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

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1341-1356
Journal homepage: http://iieta.org/journals/ts

Cross-Modal Image Enhancement and Visual-Semantic Fusion Empowered by Generative Artificial Intelligence for English Language Teaching

Yan Zhao, Yun Li, Lan Zhao*

School of Language and Cultural Communication, Baoding University of Science and Technology, Baoding 071000, China

Corresponding Author Email: lcc_zhaolan@bdlg.edu.cn

https://doi.org/10.18280/ts.430320

Received: 8 November 2025
Revised: 20 April 2026
Accepted: 3 May 2026
Available online: 30 June 2026

Keywords:
generative artificial intelligence, cross-modal fusion, diffusion models, image enhancement, semantic awareness, English language teaching

ABSTRACT

Generative image technologies offer a novel technical pathway for the intelligent construction of visual instructional materials. However, existing text-driven image generation methods inadequately accommodate the specialized semantic constraints inherent to English language teaching scenarios, with prevalent issues including tense expression distortion, misrepresentation of spatial prepositional relations, and inconsistencies between textual descriptions and visual content. Concurrently, conventional image enhancement strategies primarily focus on perceptual quality optimization, lacking attention-guidance mechanisms aligned with learners' cognitive processes, and fail to automatically rectify fine-grained semantic deviations in images, thereby falling short of the material requirements for standardized English language teaching applications. To address these challenges, a four-stage closed-loop cross-modal image processing framework was proposed, comprising grammatically decoupled generation, multi-granularity semantic fusion, cognitive saliency augmentation, and semantic contradiction self-correction. Within this framework, a tense-aware feature modulation mechanism was designed based on diffusion models to enable controllable generation of visual dynamic features corresponding to verb tenses. Syntactic graph encoding combined with attention bias constraints was employed to explicitly model textual spatial-semantic relationships, ensuring precise grammatical layout in generated images. Multi-scale semantic fusion of instructional images was realized through a dual-stream cross-modal Transformer architecture. A saliency-based adversarial augmentation model, integrated with pedagogical cognitive features, was constructed to adaptively optimize visual attention distribution. Furthermore, semantic contradiction heatmaps were generated through phrase-region alignment between textual and visual modalities, driving a diffusion-based inpainting model for fine-grained semantic correction of generated images. A dedicated English language teaching image dataset annotated with tenses and spatial relations was constructed for systematic experimentation. Experimental results demonstrated that the proposed method significantly improved both semantic accuracy and visual-cognitive adaptability of instructional images. The proposed approach outperformed current state-of-the-art algorithms in terms of scene graph preposition recognition accuracy and visual saliency matching degree, and effectively enhanced learners' grammatical acquisition efficiency. This research realizes the computable transformation of linguistic rules into controllable visual generation tasks, establishing a novel technical paradigm for domain-specific cross-modal image generation and enhancement in educational contexts.

1. INTRODUCTION

The rapid advancement of generative artificial intelligence technologies has substantially propelled the field of computer vision. Image generation algorithms, particularly those based on diffusion models, have fundamentally transformed the conventional visual content production paradigm—which previously relied heavily on manual illustration and material editing—by virtue of their superior detail modeling capabilities and cross-modal adaptability, thereby providing novel technical support for the construction of intelligent and customized educational visual resources [1, 2]. Within the modern English language teaching system, contextualized visual learning serves as a core pedagogical approach for assisting learners in comprehending grammatical rules, establishing spatial cognition, and consolidating practical language application skills. High-quality instructional images are capable of transforming abstract linguistic knowledge—such as tense variations and spatial prepositional relations—into intuitive visual scenarios, thereby effectively lowering the cognitive threshold of language acquisition [3, 4]. Distinct from general-purpose visual images, visual materials deployed in English language teaching contexts are characterized by strong domain specificity and stringent constraints. Such materials are required not only to present visually natural and realistic appearances, but also to accurately map English

grammatical rules, faithfully reproduce action tense features, spatial topological relationships among objects, and scene semantic tones, thereby meeting the demands of standardized and systematic language instruction [5, 6]. Current cross-modal generation and enhancement techniques in computer vision predominantly focus on visual quality optimization for general-purpose scenarios. An effective mapping mechanism between linguistic rules and visual generation tasks has yet to be established, rendering the precise transformation of discrete linguistic features into continuous visual controllable signals difficult to achieve. Furthermore, a quantitative evaluation framework for visual quality specifically tailored to instructional scenarios remains lacking [7, 8]. To address these issues, a specialized image processing system adapted to English language teaching scenarios is constructed in this work, integrating conditional image generation, cross-modal feature fusion, cognitive saliency modeling, and intelligent image inpainting techniques. This system not only addresses the industry pain point of customized production of educational visual resources, but also provides a novel methodological paradigm for semantically driven, precise cross-modal visual manipulation, thereby carrying substantial engineering application value and academic innovation significance [9, 10].

State-of-the-art text-driven image generation models have presently achieved high-resolution and high-fidelity visual image synthesis, and have been widely applied in cross-modal content creation for general-purpose scenarios. However, the textual semantic understanding of such models relies solely on statistical co-occurrence features derived from large-scale corpus training, without the capacity to deeply parse the logical rules and fine-grained semantic distinctions inherent to English grammar [11, 12]. Visual dynamic features corresponding to progressive and perfect verb tenses are inadequately differentiated by these models, and precise control over the topological spatial relationships of objects corresponding to various spatial prepositions is not reliably achieved. Consequently, issues such as tense expression distortion and spatial relation misalignment frequently emerge, rendering the grammatical compliance and pedagogical usability of generated images difficult to guarantee [13, 14]. In the post-processing stage, existing cross-modal image fusion techniques predominantly adopt pixel consistency and stylistic uniformity as optimization objectives, employing fixed-weight fusion strategies for multi-source image stitching and adaptation. A dynamic modulation mechanism aligned with the core semantic requirements of English language teaching has not been incorporated, and the fusion process is prone to disrupting the key object spatial relationships and scene semantic features defined by the textual input, thereby resulting in the loss of pedagogical information [15, 16]. Simultaneously, conventional image enhancement algorithms center on the improvement of perceptual visual quality for human observers, achieving enhanced image appearance through global optimization of contrast, brightness, and sharpness. The distribution patterns of learners' cognitive attention are entirely neglected, and task-oriented saliency optimization strategies are absent. This leads to insufficient visual discriminability of core instructional regions and redundant information in peripheral areas, failing to accommodate the cognitive principles of language acquisition [17, 18]. Furthermore, existing cross-modal image-text matching and image editing techniques have yet to form an automated closed-loop optimization system. Fine-grained semantic contradiction detection in images remains heavily dependent on manual screening, and the intelligent inpainting process lacks linguistic semantic constraints, such that corrected images remain susceptible to semantic biases [19, 20]. These technical deficiencies render the existing visual processing pipeline inadequate for the rigorous and standardized demands of English language teaching scenarios, necessitating the urgent construction of a comprehensive image processing chain that integrates grammatically precise generation, semantically lossless fusion, cognitively oriented enhancement, and intelligent self-correction [21-24].

To address the aforementioned technical deficiencies, four core innovations are formulated for the task of precise visual-semantic generation and enhancement in English language teaching scenarios. A grammar-decoupled conditional diffusion generation network is constructed, into which linguistic prior knowledge is explicitly incorporated during the diffusion generation process through tense-aware encoding and syntactic graph-based attention bias mechanisms, thereby enabling accurate image generation under grammatical rule constraints. An integrated module for multi-granularity cross-modal fusion and cognitive saliency augmentation is established, where semantic-preserving fusion between textbook images and generated images is achieved via a dual-stream Transformer architecture, and adaptive visual attention optimization is accomplished in conjunction with pedagogical cognitive features. An automatic semantic contradiction detection and localized repainting correction mechanism is proposed, in which contradiction heatmaps are generated based on fine-grained image-text feature alignment, and automated correction of image defects is realized through semantically constrained diffusion-based inpainting, forming a complete self-optimizing closed loop. A multidimensional quantitative evaluation framework tailored to English language teaching images is constructed, together with a comprehensively annotated dedicated dataset, and systematic validation is conducted across dimensions, including grammatical accuracy, visual quality, and pedagogical effectiveness, thereby providing a standardized evaluation benchmark for technical research in this domain.

