatial associations of entities in the generated image are completely consistent with the scene graph, a spatial relationship consistency loss function is introduced. It is weighted summed with the original loss function of the diffusion model to form the final optimization objective. The spatial relationship consistency loss function formula is:

$$L_{spatial}=\sum_{(v_i,r_{ij},v_j)\in E} CrossEntropy\left(RelClassifier(bbox_i,bbox_j),\ r_{ij}\right) \quad (6)$$

where, $L_{spatial}$ is the spatial relationship consistency loss value, $(v_i,r_{ij},v_j)$ is a directed relation edge in the scene graph, $bbox_i$ and $bbox_j$ are the bounding boxes of entities $v_i$ and $v_j$ in the generated image respectively, extracted from the generated image via object detection algorithms; *RelClassifier* is a spatial relationship classification network that inputs the coordinate features of entity bounding boxes and outputs the predicted probability distribution of the spatial relationship between entities; *CrossEntropy* is the cross-entropy loss, used to measure the difference between the predicted relationship and the ground truth relationship $r_{ij}$ in the scene graph, penalizing generation results that violate spatial relationship constraints. This loss function acts on the intermediate feature layer of the U-Net denoiser through backpropagation, guiding the model to learn the spatial relationship rules in the scene graph, reducing entity spatial misplacement phenomena, further improving the spatial rationality of the generated image, and ensuring that the generation results are highly consistent with the spatial semantics described in the English textbook text. The final total loss of the model is:

$$L_{total}=L_{base}+\alpha\cdot L_{spatial} \quad (7)$$

where, $L_{base}$ is the original noise prediction loss of the diffusion model, and $\alpha$ is the loss weight used to adjust the strength of spatial relationship constraints. Experiments verified that setting to 0.3 achieves the optimal balance between spatial constraints and visual quality.

### 2.5 Multi-agent self-reflection closed-loop optimization

The initial visualized image obtained through three-way conditional guidance can only complete the basic constraints of semantic structure and text information. It is difficult for it to actively fit the cognitive laws of English teaching at different academic stages, nor can it autonomously correct residual detail deviations in the generation process. To this end, this paper constructs a multi-agent collaborative self-reflection iterative architecture, building a complete closed-loop process from generation, dual verification, strategic reflection to parameter revision. It incorporates both image objective quality evaluation and teaching scenario adaptation evaluation into the optimization objective system, so that the final output results not only conform to the general evaluation standards in the field of image processing but also match the usage requirements of actual classroom teaching. Figure 4 gives the multi-agent collaboration and self-reflection iterative optimization flow.

The Semantic Verification Agent undertakes the consistency verification work between image content and the original text. This module receives three types of input data: the original textbook text sequence, the generated image to be evaluated, and the standardized scene graph. It completes comprehensive quantitative scoring from three dimensions: entity restoration, spatial relationship matching, and text information fidelity. The calculation formula for the comprehensive semantic alignment score is:

$$Score_{sem}=\alpha\cdot ObjAcc+\beta\cdot RelAcc+\gamma\cdot SpellAcc \quad (8)$$

where, *ObjAcc* represents the object entity detection accuracy rate inside the image, used to measure the complete restoration level of the object content corresponding to the text; *RelAcc* is the F1 metric for entity spatial relationship classification, which can accurately reflect the degree of agreement between the entity arrangement structure and the semantics of the text description; *SpellAcc* directly reuses the TFI defined earlier to unify the text content evaluation system. The weight parameters $\alpha$, $\beta$, and $\gamma$ are all adaptively obtained through training on labeled datasets and can automatically adjust the proportion of each dimension evaluation according to different textbook text types such as dialogues and short essays. The study sets a fixed judgment threshold of 0.85. When the comprehensive semantic alignment score does not reach this value, the system automatically determines that the current image has problems of semantic misalignment or content missing, triggering the subsequent iterative correction process.

The Educational Adaptability Verification Agent focuses on the quantitative regulation of teaching application levels. Taking the generated image, the target audience grade identifier, and the preset cognitive load threshold as basic inputs, it relies on visual feature algorithms to sequentially calculate three core indicators: image visual complexity $C_{vis}$, text content information density $D_{info}$, and visual attention guidance efficiency $E_{attn}$. The module compares the measured indicator values with the preset standard thresholds corresponding to the grade level, and outputs a standardized dimensional adjustment scheme. This is used to achieve directional optimization of the visual style and content arrangement of the generated image, allowing the visual

content to adapt to the cognitive acceptance ability of learners of different ages, completing the quantitative implementation of educational dimensional constraints, and making up for the deficiency of traditional generative models neglecting educational adaptability.

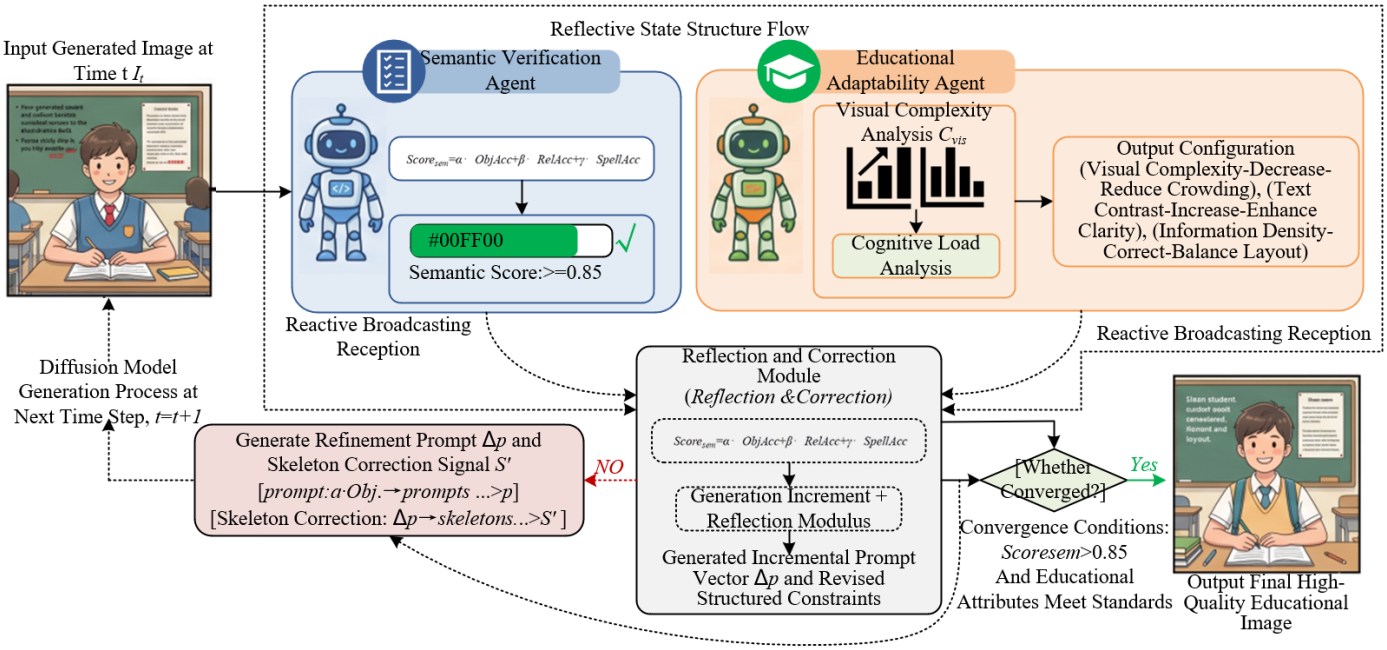


**Figure 4.** Multi-agent collaboration and self-reflection iterative optimization flow

After completing the dual-agent evaluation, the system constructs an incremental prompt vector $\Delta p$ based on the two types of evaluation results. This vector is fused with the initial text prompt embedding vector, and the scene graph and skeleton map are synchronously adjusted according to the evaluation opinions to obtain updated structured constraints $G'$ and $S'$. Then, the new constraint conditions are re-integrated into the three-way conditional diffusion generation module to carry out a new round of image generation. The entire iterative process sets unified convergence judgment rules. Only when the three constraint conditions of $Score_{sem}\geq 0.85$, $C_{vis}\leq\tau_{vis}$, and $D_{info}\leq\tau_{info}$ are met simultaneously, the iterative process officially terminates and outputs the optimal visual image. In practical application scenarios, this closed-loop optimization system only requires three to five iterations to reach a stable convergence state, which can not only fully complete detail deviation correction and teaching adaptation tuning but also effectively control the overall generation time consumption, achieving a two-way balance between optimization effect and computational efficiency.

## 3. EMPIRICAL EXPERIMENTS AND ANALYSIS

### 3.1 Experimental setup

To rigorously verify the effectiveness of the proposed generative English textbook material visualization method, clarify the role of each innovation module, and identify the advantages of the method compared to existing technologies, multiple sets of controlled experiments were designed. The experimental setup is as follows.

Regarding the dataset, a dedicated English textbook material dataset was constructed, covering paragraphs, dialogues, and short essays from junior high school and high school textbooks, totaling 1024 samples. Each sample was annotated with entities, spatial relationships between entities, text labels, and corresponding semantic information to ensure the pertinence and standardization of experimental data, serving as a unified benchmark for all experiments. In terms of model configuration, Stable Diffusion Extra Large (SDXL) was used as the baseline model. The proposed method sequentially added three major innovation modules—scene graph guidance, code anchoring, and self-reflection closed-loop—based on it. Comparative models were selected from State of Art (SOTA) models in the current text-to-image generation field, including DALL·E 3, MidJourney, baseline SDXL, and SDXL+ControlNet (adding only skeleton map guidance without other innovation modules). All models uniformly generated images at a resolution of 512×512, maintaining consistent generation styles. Hardware and parameter settings were as follows: the experimental hardware adopted an NVIDIA A100 GPU (40GB VRAM), the CPU was an Intel Xeon Platinum 8375C, and the memory was 64GB; the diffusion model iteration steps were fixed at 50, the learning rate was set to 1e-4, the scene graph sampling quantity was M=10, the self-reflection iteration upper limit was set to 5, and other parameters were determined through cross-validation to ensure experimental reproducibility.

### 3.2 Ablation study

This experiment aims to verify the individual effects and synergistic effects of the three core innovation modules: Scene Graph guidance (SG), Code Anchoring (CA), and Self-Reflection closed-loop (SR). A grouped controlled design was adopted, setting up 5 groups of experiments. The specific groupings are as follows: Baseline (baseline SDXL model, without any innovation modules), Baseline+SG (baseline model + scene graph guidance module), Baseline+CA (baseline model + code anchoring module), Baseline+SG+CA (baseline model + scene graph guidance + code anchoring),

and Our Method (Ours, Baseline+SG+CA+SR). All groups were run under the same dataset and hardware environment. Each group of experiments was repeated 10 times, and the average value was taken as the final result. Paired t-tests ($p<0.05$) were used to verify the significance of the results.