The overall organizational structure of this work is as follows: In Chapter 2, the core principles and technical details of the proposed four-stage closed-loop image processing framework are elaborated, with comprehensive descriptions of the network architectures, computational logic, and optimization functions of each innovative module. In Chapter 3, extensive experimental validation is conducted, with detailed descriptions of the dataset and experimental parameter configurations. The superiority of the proposed method and the effectiveness of each module are verified through comparative experiments, ablation studies, and user evaluation experiments. In Chapter 4, the experimental results are analyzed and discussed, with interpretations of the model working mechanisms, application advantages, and existing limitations. In Chapter 5, the research contributions are summarized, and future research directions are outlined.

2. METHODOLOGY

In response to the inadequacies of existing cross-modal generation and image enhancement techniques in accommodating the specialized semantic constraints and cognitive learning requirements of English language teaching

scenarios, an end-to-end closed-loop cross-modal image enhancement and visual-semantic fusion framework is constructed. This framework is composed of four deeply coupled and progressively cascaded core modules: grammar-decoupled diffusion generation, multi-granularity visual-semantic fusion, pedagogical saliency-based adversarial augmentation, and semantic contradiction detection and repainting. Standardized English language teaching textual inputs are employed as the primary foundation, with optional reference textbook images incorporated for collaborative computation. Through a structured visual processing pipeline, discrete linguistic grammatical rules are transformed into continuously controllable visual generation features, fundamentally resolving the persistent issues of tense expression ambiguity and inaccurate spatial semantic modeling inherent to general-purpose generation models. The framework sequentially accomplishes grammatically precise initial instructional image generation, multi-scale feature fusion with textbook semantic consistency, cognitively oriented adaptive visual enhancement, and automated correction and restoration of fine-grained semantic biases. Each functional module is interconnected in a sequential manner and jointly optimized. Concurrently, through a training strategy combining stage-wise pre-training and global joint fine-tuning, efficient propagation of semantic features and gradient information throughout the complete processing pipeline is ensured. Ultimately, high-quality images characterized by grammatical accuracy, visual adaptability, and pedagogical practicality are generated, providing a systematic and generalizable technical solution for semantically driven cross-modal visual generation and enhancement tasks in educational contexts.

2.1 Grammar-decoupled conditional diffusion generation module

General-purpose latent diffusion models, which learn modal association features from large-scale image-text data, are capable of achieving high-quality visual image synthesis. However, these models are limited to learning statistical association patterns between textual and visual content, without the capacity to accurately model the tense logic and spatial topological relationships among objects that are inherent to English grammar. Consequently, semantic deficiencies such as action state distortion and spatial relation misalignment frequently emerge in the generated instructional images. To address this issue, a grammar-decoupled conditional generation mechanism is constructed based on a pre-trained diffusion model, through which linguistic rules are transformed into differentiable visual modulation signals. By means of a dual-branch architecture comprising tense feature dynamic modulation and syntactic structure attention constraints, grammatical prior knowledge is explicitly embedded during the diffusion denoising process, thereby overcoming the limitations of statistical semantic matching and achieving accurate instructional image generation under grammatical rule constraints. The network architecture of the grammar-decoupled conditional diffusion generation module is illustrated in Figure 1.

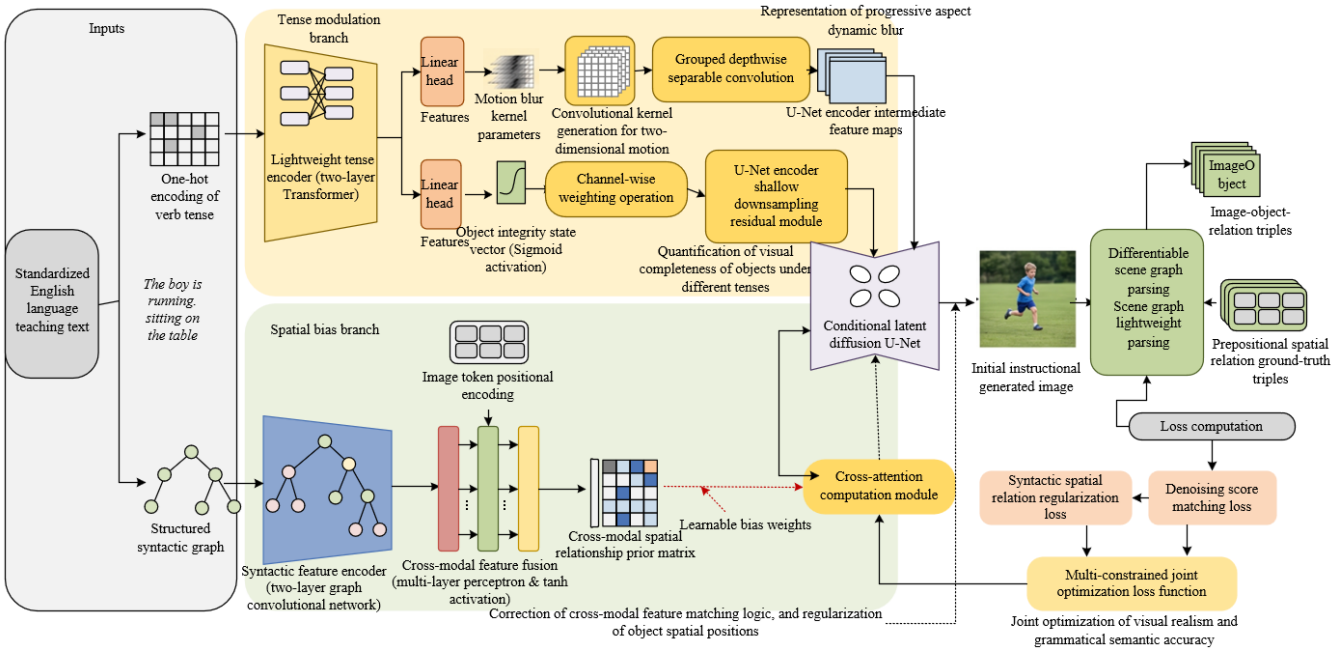

Figure 1. Network architecture of the grammar-decoupled conditional diffusion generation module

A lightweight tense encoder is designed to accomplish the visual mapping of linguistic tense features. The encoder is primarily composed of two stacked Transformer encoding units, with the one-hot encoding vector of verb tense taken as input. Through feature encoding, the visual representation patterns corresponding to different tenses are extracted, and two independent linear output heads are configured to achieve feature-decoupled modeling. From the first linear head, a motion blur kernel parameter vector is output, based on which a dynamically constructed two-dimensional motion convolution kernel of configurable size is built. Simultaneously, non-negativity and normalization constraints are imposed on the convolution kernel parameters to ensure the physical plausibility of the motion blur features and to accurately characterize the motion trajectory features of dynamic actions. From the second linear head, in conjunction with a Sigmoid activation function, an object integrity state vector along the channel dimension is output. The

dimensionality of this vector is maintained consistently with that of the intermediate feature channels of the diffusion U-Net, serving to quantify the visual completeness features of objects under different tenses. During the progressive denoising process of the diffusion model, tense feature injection is performed after the shallow downsampling residual module of the U-Net encoder. First, the object completeness information of the feature maps is modulated through a channel-wise weighting operation, defined as:

$$h'=h\odot\omega \tag{1}$$

where, h denotes the original network feature map, ω denotes the object integrity state vector, and h' denotes the tense-modulated feature map. On this basis, grouped depthwise separable convolution is applied to the feature maps using the dynamically generated motion blur kernel, imposing a unified motion blur constraint on all channel features. Tense information is thereby encoded into the low-level texture and edge features of the image, achieving differentiated visual representations characterized by dynamic blur for the progressive aspect and static completeness for the perfective aspect.