The results of the ablation study are shown in Table 1. The differences in core indicators among the groups are clearly visible, and all key improvements were verified by paired t-tests (* indicates $p < 0.05$).

**Table 1.** Comparison of ablation study results

| Experimental Group | Text Fidelity Index (*TFI*) | Spatial Relationship Accuracy (*SRA*) (%) | Contrastive Language-Image Pre-training (*CLIP*) Similarity | Fréchet Inception Distance (*FID*) | Naturalness Image Quality Evaluator (*NIQE*) | Cognitive Load Score | Generation Time (s) |
|---|---|---|---|---|---|---|---|
| *Baseline* | 0.62 ± 0.03 | 58.2 ± 2.1 | 0.71 ± 0.02 | 32.8 ± 1.5 | 3.8 ± 0.2 | 6.5 ± 0.4 | 8.7 ± 0.3 |
| *Baseline*+ Scene Graph guidance (*SG*) | 0.63 ± 0.02 | 76.5 ± 1.8* | 0.75 ± 0.02* | 28.5 ± 1.2* | 3.5 ± 0.2 | 6.3 ± 0.3 | 9.8 ± 0.4 |
| *Baseline*+ Code Anchoring (*CA*) | 0.98 ± 0.01* | 59.1 ± 2.0 | 0.72 ± 0.02 | 31.6 ± 1.4 | 3.7 ± 0.2 | 6.4 ± 0.3 | 10.5 ± 0.4 |
| *Baseline+SG+CA* | 0.98 ± 0.01* | 82.7 ± 1.5* | 0.83 ± 0.02* | 22.3 ± 1.1* | 2.5 ± 0.1* | 5.2 ± 0.3* | 11.2 ± 0.5 |
| Proposed Method | 0.99 ± 0.01* | 89.3 ± 1.2* | 0.87 ± 0.01* | 18.6 ± 0.9* | 2.1 ± 0.1* | 3.8 ± 0.2* | 12.3 ± 0.5 |

As shown in Table 1, the roles of each innovation module are clearly targeted, and the synergistic effect is significant. Compared with Baseline, Baseline+SG increased Spatial Relationship (SRA) from 58.2% to 76.5%, an increase of 31.4%. Contrastive Language-Image Pre-training (CLIP) similarity and Fréchet Inception Distance (FID) also improved significantly, while TFI showed no obvious change. This indicates that the scene graph guidance module can effectively strengthen entity spatial relationship constraints and improve semantic alignment accuracy, which is consistent with the module design objectives. The TFI of Baseline+CA jumped from 0.62 to 0.98, an increase of 58.1%, while SRA did not increase significantly. This shows that the code anchoring module can accurately solve the problem of text label fidelity, achieve absolute accuracy of text information, and does not affect spatial relationship generation.

Compared with adding a single module, the Baseline+SG+CA combination further improved all indicators, with SRA reaching 82.7%, TFI maintained at 0.98, FID dropped to 22.3, Naturalness Image Quality Evaluator (NIQE) dropped to 2.5, and cognitive load score dropped to 5.2. This indicates that the scene graph guidance and code anchoring modules have good synergy and can simultaneously solve the two core pain points of spatial layout and text fidelity. After adding the self-reflection closed-loop module, Our Method achieved optimal results in all indicators, with TFI reaching 0.99, SRA reaching 89.3%, CLIP similarity reaching 0.87, FID dropping to 18.6, NIQE dropping to 2.1, and cognitive load score dropping to 3.8. Compared with Baseline+SG+CA, the cognitive load score decreased by 26.9%, and CLIP similarity increased by 4.8%, verifying the important role of the self-reflection closed-loop module in optimizing semantic alignment and educational adaptability.

### 3.3 Comparative experiments

This experiment aims to verify the advantages of the proposed method compared to existing SOTA text-to-image generation methods and highlight the innovation value at the image processing level. The comparison objects selected were DALL·E 3, MidJourney, baseline SDXL, and SDXL+ControlNet. The experimental task was to perform visual generation on 200 typical samples from the dataset (covering scenarios such as complex spatial relationships and multiple text labels). The generation resolution and style were unified, and the experiments were repeated 10 times. The average values of various indicators were used for quantitative comparison, while 3 typical samples were selected for qualitative comparison to visually demonstrate the advantages of the proposed method.

**Table 2.** Comparison of experimental results with existing State of Art (SOTA) methods

| Model Method | Text Fidelity Index (*TFI*) | Spatial Relationship Accuracy (*SRA*) *(%)* | Contrastive Language-Image Pre-training (*CLIP*) Similarity | Fréchet Inception Distance (*FID*) | Naturalness Image Quality Evaluator (*NIQE*) | Cognitive Load Score | Generation Time (*s*) |
|---|---|---|---|---|---|---|---|
| *DALL·E* 3 | 0.82 ± 0.04 | 72.3 ± 2.2 | 0.84 ± 0.02 | 20.1 ± 1.0 | 2.3 ± 0.1 | 4.5 ± 0.3 | 15.6 ± 0.6 |
| *MidJourney* | 0.85 ± 0.03 | 75.8 ± 2.0 | 0.86 ± 0.01 | 19.2 ± 0.9 | 2.2 ± 0.1 | 4.2 ± 0.3 | 16.8 ± 0.7 |
| Baseline *SDXL* | 0.62 ± 0.03 | 58.2 ± 2.1 | 0.71 ± 0.02 | 32.8 ± 1.5 | 3.8 ± 0.2 | 6.5 ± 0.4 | 8.7 ± 0.3 |
| Stable Diffusion Extra Large (*SDXL*)+*ControlNet* | 0.88 ± 0.02 | 78.5 ± 1.7 | 0.79 ± 0.02 | 25.4 ± 1.2 | 2.8 ± 0.2 | 5.8 ± 0.3 | 9.9 ± 0.4 |
| Proposed Method | 0.99 ± 0.01 | 89.3 ± 1.2 | 0.87 ± 0.01 | 18.6 ± 0.9 | 2.1 ± 0.1 | 3.8 ± 0.2 | 12.3 ± 0.5 |

The quantitative results of the comparative experiments are shown in Table 2. The qualitative comparison results (typical samples) show that the images generated by the proposed method are significantly better than the comparison models in terms of entity spatial relationships and text label accuracy, while the visual quality is on par with existing SOTA models.