To accurately control the object spatial relationships corresponding to prepositions in the text, a structured syntactic graph is constructed from the input teaching text. Syntactic feature encoding is accomplished via a two-layer graph convolutional network, through which the dependency constraint relationships among words are extracted, yielding high-dimensional syntactic feature vectors for each word node. In conjunction with image token positional encoding and textual syntactic features, a cross-modal spatial relationship prior matrix is constructed to quantify the association strength between image visual regions and textual semantic units. The matrix is computed as follows:

$$M_{syn}(p,q)=MLP([e_p^{pos};h_q^{syn}]) \tag{2}$$

where, e_p^{pos} denotes the positional encoding of the p-th image visual token, $h^{syn}{}_q$ denotes the syntactic feature of the q-th text token, and a multilayer perceptron is employed to fuse cross-modal features and output an association score, which is then normalized via a tanh activation function to yield a normalized bias weight. This prior matrix is incorporated into the cross-attention computation of the diffusion model as learnable weights, through which the cross-modal feature matching logic is corrected. The optimized formulation can be expressed as:

$$Attention(Q,K,V)=softmax\left(\frac{QK^{\top}}{\sqrt{d}}+\lambda M_{syn}\right)V \tag{3}$$

where, Q, K, and V denote the query, key, and value vectors of the attention mechanism, respectively; d denotes the feature dimensionality; and λ denotes a learnable bias intensity coefficient. Through this mechanism, cross-regional associations constrained by syntactic rules are adaptively reinforced during the image layout generation stage, and object spatial positional relationships are regulated at the feature matching level, thereby effectively circumventing the spatial semantic confusion issues common to general-purpose models.

To simultaneously ensure the visual quality and grammatical semantic accuracy of generated images, a multi-constrained joint optimization loss function is constructed in this module. The standard denoising score matching loss of the diffusion model is adopted as the foundational optimization objective to guarantee the visual realism and textural fidelity of image generation. Concurrently, a differentiable scene graph parser is introduced to perform structured semantic parsing of generated images, from which image object relation triples are extracted. These triples are then matched against the prepositional spatial relation ground-truth triples corresponding to the teaching text, and a syntactic spatial relation regularization loss is formulated to quantify the cross-modal semantic matching deviation. The overall training loss function of the module is defined as:

$$L_{gen}=L_{denoise}+\beta L_{syn} \tag{4}$$

where, $L_{denoise}$ denotes the denoising score matching loss, L_{syn} denotes the syntactic spatial relation regularization loss, and β denotes a loss balancing weight that coordinates the optimization priorities between visual generation quality and grammatical semantic accuracy. Through iterative optimization of the joint loss, the model is enabled to adhere rigorously to English grammatical rules during scene generation while preserving visual perceptual quality, ultimately producing a semantically accurate initial instructional generated image I_{gen}.

2.2 Multi-granularity visual-semantic fusion module

The initial instructional images generated through grammar-decoupled diffusion generation possess precise grammatical semantic features. However, significant discrepancies exist in visual style and scene element layout relative to standardized textbook images, and direct application to teaching scenarios is prone to issues such as visual system fragmentation and inadequate material adaptability. Conventional image fusion methods predominantly perform pixel-level superposition and style unification based on fixed weights, without incorporating task-specific semantic features to dynamically regulate the fusion process, thereby readily disrupting the established object spatial topological relationships and grammatical semantic information within the images. To address this issue, a multi-granularity cross-modal visual-semantic fusion architecture is constructed in this module. Through hierarchical feature modeling and a text-driven dynamic gating mechanism, adaptive fusion between generated images and reference textbook images is achieved, whereby the core pedagogical semantics are fully preserved while the visual presentation is unified, enabling co-optimization of visual adaptability and semantic fidelity. A schematic diagram of the multi-granularity visual-semantic fusion and alignment mechanism is illustrated in Figure 2.

Multi-scale feature extraction of the dual input images is accomplished in this module based on a contrastive language-image pre-training visual encoder, through which visual representation information at different levels is fully exploited. By extracting output features from multiple layers of the encoder, a hierarchical feature set covering local texture details to global scene semantics is constructed, corresponding respectively to the multi-granularity feature representations of the generated image and the reference textbook image. Features at all levels are uniformly fed into a parameter-shared cross-modal Transformer fusion unit, ensuring consistency

and stability in the fusion logic across different scales. To achieve text-semantic-guided dynamic fusion, a feature-wise linear modulation gating modulation structure is introduced, in which the global semantic embedding of the teaching text is adopted as the modulation reference to generate a spatial-dimensional hybrid mask and channel-wise affine transformation parameters, respectively. First, adaptive weighted fusion of the dual image features is accomplished through the spatial mask, with the specific computation defined as:

$$\tilde{F}^{l}_{fused}=\alpha^{l}\odot F^{l}_{gen}+(1-\alpha^{l})\odot F^{l}_{ref} \tag{5}$$

where, α^{l} denotes the spatial hybrid mask corresponding to the l-th layer feature, with values in the range [0, 1], used to adaptively allocate the fusion weights between the generated features and the reference textbook features; F^{l}_{gen} and F^{l}_{ref} denote the l-th layer features of the generated image and the textbook image, respectively. On this basis, semantic rectification and feature distribution optimization of the fused features are accomplished through adaptive instance normalization, with the computation process defined as:

$$F^{l}_{fused}=\gamma^{l}\odot\frac{\tilde{F}^{l}_{fused}-\mu}{\sigma}+\beta^{l} \tag{6}$$

where, μ and σ denote the channel-wise mean and standard deviation of the fused features, respectively, and γ^{l} and β^{l} denote the channel-wise scaling and bias parameters, used to adjust the semantic tendency and feature distribution of the fused features. The multi-scale optimized fused features are upsampled and reconstructed through a lightweight feature pyramid decoder, yielding a preliminary fused image.

To fundamentally prevent the fusion process from disrupting the key spatial semantic relationships essential to English language teaching, a differentiable scene graph alignment loss is introduced in this module to enforce high-level semantic constraints. Through a unified scene graph parser, the structured semantic graph of the teaching text and the visual scene graph of the fused image are respectively extracted. The sets of spatial prepositional relation edges in both graphs are filtered, and bidirectional optimal node matching is accomplished via the Hungarian algorithm, establishing precise correspondence between textual semantic relations and visual entity relations. Weighted constraints are imposed on the core spatial topological relationships in English language teaching to reinforce the model's capacity for preserving key semantics. The loss function is defined as follows:

$$L_{sg}=\sum_{(r_t,r_f)\in Match} w_{r_t}\cdot CE(r_f,r_t) \tag{7}$$

where, r_t and r_f denote the textual semantic relation and the image visual relation, respectively; w_{r_t} denotes the semantic relation weight coefficient, with higher weight constraints assigned to prototypical spatial prepositional relations; and CE denotes the cross-entropy loss function, used to quantify the classification deviation of matched semantic pairs. This loss is capable of approximating the scene graph edit distance in a differentiable form, precisely supervising semantic integrity throughout the fusion process.

A multi-constrained joint optimization strategy is adopted for the overall module to perform iterative parameter updates. Pixel-level L1 reconstruction loss and high-level perceptual loss are combined to ensure detail fidelity and visual naturalness of the fused image. An adversarial loss is employed to optimize overall style consistency, complemented by the scene graph alignment loss to constrain semantic fidelity. These multi-dimensional loss functions cooperate synergistically with complementary constraints, enabling the model to rigorously preserve the tense features and spatial semantic relationships established during the grammar-decoupled generation stage while adapting to the visual style of textbook images. Ultimately, a fused image I_{fused} characterized by both visual adaptability and pedagogical semantic accuracy is produced.