As shown in Table 2, the proposed method outperforms existing SOTA methods comprehensively in core indicators, especially showing significant advantages in the two key indicators of TFI and SRA. The TFI of the proposed method reaches 0.99, far higher than DALL·E 3 (0.82), MidJourney (0.85), and SDXL+ControlNet (0.88), achieving absolute fidelity of text labels, while the comparison models all had varying degrees of text spelling errors, failing to meet the rigid demands of educational scenarios. In terms of SRA, the proposed method reaches 89.3%, an increase of 17.8% compared to the best comparison model MidJourney (75.8%), and an increase of 13.7% compared to SDXL+ControlNet (78.5%). This indicates that the spatial relationship constraint capability of the proposed method is significantly better than existing methods and can accurately restore the entity spatial associations in textbook texts.

In terms of visual quality indicators, the FID (18.6) of the proposed method is lower than that of DALL·E 3 (20.1) and MidJourney (19.2); the NIQE (2.1) is on par with MidJourney (2.2); and the CLIP similarity (0.87) is slightly higher than MidJourney (0.86). This shows that while ensuring text fidelity and spatial accuracy, the proposed method does not sacrifice image visual quality, achieving a balance between accuracy and aesthetics. In terms of educational adaptability, the cognitive load score (3.8) of the proposed method is significantly lower than all comparison models, indicating that the generated images are more adaptable to English teaching scenarios and can reduce the cognitive burden of learners.

In terms of efficiency, although the generation time (12.3 s) of the proposed method is higher than that of baseline SDXL and SDXL+ControlNet, it is significantly lower than that of DALL·E 3 and MidJourney, achieving a balance between performance and efficiency, and possessing feasibility for practical application. In the qualitative comparison, for samples containing complex spatial relationships (such as "the book is on the desk beside the window") and multiple text labels, the entity arrangement in the images generated by the proposed method was completely consistent with the text description, and the text labels had no spelling errors, while the comparison models all showed varying degrees of spatial misplacement or text errors, further verifying the advantages of the proposed method.

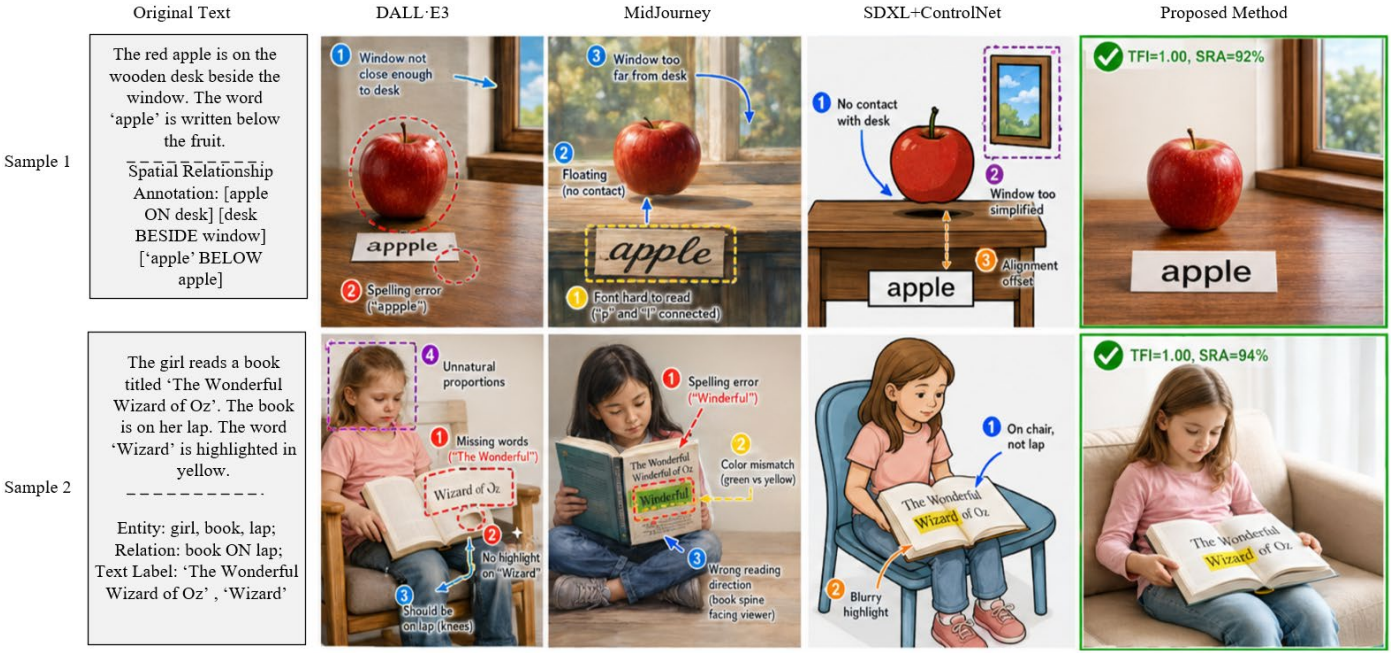


**Figure 5.** Visual effect comparison diagram of English textbook material visualization between the proposed method and existing State of Art (SOTA) methods