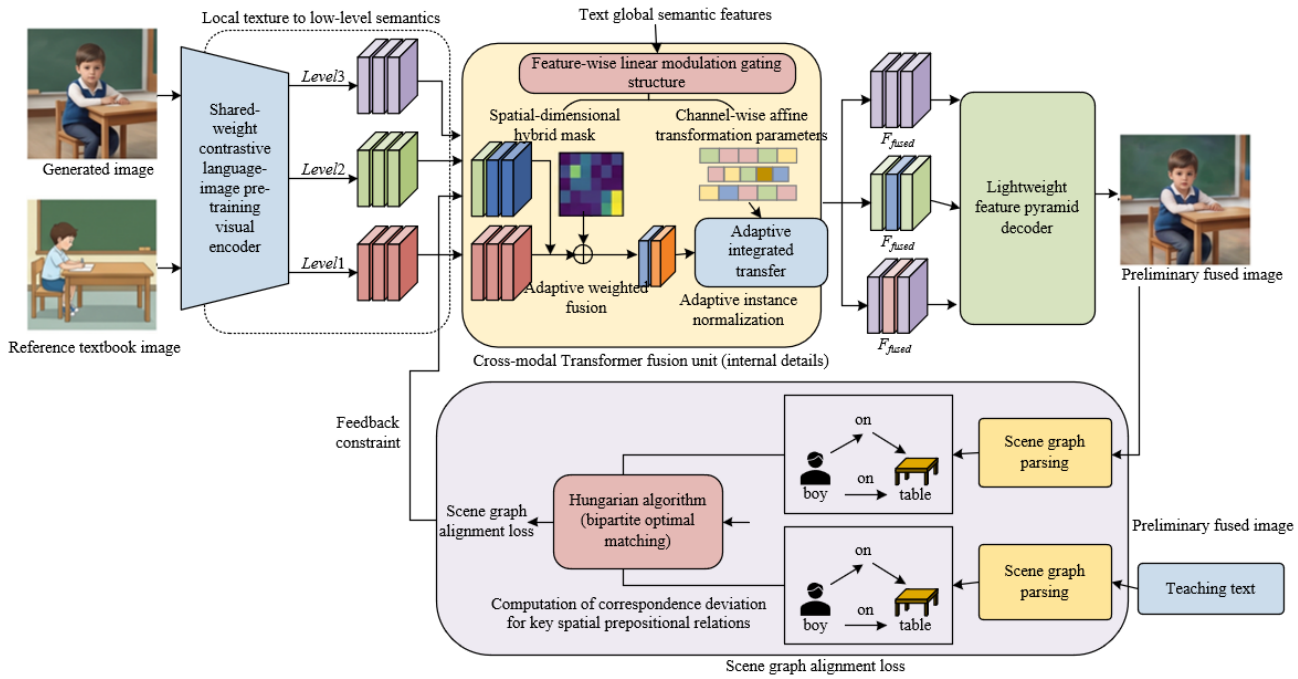

Figure 2. Schematic diagram of the multi-granularity visual-semantic fusion and alignment mechanism

2.3 Pedagogical saliency-driven adversarial augmentation module

Following multi-granularity semantic fusion, the processed images fully conform to English grammatical rules and textbook visual styles, with overall semantic structure and scene layout exhibiting pedagogical usability. However, the visual energy distribution in conventional fusion outputs tends toward uniformity, failing to align with the cognitive attention patterns of language learning. Conventional image enhancement methods predominantly employ globally uniform contrast and brightness adjustment strategies, with optimization objectives confined to general-purpose visual quality improvement, without the capacity to perform differentiated visual modulation according to the importance levels of teaching knowledge points. This results in insufficient visual saliency for key learning objects such as preposition-associated entities and core action regions, while secondary scene elements are prone to creating visual interference. To address these issues, learner cognitive visual patterns are integrated with an adversarial augmentation mechanism in this module, and a text-conditioned saliency-guided adaptive enhancement network is constructed. Pixel-level non-uniform visual optimization is driven by pedagogical tasks, achieving deep adaptation between image visual presentation and the attention mechanisms of language learning. The network architecture of the pedagogical saliency-driven adversarial augmentation module is illustrated in Figure 3.

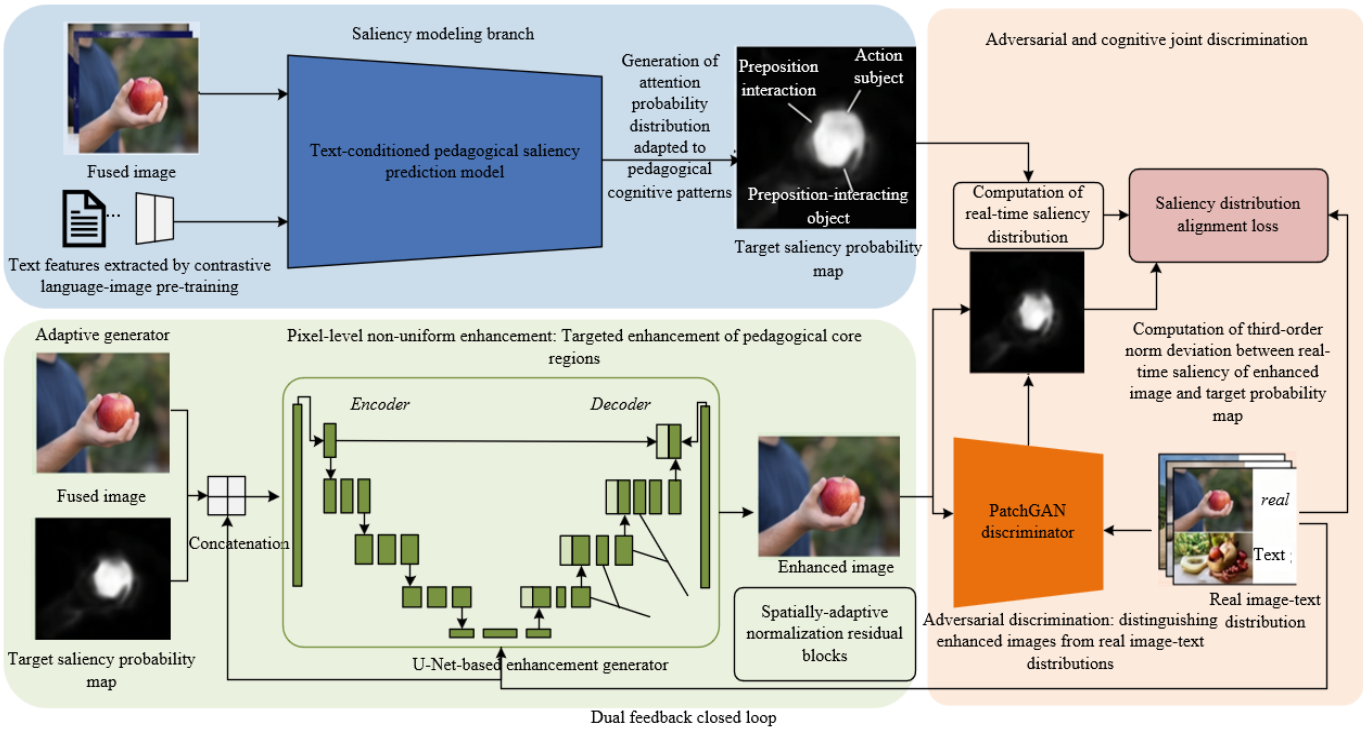

Figure 3. Network architecture of the pedagogical saliency-driven adversarial augmentation module

A text-conditioned pedagogical saliency prediction model is constructed in this study to achieve domain-specific visual attention distribution modeling for teaching scenarios. The model is built upon a cross-modally adapted SalGAN architecture, in which the conventional approach of relying solely on image pixel features for saliency detection is abandoned, and textual semantic priors are introduced to guide saliency region prediction. Global semantic features of the teaching text are extracted via a contrastive language-image pre-training text encoder, and the textual semantic information is progressively embedded into the multi-scale visual features of the saliency encoder through feature adaptive modulation layers, establishing a mapping relationship between textual grammatical semantics and visual attention regions. The model is pre-trained on a dedicated dataset comprising real learner eye-tracking data and expert pedagogical annotations, and is capable of outputting a standardized target attention probability distribution based on the input text and fused image. The computation can be expressed as:

$$P_{target}=S(I_{fused}, T) \tag{8}$$

where, S denotes the pedagogical saliency prediction network, I_{fused} denotes the input fused image, T denotes the teaching text, and P_{target} denotes the pixel-wise saliency probability map. This map is capable of precisely characterizing the effective attention regions in English learning scenarios, assigning high probability weights to core pedagogical targets such as preposition-interacting objects and action subjects, thereby providing precise cognitive supervisory signals for subsequent targeted enhancement.

A saliency-conditional adversarial augmentation network is constructed to achieve co-optimization of visual quality enhancement and cognitive attention alignment. The enhancement generator adopts a lightweight U-Net architecture integrated with spatially-adaptive normalization residual blocks. Channel-wise concatenation of the fused image and the target saliency probability map is performed as the joint input for feature decoding and image reconstruction. Through the spatially-adaptive normalization mechanism, differentiated enhancement requirements across saliency regions are accommodated, and a cognitively optimized enhanced image is produced. The discriminator employs a

PatchGAN architecture, using paired features of images and saliency maps as discriminative criteria to distinguish, at the local patch level, between natural real image-text matching distributions and the visually enhanced distributions produced by the model, thereby constraining the local visual realism of the enhancement results. The overall adversarial optimization objective function is defined as:

$$\min_{G_{enh}} \max_{D} \begin{matrix} E\left[\log D\left(I_{fused}, P_{target}\right)\right] + \\ E\left[log\left(1\text{-}D(I_{enh}, P_{target})\right)\right] + \eta L_{sal} \end{matrix} \quad (9)$$

where, G_{enh} and D denote the enhancement generator and the discriminator, respectively; E denotes the mathematical expectation operator; and η denotes a weight balancing coefficient used to adjust the optimization priorities between the adversarial loss and the saliency regularization loss. The first two terms in the formulation constitute the classical adversarial training loss, responsible for constraining the visual naturalness and stylistic consistency of the enhanced images.

To further strengthen the optimization constraint at the cognitive level, a saliency distribution alignment loss is introduced in this study to quantify the deviation between the enhanced image and the standard pedagogical attention distribution. The specific formulation is expressed as:

$$L_{sal} = \left\| S(I_{enh}, T) - P_{target} \right\|_2^2 \quad (10)$$

where, $S(I_{enh},T)$ denotes the real-time saliency distribution corresponding to the enhanced image. The overall deviation between the predicted distribution and the target distribution is minimized via the squared L2 norm. This loss is capable of precisely guiding the enhancement strategy at the pixel level, enabling the model to adaptively improve the clarity, contrast, and detail expressiveness of pedagogical core regions, while simultaneously attenuating the visual weight of redundant background information. Under the dual regulation of adversarial and cognitive constraints, the network is enabled to transcend the limitations of globally homogeneous enhancement, producing visual content that conforms to human cognitive habits in language learning, and ultimately yielding a high-pedagogical-validity enhanced image I_{enh}.

2.4 Semantic contradiction detection and localized repainting correction module

The aforementioned generation, fusion, and augmentation modules effectively ensure the macro-level grammatical structure, visual style adaptability, and cognitive attention rationality of instructional images. However, semantic deviations at the pixel level and in fine-grained entity attributes remain difficult to avoid. Stochastic perturbations in the latent space during image generation are prone to causing inconsistencies between detailed attributes—such as color, quantity, and material—and the textual instructional descriptions. Such subtle semantic contradictions cannot be corrected through macro-level semantic constraints or global visual optimization. Moreover, conventional cross-modal editing methods rely on manual localization of defective regions, lacking an automated detection and correction pipeline. To address this issue, a fine-grained image-text alignment-driven semantic error correction mechanism is constructed in this module. Through fully automatic contradiction region detection, dynamic mask construction, and semantically constrained diffusion-based repainting strategies, precise correction of detailed semantic deviations in instructional images is achieved, thereby completing the closed-loop optimization capability of the overall visual processing pipeline. A flowchart of the semantic contradiction detection and fine-grained diffusion-based repainting process is illustrated in Figure 4.

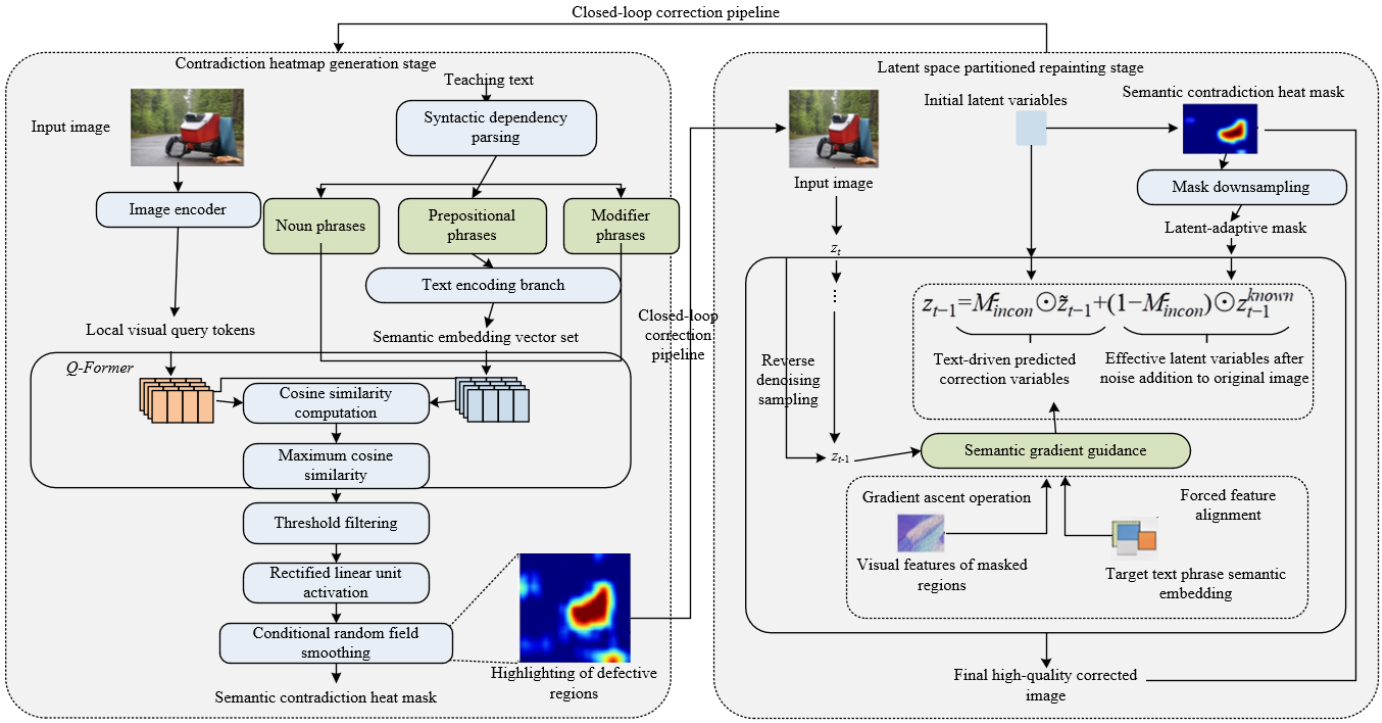

Figure 4. Flowchart of semantic contradiction detection and fine-grained diffusion-based repainting