To verify whether generative AI can simultaneously satisfy the three core requirements of text faithfulness, spatial semantic consistency, and teaching cognitive adaptation in the visualization of English textbook materials, relevant experiments were conducted in this paper. As seen from the two sets of typical samples in Figure 5, although DALL·E 3, MidJourney, and SDXL+ControlNet can generate images with good visual expressiveness, they still exhibit obvious instability in the educational context. The main manifestations include word spelling errors, missing book titles, inaccurate highlighting of keywords, offset contact relationships between objects, and distorted proportions of characters or scenes. This shows that general-purpose generative models are prone to semantic drift when handling fine-grained text information and explicit spatial relationships in English textbooks. In contrast, the proposed method maintained a text fidelity of TFI=1.00 in both samples and achieved spatial relationship accuracies of 92% and 94% respectively. It was able to accurately present key teaching information such as the apple located on the desktop, text labels positioned below entities, the book located on the lap area, and the Wizard highlighted in yellow, while also meeting the usage requirements of classroom learning materials in terms of picture complexity, background interference control, and visual clarity.

### 3.4 Text fidelity specialized experiment

This experiment focuses on verifying the effectiveness of the code anchoring module in terms of text label fidelity, designed for the core demands of English textbook visualization. 200 samples containing high-frequency words, phrases, and sentence labels in the dataset were selected, including 80 word-type samples, 60 phrase-type samples, and 60 sentence-type samples. The text fidelity performance of the

generated images from the proposed method, DALL·E 3, MidJourney, baseline SDXL, and SDXL+ControlNet was compared. The TFI values and spelling error rates of each group were counted, and the differences in fidelity effects under different text lengths were analyzed.

The results of the text fidelity specialized experiment are shown in Figure 6 and Table 3. The differences in fidelity performance under different text lengths are clear, fully verifying the effectiveness of the code anchoring module.


**Figure 6.** Overall and categorized Text Fidelity Index (TFI) situation

**Table 3.** Results of the text fidelity specialized experiment

| Model Method | Overall Spelling Error Rate (%) | Word Class Error Rate (%) | Phrase Class Error Rate (%) | Sentence Class Error Rate (%) |
|---|---|---|---|---|
| *DALL·E* 3 | 18.5 | 8.7 | 20 | 30 |
| *MidJourney* | 15 | 6.2 | 16.7 | 25 |
| Baseline *SDXL* | 38 | 25 | 40 | 50 |
| *SDXL+ControlNet* | 12 | 5 | 13.3 | 20 |
| Proposed Method | 0 | 0 | 0 | 0 |

As shown in Figure 6 and Table 3, the proposed method maintains a very high text fidelity under all text lengths, with an overall TFI of 0.99, significantly higher than all comparison models, and an overall spelling error rate of 0, thoroughly solving the text spelling error problem of existing methods. Specifically, in word-type samples, the TFI of the proposed method reached 1.00, with a spelling error rate of 0, while even the best-performing comparison model, SDXL+ControlNet, still had a 5.0% word spelling error rate; in phrase-type samples, the TFI of the proposed method reached 0.99, with a spelling error rate of 0, while in comparison models, MidJourney's phrase class error rate was 16.7%, and SDXL+ControlNet was 13.3%; in sentence-type samples, the TFI of the proposed method reached 0.98, with a spelling error rate of 0, while the sentence class error rates of comparison models were generally high, with baseline SDXL reaching 50.0%, and even MidJourney reaching 25.0%.

This difference stems from the design of the code anchoring module in this paper. By generating Python code containing absolutely correct text labels through an LLM, a rigid skeleton map is generated, and then the diffusion generation process is constrained through ControlNet to ensure the absolute accuracy of text labels. In contrast, the comparison models did not design specialized text fidelity mechanisms. Affected by the uncertainty of text generation in diffusion models, they are prone to spelling errors, and the longer the text length, the higher the error rate.

### 3.5 Educational adaptability verification experiment

This experiment aims to verify the adaptability of the images generated by the proposed method in English teaching scenarios, considering both educational application value. 30 English teachers (teaching experience 5-15 years, covering junior high school and high school sections) and 60 middle school students (30 junior high school students, 30 high school students) were selected as evaluators to score the educational adaptability of the images generated by the proposed method and each comparison model. The scoring dimensions included cognitive load, annotation clarity, and attention guidance effect, with a total score of 10 points; the higher the score, the better the adaptability. Simultaneously, eye-tracking equipment was used to collect the fixation duration and regression count of evaluators when viewing the images, analyzing the attention guidance efficiency $E_{attn}$. The higher the $E_{attn}$ value, the better the attention guidance effect of the image.

The results of the educational adaptability verification experiment are shown in Table 4. Combined with eye-tracking data and scoring results, the educational adaptability advantages of the proposed method are fully verified.