Fine-grained matching between local image regions and textual semantic phrases is accomplished in this module based on the Q-Former architecture, enabling pixel-level localization of semantic contradiction regions. Multiple sets of learnable visual query tokens are output by the image encoder, with each set independently corresponding to a local semantic region of the image, forming a global visual-semantic representation sequence. Simultaneously, syntactic dependency parsing is performed on the input teaching text, from which basic semantic units such as noun phrases, prepositional phrases, and attribute modifier phrases are extracted, and independent semantic embeddings for each phrase are generated through the text encoding branch. To quantify the degree of local image-text matching, one-to-one alignment between visual tokens and text phrase features is performed, cosine similarity is computed, and the maximum value is taken as the single-region alignment score, thereby characterizing the degree of correspondence between local visual content and textual semantics. A semantic contradiction heat mask is constructed based on the pixel-level alignment scores, with the computation defined as:

$$M_{incon}(x,y)=ReLU(\tau-S_{align}(x,y)) \tag{11}$$

where, $S_{align}(x,y)$ denotes the full-resolution image-text alignment score map, and τ denotes a preset semantic matching threshold used to distinguish between well-matched regions and semantically mismatched regions. The rectified linear unit activation function is capable of filtering out positive matching deviations, retaining only pixel responses where semantic contradictions exist. To eliminate single-pixel noise interference and ensure the spatial continuity of defective regions, a conditional random field is introduced in this study to perform smooth optimization of the initial heat mask, precisely localizing defective regions such as entity attribute errors and fine-grained semantic mismatches, thereby providing precise mask constraints for subsequent localized repainting.

Based on the generated semantic contradiction mask, a latent-space diffusion-based inpainting model is adopted in this study to accomplish local image correction, whereby semantic repainting of defective regions is achieved while preserving the valid semantic content and visual structure of the original image. The pixel-space contradiction mask is downsampled to the latent-space resolution of the diffusion model, yielding a latent-adaptive mask. During each iteration of the reverse diffusion sampling process, partitioned updates of the latent variables are performed, with the fusion formulation expressed as:

$$z_{t-1}=M^{z}_{incon}\odot\tilde{z}_{t-1}+(1-M^{z}_{incon})\odot z^{known}_{t-1} \tag{12}$$

where, M^{z}_{incon} denotes the latent-space semantic contradiction mask, $\tilde{z}_{t-1}$ denotes the corrected latent variable predicted by the model based on text semantics, and z^{known}_{t-1} denotes the effective latent variables of the original image after noise addition. This computation enables dynamic repainting of defective regions while preserving intact regions unchanged, avoiding the disruption of already optimized grammatical structures and visual features through global reconstruction. To further constrain the semantic accuracy of the repainted content, a semantic gradient guidance mechanism is introduced in the post-sampling stage, through which latent features are optimized via gradient ascent to maximize the similarity between the visual features of the masked region and the target text phrase. The optimization process can be expressed as:

$$z_t\leftarrow z_t+\delta\cdot\nabla_{z_t} log\, sim\,(v_{masked}(z_t),t_{phrase}) \tag{13}$$

where, δ denotes the gradient update step size, v_{masked} denotes the local visual token features decoded from the masked region, and t_{phrase} denotes the semantic embedding of the corresponding textual modifier phrase. Concurrently, a semantic correction loss is introduced to perform end-to-end fine-tuning of the denoising module of the diffusion model, continuously enhancing the model's capacity for fine-grained semantic restoration.

The four functional modules proposed in this work form a complete processing pipeline characterized by progressive cascading and deep coupling. The overall data flow follows an optimization logic driven by the teaching text, sequentially accomplishing grammatically precise generation, textbook semantic fusion, cognitive attention enhancement, and fine-grained semantic error correction. Model training is conducted using a strategy of stage-wise pre-training combined with global joint fine-tuning. The module-specific loss functions jointly constrain the network parameter updates, ensuring stable propagation and lossless iteration of linguistic priors, spatial semantic relationships, and cognitive visual features throughout the pipeline. The entire architecture achieves full-process modeling from discrete English grammatical rules to continuously controllable visual generation, ultimately producing a high-quality image I_{out} characterized by fully aligned fine-grained semantics and full suitability for teaching scenarios, thereby providing a comprehensive technical paradigm for the precise generation and intelligent optimization of educational cross-modal visual content.

3. EXPERIMENTS

To systematically validate the effectiveness of the proposed cross-modal image enhancement and visual-semantic fusion framework, multi-dimensional comparative experiments and ablation studies were conducted, centering on grammatical semantic generation accuracy, cross-modal fusion quality, cognitive saliency augmentation effectiveness, fine-grained semantic restoration performance, and practical pedagogical application value. Through quantitative metric evaluation and qualitative visual analysis of the results, the technical gains of each core module and the overall superiority of the proposed framework were comprehensively verified.

3.1 Experimental setup

A dedicated image-text dataset tailored for English grammar teaching, designated as ET-Image, was constructed in this study. The dataset comprised a total of 15,000 high-quality image-text paired samples, comprehensively covering 12 types of English verb tense variants and 20 types of high-frequency spatial prepositional semantic relations. The data samples integrated publicly available teaching materials and standardized textbook illustration resources. All images were annotated with structured scene graphs, tense semantic labels, and learner eye-tracking fixation points, enabling simultaneous support for image semantic accuracy evaluation, visual saliency matching assessment, and pedagogical

effectiveness validation, thereby effectively addressing the deficiency of fine-grained English grammar annotations in existing general-purpose image-text datasets.

The overall framework of this work was built upon the pre-trained diffusion model Stable Diffusion v2 as the core generation network. The tense-aware encoder and the syntactic graph convolutional network were initialized through independent pre-training strategies. The cross-modal feature extraction module employed a contrastive language-image pre-training model with a ViT-B/32 architecture, and the saliency prediction network was task-adaptively fine-tuned based on the SalGAN architecture. All experimental images were uniformly configured at a resolution of 512 × 512. The AdamW optimizer was adopted for network optimization, with a unified learning rate decay strategy and batch size settings, ensuring experimental environment consistency across all comparative experiments and ablation studies, and eliminating hyperparameter interference.

A comprehensive evaluation system was constructed from five dimensions: image generation quality, grammatical semantic accuracy, fusion and augmentation performance, semantic restoration effectiveness, and pedagogical application validity. Image generation quality was assessed using general-purpose visual evaluation metrics, including Fréchet inception distance, inception score, and learned perceptual image patch similarity. Grammatical semantic accuracy was evaluated through tense classification accuracy, scene graph preposition relation recall, and image-text matching score. Fusion and saliency evaluation metrics included scene graph preservation rate, normalized saliency score, and correlation coefficient. Semantic restoration performance was measured by defect region detection precision, phrase alignment score, and local Fréchet inception distance fluctuation magnitude. Pedagogical effectiveness metrics encompassed learner answer accuracy rate, cognitive response time, eye-tracking path entropy, and subjective evaluation scores.

3.2 Grammatical semantic accuracy validation experiments

This experiment was designed to validate the modeling capability of the proposed grammar-decoupled diffusion generation module with respect to English tense features and spatial prepositional relations. Mainstream generative models were selected as comparative baselines, including the original Stable Diffusion v2, the text-semantically enhanced Stable Diffusion XL, and state-of-the-art language-conditioned layout generation models. English grammar teaching texts were uniformly input across all models, and the grammatical semantic precision and visual quality of the generated images were quantitatively compared. The experimental quantitative results are presented in Figure 5 and Table 1.

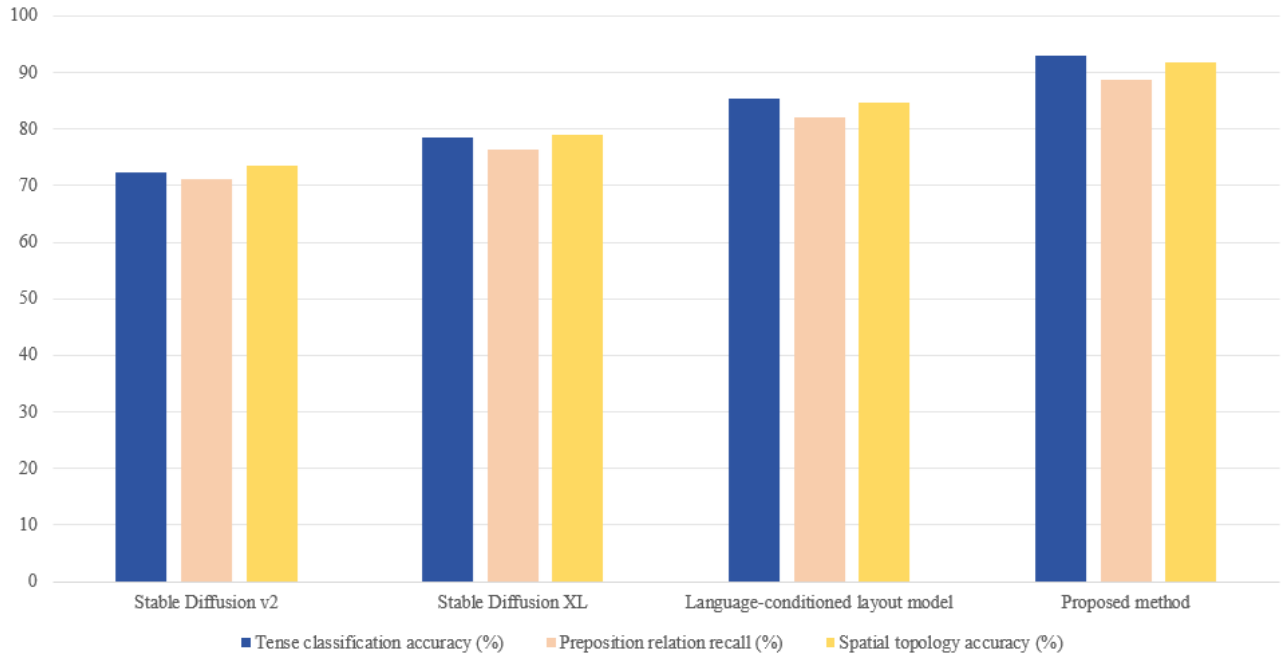

Figure 5. Comparison of grammatical semantic precision metrics across different models

Table 1. Comparison of grammatical semantic generation performance across different models

Model	Contrastive Language-Image Pre-Training Matching Score	Fréchet Inception Distance	Learned Perceptual Image Patch Similarity
Stable Diffusion v2	0.762	23.14	0.185
Stable Diffusion XL	0.791	19.87	0.152
Language-conditioned layout model	0.825	16.52	0.126
Proposed method	0.874	12.36	0.093

Note: The full name of "preposition relation recall" is scene graph preposition relation recall. Spatial topology accuracy is used to measure the degree of consistency between the relative positional relationships among objects and the textual descriptions. Lower Fréchet inception distance values indicate higher visual realism, and lower learned perceptual image patch similarity values indicate higher perceptual similarity.

From the quantitative results presented in Figure 5 and Table 1, it is evident that the proposed method significantly outperforms all comparative models across the core grammatical semantic metrics. In terms of tense feature modeling, a tense classification accuracy of 93.1% is achieved, representing a 20.7% improvement over the baseline diffusion model, thereby demonstrating that the tense-aware dynamic modulation mechanism effectively establishes the mapping relationship between linguistic tenses and visual dynamic features, enabling precise differentiation between the visual representations of progressive and perfective aspects. At the level of spatial semantic modeling, a scene graph preposition relation recall of 88.7% and a spatial topology accuracy of 91.8% are attained, substantially surpassing those of conventional models. This indicates that the syntactic graph bias attention mechanism effectively constrains the image layout logic, circumventing preposition-corresponding object positional misalignment from both entity relationship and topological structure dimensions. Concurrently, the proposed method achieves the lowest Fréchet inception distance and learned perceptual image patch similarity values and the highest image-text matching score, demonstrating that the incorporation of grammatical priors does not compromise visual perceptual quality, thereby achieving joint optimality in both semantic precision and visual quality.

Ablation study results reveal that when either the tense encoder or the syntactic graph attention bias module is removed, substantial degradation is observed in both tense recognition accuracy and preposition relation recall, accompanied by simultaneous declines in all visual metrics. This confirms that the two core mechanisms serve as indispensable pillars for grammatically precise generation. Qualitative visualization results demonstrate that the proposed method accurately reconstructs the visual topological structures of prototypical spatial relations such as "on," "in," and "under," without the semantic confusion issues commonly observed in conventional models.

3.3 Visual-semantic fusion and saliency augmentation experiments

This experiment was designed to validate the performance of the multi-granularity semantic fusion and pedagogical saliency augmentation modules. Traditional image fusion algorithms and mainstream visual enhancement models were selected as comparative methods, including Poisson fusion, deep Poisson fusion, direct image stitching, histogram equalization, and the Geometry-Aware Generative Adversarial Networks (GAGAN) enhancement algorithm. The generated image and the reference textbook image were fixed as inputs across all experiments, and evaluation was conducted from three dimensions: semantic fidelity, visual quality, and cognitive saliency matching degree. The experimental results are presented in Figure 6 and Table 2.

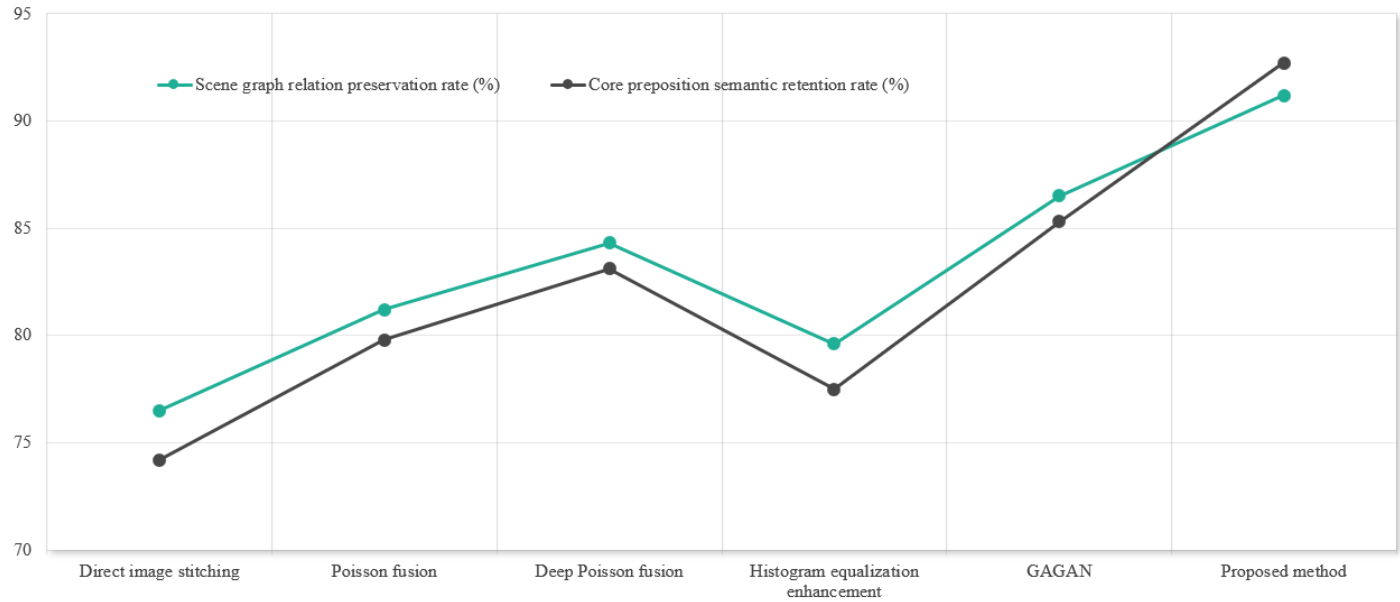

Figure 6. Quantitative comparison results of scene graph relation preservation rate and core preposition semantic retention rate

Table 2. Quantitative comparison results of other image fusion and saliency augmentation methods

Method	Fréchet Inception Distance	Learned Perceptual Image Patch Similarity	Normalized Scanpath Saliency	Correlation Coefficient
Direct image stitching	21.35	0.172	0.71	0.68
Poisson fusion	18.62	0.148	0.73	0.70
Deep Poisson fusion	17.15	0.131	0.75	0.72
Histogram equalization enhancement	19.83	0.145	0.78	0.75
GAGAN	15.48	0.112	0.81	0.79
Proposed method	11.94	0.087	0.89	0.86

Note: Core preposition semantic retention rate is specifically calculated for the spatial prepositional relations emphasized in teaching focal points. Higher values for the three metrics (normalized scanpath saliency, correlation coefficient, and information gain) indicate greater alignment between the saliency distribution and real learner fixation patterns.