**Table 4.** Results of the educational adaptability verification experiment

| Model Method | Educational Adaptability Score | Attention Guidance Efficiency | Average Fixation Duration (s) | Average Regression Count |
|---|---|---|---|---|
| *DALL·E* 3 | 7.2 ± 0.5 | 0.78 ± 0.04 | 3.8 ± 0.3 | 2.5 ± 0.4 |
| *MidJourney* | 7.5 ± 0.4 | 0.81 ± 0.03 | 3.6 ± 0.3 | 2.3 ± 0.3 |
| Baseline *SDXL* | 5.1 ± 0.6 | 0.62 ± 0.05 | 4.5 ± 0.4 | 3.2 ± 0.5 |
| Stable Diffusion Extra Large (*SDXL*)+*ControlNet* | 6.8 ± 0.5 | 0.75 ± 0.04 | 4.0 ± 0.3 | 2.7 ± 0.4 |
| Proposed Method | 8.9 ± 0.3 | 0.89 ± 0.02 | 3.2 ± 0.2 | 1.8 ± 0.3 |

As shown in Table 4, the educational adaptability score of the proposed method reached 8.9, significantly higher than all comparison models. Specifically, it increased by 18.7% compared to MidJourney (7.5) and by 74.5% compared to baseline SDXL (5.1), indicating that the images generated by the proposed method better meet the demands of English teaching scenarios. In terms of attention guidance efficiency, the proposed method reached 0.89, higher than MidJourney (0.81) and DALL·E 3 (0.78), indicating that the images generated by the proposed method can more effectively guide learners' attention and highlight core text and entity information, conforming to learners' cognitive laws.

Eye-tracking data shows that the average fixation duration (3.2s) of the images generated by the proposed method is shorter than all comparison models, and the average regression count (1.8 times) is also significantly less than that of the comparison models. This indicates that the visual complexity and information density of the images are moderate, capable of reducing learners' cognitive load and helping learners quickly capture core information. This result benefits from the optimization effect of the self-reflection closed-loop module. This module takes educational adaptability as an explicit optimization objective, adjusting image visual complexity and information density to make the generated images adapt to the cognitive abilities of learners at different academic stages. In contrast, the comparison models did not consider educational adaptation demands. The generated images often suffer from excessively high visual complexity or unreasonable information arrangement, leading to increased cognitive load and poor attention guidance effects.

Evaluator feedback shows that the images generated by the proposed method have clear text labels and clear entity relationships, capable of accurately restoring text scenarios, facilitating teachers to carry out teaching activities, and helping students quickly understand the semantics of textbook texts, further verifying the application value of the proposed method in educational scenarios.

### 3.6 Efficiency optimization experiment

This experiment aims to verify the generation efficiency of the proposed method, analyze the impact of the number of self-reflection iterations on generation time and indicator performance, and verify the optimization effect of the dynamic early exit mechanism. The experimental task was to select 100 typical samples and set the number of self-reflection iterations to 1, 2, 3, 4, and 5, respectively. The generation time and core indicators (TFI, SRA, FID) under different iteration counts were compared. Simultaneously, a dynamic early exit mechanism was introduced (when the semantic alignment score $Score_{sem} \geq 0.85$ and the educational adaptability indicators met the standard, the iteration was terminated early). The generation time and performance changes before and after optimization were compared.

The results of the efficiency optimization experiment are shown in Table 5. The relationship between iteration count, generation efficiency, and performance is clear, and the optimization effect of the dynamic early exit mechanism is significant.

**Table 5.** Results of the efficiency optimization experiment

| Iteration Count | Text Fidelity Index (TFI) | Spatial Relationship Accuracy (SRA) (%) | Fréchet Inception Distance (FID) | Generation Time (s) | Generation Time After Dynamic Early Exit Optimization (s) | Time Optimization Rate (%) |
|---|---|---|---|---|---|---|
| 1 | 0.97 ± 0.01 | 83.5 ± 1.4 | 21.3 ± 1.0 | 9.8 ± 0.4 | 9.8 ± 0.4 | 0 |
| 2 | 0.98 ± 0.01 | 86.7 ± 1.3 | 19.8 ± 0.9 | 11.2 ± 0.4 | 10.5 ± 0.4 | 6.3 |
| 3 | 0.99 ± 0.01 | 89.3 ± 1.2 | 18.6 ± 0.9 | 12.3 ± 0.5 | 10.8 ± 0.4 | 12.2 |
| 4 | 0.99 ± 0.01 | 89.5 ± 1.2 | 18.5 ± 0.9 | 13.5 ± 0.5 | 11.0 ± 0.4 | 18.5 |
| 5 | 0.99 ± 0.01 | 89.6 ± 1.2 | 18.4 ± 0.9 | 14.7 ± 0.6 | 11.2 ± 0.5 | 23.8 |

As shown in Table 5, with the increase in the number of self-reflection iterations, the core performance indicators of the proposed method gradually stabilized, while the generation time increased linearly. When the iteration count increased from 1 to 3, TFI increased from 0.97 to 0.99, SRA increased from 83.5% to 89.3%, and FID decreased from 21.3 to 18.6, showing significant performance improvement; when the iteration count exceeded 3, TFI remained at 0.99, SRA increased to above 89.5%, and FID basically stabilized between 18.4 and 18.5, with a performance improvement amplitude of less than 1%. However, the generation time increased from 12.3s to 14.7s, an increase of 19.5%, indicating that 3 iterations can already meet performance demands, and excessive iterations would cause a waste of computing resources.

The dynamic early exit mechanism can effectively optimize generation efficiency, achieving time reduction under different iteration counts without affecting performance. When the iteration count was 5, the dynamic early exit mechanism could shorten the generation time from 14.7s to 11.2s, with a time optimization rate of 23.8%; when the iteration count was 3, the time optimization rate reached 12.2%, shortening the generation time to 10.8s while maintaining the optimal performance of TFI=0.99 and SRA=89.3%. This result shows that the dynamic early exit mechanism can accurately judge the iteration convergence

status. Under the premise of ensuring that performance meets the standard, it effectively reduces invalid iterations, solves the generation delay problem caused by multiple rounds of self-reflection iteration, and makes the generation efficiency of the proposed method meet the demands of practical teaching applications.

## 4. DISCUSSION AND LIMITATIONS

The generative English textbook material visualization method proposed in this paper has formed significant advantages in terms of technological innovation in image processing and adaptation to educational scenarios, possessing important technical value and application prospects. At the technical level, aiming at the core pain points of existing text-to-image generation technologies, this method constructs a dependency-aware discrete diffusion scene graph generation mechanism, realizing precise modeling of entity spatial relationships and compensating for the defect of insufficient spatial constraints in traditional methods; the code anchoring mechanism, through the deep integration of structured code generation and conditional constraints, builds an absolute fidelity barrier for text labels, solving the industry challenge of inaccurate text generation in diffusion models; the three-way conditional guided diffusion generation strategy realizes the collaborative constraints of multi-source information including text embedding, scene graph, and skeleton map, balancing image accuracy and visual aesthetics, enriching the innovative application of text-to-image generation technology in structured constraint scenarios. At the application level, this method can accurately adapt to the visualization demands of English textbook materials. The generated images not only meet the rigid requirements of text fidelity and spatial accuracy but also possess good educational adaptability. They can be directly applied to immersive English teaching to improve teaching efficiency and learning experience. Meanwhile, its core technical framework can be migrated to the visualization of textbook materials in other subjects such as Chinese and Science, possessing broad application potential. At the interdisciplinary level, based on the cross-integration of image processing and education, this paper constructs a complete research paradigm of "technological innovation-empirical verification-application landing," providing referenceable technical ideas and experimental design references for similar interdisciplinary research, and promoting the standardized application of generative AI technology in the field of educational visualization.

Although the proposed method demonstrates significant advantages in multiple experiments, there are still some limitations that cannot be ignored, reflecting the objectivity and rigor of the research. The scene graph parser is currently only fine-tuned on junior high school and high school English textbook corpora, and its generalization ability is limited by the domain corpus. When migrating to textbook materials in other subjects such as Science and History, due to differences in subject text semantic features and entity relationships, the scene graph generation accuracy may decrease, requiring further collection of domain-specific labeled data for optimization. In terms of generation efficiency, multi-round self-reflection iterations and scene graph Monte Carlo sampling inevitably increase end-to-end generation time. Although the dynamic early exit mechanism has achieved a certain degree of optimization, the generation speed is still difficult to meet real-time visualization demands and has limitations in large-scale textbook material batch generation scenarios. Educational adaptability optimization is still at a basic level, currently only focusing on general indicators such as cognitive load and attention guidance, without targeted design combined with specific teaching objectives such as English vocabulary memorization and grammar understanding, making it difficult to fully exert the supporting role of visualization technology in specific teaching links. In addition, the experimental dataset only covers English textbook materials from junior high school and high school sections, with limited sample size and coverage, making it difficult to fully verify the generalization performance of the method in different scenarios, which may affect the universality of the research conclusions.

Aiming at the above limitations and combining the research orientation in the field of image processing with the actual demands of educational scenarios, this paper proposes the following specific and feasible future research directions. To improve the cross-domain generalization ability of the scene graph parser, domain adaptive learning methods will be introduced. Through transfer learning and feature alignment of cross-disciplinary corpora, reliance on specific domain labeled data will be reduced to achieve accurate scene graph generation for textbook materials in different disciplines. To further improve generation efficiency, the deep integration of lightweight diffusion models and the self-reflection closed-loop will be explored, designing a more efficient iteration termination judgment mechanism, and combining model quantization and pruning technologies to reduce computational overhead, meeting real-time generation and batch generation demands. To deepen educational adaptability optimization, combining specific objectives of English teaching, a more refined educational adaptability indicator system will be designed to achieve deep matching between visualized images and specific teaching links such as vocabulary teaching and grammar teaching, enhancing the teaching value of the technology. Meanwhile, the coverage of the dataset will be expanded to include various material types such as primary school English, English picture books, and listening materials, further verifying the generalization of the method, and expanding application scenarios to promote the deep landing and innovative development of generative AI technology in the field of educational visualization.

## 5. CONCLUSION

Aiming at the core pain points of uncontrollable spatial relationships, low text fidelity, and insufficient educational adaptability in the visualization of English textbook materials, this paper proposes a generative AI visualization method integrating scene graph guidance, code anchoring, and multi-agent self-reflection. This paper elaborates on the technical details of core innovation modules such as dependency-aware discrete diffusion scene graph generation, three-way conditional guided diffusion generation, code anchoring fidelity, and multi-agent self-reflection closed-loop. By constructing reasonable mathematical models and constraint mechanisms, triple precise control over text semantics, spatial relationships, and educational adaptability is realized, providing a complete technical solution for the efficient visualization of English textbook materials.

Multiple sets of empirical experimental results show that

this method significantly outperforms existing mainstream text-to-image generation methods in core image processing indicators such as text fidelity and spatial relationship accuracy. The text fidelity is close to 1.0, and the spatial relationship accuracy rate is improved by more than 17% compared to the existing optimal method. Meanwhile, it possesses good educational adaptability and generation efficiency, and can effectively meet the rigid demands of English teaching scenarios. This research not only enriches the innovative application of generative AI in structured constraint scenarios, providing new ideas for the optimization of text-to-image generation technology, but also offers a referenceable technical paradigm and empirical reference for the cross-integration of image processing and education. At the same time, limitations such as insufficient generalization of the scene graph parser and the need for further improvement in generation efficiency objectively exist. Future work will continue to refine the method through domain adaptive learning and lightweight model optimization to promote the innovative development and deep landing of educational material visualization technology.

## REFERENCES


[1] Zhou, L., Deng, X., Ning, Z., Zhao, H., Wei, J., Leung, V.C.M. (2025). When generative AI meets semantic communication: Optimizing radio map construction and distribution in future mobile networks. IEEE Network, 39(3): 47-55. https://doi.org/10.1109/mnet.2025.3529513

[2] Heitmann, M., Jansen, T.P., Reisenbichler, M., Schweidel, D.A. (2025). Picture perfect: Engaging customers with visual generative AI. Journal of Marketing. https://doi.org/10.1177/00222429251356993

[3] Dai, C., Ke, F., Dai, Z., Pachman, M. (2022). Improving teaching practices via virtual reality-supported simulation-based learning: Scenario design and the duration of implementation. British Journal of Educational Technology, 54(4): 836-856. https://doi.org/10.1111/bjet.13296

[4] Peltonen, L., Hu, G. (2024). Linguacultural competence in business English communication: The case of a business English textbook in China. Language, Culture and Curriculum, 38(1): 19-37. https://doi.org/10.1080/07908318.2024.2372594

[5] Petersson, J., Sayers, J., Rosenqvist, E., Andrews, P. (2022). Analysing English year-one mathematics textbooks through the lens of foundational number sense: A cautionary tale for importers of overseas-authored materials. Oxford Review of Education, 49(2): 262-280. https://doi.org/10.1080/03054985.2022.2064443

[6] Gendy, G., He, G., Sabor, N. (2024). Diffusion models for image super-resolution: State-of-the-art and future directions. Neurocomputing, 617: 128911. https://doi.org/10.1016/j.neucom.2024.128911

[7] Liu, B., Shao, S., Li, B., Bai, L., Xu, Z., Xiong, H., Kwok, J.T., Helal, S., Xie, Z. (2026). Alignment of diffusion models: Fundamentals, challenges, and future. ACM Computing Surveys, 58(9): 1-37. https://doi.org/10.1145/3796982

[8] Levin, O., Frei-Landau, R., Flavian, H., Miller, E.C. (2023). Creating authenticity in simulation-based learning scenarios in teacher education. European Journal of Teacher Education, 48(2): 291-312. https://doi.org/10.1080/02619768.2023.2175664

[9] Koh, E., Zhang, L., Lee, A.V.Y., Wang, H. (2024). Revolutionizing word clouds for teaching and learning with generative artificial intelligence: Cases from China and Singapore. IEEE Transactions on Learning Technologies, 17: 1390-1401. https://doi.org/10.1109/tlt.2024.3385009

[10] Gonsalves, R.A. (1981). Digital image-processing in education. In Proc. SPIE 0301, Design of Digital Image Processing Systems, pp. 22-28. https://doi.org/10.1117/12.932599

[11] Ma, L. (2019). Research on distance education image correction based on digital image processing technology. EURASIP Journal on Image and Video Processing, 2019(1): 18. https://doi.org/10.1186/s13640-019-0416-9

[12] Wei, Z. (2023). Metaverse-based online English teaching scheme in multi-source and cross-domain environment. Fractals, 31(6): 2340153. https://doi.org/10.1142/s0218348x23401539

[13] Ma, L. (2021). An immersive context teaching method for college English based on artificial intelligence and machine learning in virtual reality technology. Mobile Information Systems, 2021: 1-7. https://doi.org/10.1155/2021/2637439

[14] Küçükaydin, M.A. (2025). A social semiotic analysis of visual representation in science textbooks for Maarif Model for Turkey's century. Women's Studies International Forum, 114: 103229. https://doi.org/10.1016/j.wsif.2025.103229

[15] Liu, Q., Dong, Y., Pei, Y., Zheng, L., Zhang, L. (2023). Long and short-range relevance context network for semantic segmentation. Complex & Intelligent Systems, 9(6): 7155-7170. https://doi.org/10.1007/s40747-023-01103-6

[16] Ren, J., Zhang, Q., Kang, B., Zhong, Y., He, M., Ge, Y., Bi, H. (2025). Semantic-spatial guided context propagation network for camouflaged object detection. Applied Intelligence, 55(5): 1-15. https://doi.org/10.1007/s10489-025-06264-0

[17] Li, X.D. (2022). A hybrid online and offline approach to teaching spoken English based on modern educational technology. Mathematical Problems in Engineering, 2022: 3803436. https://doi.org/10.1155/2022/3803436

[18] Diamond, F. (2020). Cultural memory in English teaching: A critical autobiographical inquiry. English Teaching: Practice & Critique, 19(2): 231-244. https://doi.org/10.1108/etpc-05-2019-0061

[19] Gao, X.B., Wang, T., Li, J. (2005). A content-based image quality metric. In Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, pp. 231-240. https://doi.org/10.1007/11548706_25

[20] Fry, E.W.S., Triantaphillidou, S., Jenkin, R.B., Jacobson, R.E., Jarvis, J.R. (2019). Scene-and-process-dependent spatial image quality metrics. Journal of Imaging Science and Technology, 63(6): 060407. https://doi.org/10.2352/j.imagingsci.technol.2019.63.6.060407