From the data presented in Figure 6 and Table 2, it is evident that the proposed fusion and augmentation method achieves optimal comprehensive performance. At the semantic fusion level, a scene graph relation preservation rate of 91.2% and a core preposition semantic retention rate of 92.7% are attained by the proposed method, representing improvements of 4.7% and 7.4%, respectively, over the suboptimal model. Conventional fusion algorithms, which focus solely on pixel-level and stylistic consistency, are prone to disrupting the core spatial semantic relationships of instructional images. In contrast, the proposed method, through the scene graph alignment loss constraining the fusion process, maximally preserves the grammatical semantic structure, achieving synergistic unification of textbook style adaptation and semantic fidelity. At the visual cognitive augmentation level, the proposed method significantly outperforms global enhancement algorithms across the three saliency evaluation metrics of normalized scanpath saliency, correlation coefficient, and information gain, demonstrating that the text-driven adversarial saliency augmentation mechanism is capable of adaptively optimizing the image visual attention distribution according to the importance weights of teaching knowledge points, precisely guiding learner gaze toward the core instructional regions.

Ablation study results indicate that when the saliency alignment loss function is removed, the enhancement strategy of the model degenerates to globally uniform enhancement, with all three saliency metrics exhibiting substantial declines, and the images fail to highlight pedagogical focal points such as preposition-interacting objects and core action regions. User eye-tracking data further corroborate that the images processed by the proposed method effectively increase learner fixation duration on core instructional regions, significantly optimizing visual cognitive efficiency.

3.4 Semantic contradiction detection and restoration experiments

To validate the correction capability of the fine-grained semantic contradiction detection and localized repainting module, quantitative evaluation was conducted in this experiment from three dimensions: defect detection precision, semantic restoration effectiveness, and visual fidelity preservation, while simultaneously monitoring the impact of the restoration operations on overall image visual quality. The core metric results are presented in Table 3.

Table 3. Quantitative results of semantic contradiction detection and restoration

Evaluation Dimension	Metric	Numerical Result
Defect detection performance	mAP@IoU = 0.5 (%)	87.6
	mAP@IoU = 0.7 (%)	72.3
	Defect localization mean pixel error (px)	8.2
Semantic alignment restoration performance	Global phrase alignment score (pre-restoration)	0.42
	Global phrase alignment score (post-restoration)	0.81
	Color attribute defect alignment improvement rate (%)	92.5
	Quantity attribute defect alignment improvement rate (%)	85.7
Visual fidelity preservation performance	Local Fréchet inception distance fluctuation magnitude pre- vs. post-restoration (%)	4.20
	Global learned perceptual image patch similarity change pre- vs. post-restoration	0.012
	Structural similarity index preservation rate (%)	96.8

Note: mAP denotes mean average precision, and IoU denotes intersection over union. The alignment improvement rate refers to the proportion of the total defect deviation accounted for by the score improvement after restoration. The structural similarity index preservation rate denotes the ratio of structural similarity between the post-restoration image and the original image.

As can be observed from Table 3, the proposed module demonstrates high-precision semantic defect detection capability. A mean average precision of 87.6% is achieved for defect region detection under the intersection over union = 0.5 condition, with 72.3% detection precision maintained even under the more stringent intersection over union = 0.7 criterion, and an average pixel localization error of only 8.2 pixels. Such performance enables precise localization of fine-grained semantic errors in images, including entity color, quantity, and attribute mismatches. Following diffusion-based localized repainting correction, the image-text phrase alignment score for the defective regions is improved from 0.42 to 0.81, with a color attribute defect alignment improvement rate of 92.5%, representing a substantial leap in semantic matching precision. This demonstrates that the semantic gradient-guided diffusion-based restoration mechanism effectively rectifies various types of detailed semantic contradictions. Simultaneously, the local Fréchet inception distance fluctuation pre- and post-restoration is only 4.20%, the global learned perceptual image patch similarity change is controlled within 0.012, and the structural similarity index preservation rate reaches 96.8%, indicating that the localized repainting operation selectively corrects only the defective regions without disrupting the overall scene structure, stylistic characteristics, or correct semantic content of the image, thereby achieving an optimal balance between precise error correction and visual fidelity preservation. Qualitative case results further demonstrate that the module effectively corrects typical issues such as entity color errors and attribute mismatches, and is fully capable of accommodating the fine-grained semantic correction requirements of English language teaching images.

3.5 System-wide ablation experiments

To independently validate the individual contributions of the four core modules, the complete system was configured alongside four ablation baseline models, in which the grammar-decoupled generation module, the multi-granularity semantic fusion module, the saliency augmentation module, and the semantic contradiction restoration module were respectively removed. Semantic accuracy, visual quality, and pedagogical effectiveness metrics were uniformly evaluated across all models. The ablation experimental results are presented in Table 4.

From the ablation data, the functional value and progressive logic of each module can be clearly identified. When the grammar-decoupled generation module is removed, the most

substantial performance degradation is observed across all metrics, with tense and spatial semantic generation precision exhibiting severe declines, directly resulting in a significant reduction in learner answer accuracy. This confirms that grammatically controllable generation serves as the core foundation of the entire framework. When the semantic fusion module is removed, the stylistic compatibility between images and textbooks is diminished, the integrity of scene semantics is compromised, and instructional adaptability is reduced, accompanied by an obvious degradation in visual quality metrics. When the saliency augmentation module is removed, the cognitive guidance capability of images is lost and the visual focal points become blurred, with the normalized scanpath saliency metric exhibiting a marked decline, rendering the framework unable to adapt to learner cognitive patterns. When the semantic restoration module is removed, fine-grained semantic defects persist in the images, marginally affecting instructional effectiveness and semantic precision, while exerting the least impact on overall visual quality.

The overall ablation results fully demonstrate that the four modules are progressively complementary and deeply coupled. The absence of any single module results in performance degradation in the corresponding dimension of the model. The complete four-stage closed-loop architecture constitutes a necessary condition for achieving optimal instructional image generation and augmentation performance.

Table 4. Results of full-module ablation experiments

Model configuration	Learner Answer Accuracy (%)	Preposition Relation Recall (%)	Tense Accuracy (%)	Normalized Scanpath Saliency	Fréchet Inception Distance
Complete model (proposed)	89.4	88.7	93.1	0.89	12.36
Removal of the grammar-decoupled generation module	72.1	70.5	71.8	0.76	22.45
Removal of the multi-granularity semantic fusion module	78.5	79.2	88.2	0.79	16.72
Removal of the pedagogical saliency augmentation module	81.3	83.6	91.5	0.77	13.08
Removal of the semantic contradiction restoration module	85.2	85.1	92.0	0.83	12.97

3.6 User evaluation experiments in teaching scenarios

To validate the practical value of the proposed method in real English language teaching scenarios, a user-controlled experiment was conducted with 30 English learners recruited for this study. Four categories of instructional images were configured as comparative materials: images generated by the proposed method, images generated by the original diffusion model, original textbook images, and images without fusion or augmentation processing. Pedagogical effectiveness was comprehensively evaluated through three approaches: grammar learning task testing, eye-tracking data acquisition, and multi-dimensional subjective scoring. The user evaluation results are presented in Figure 7 and Table 5.

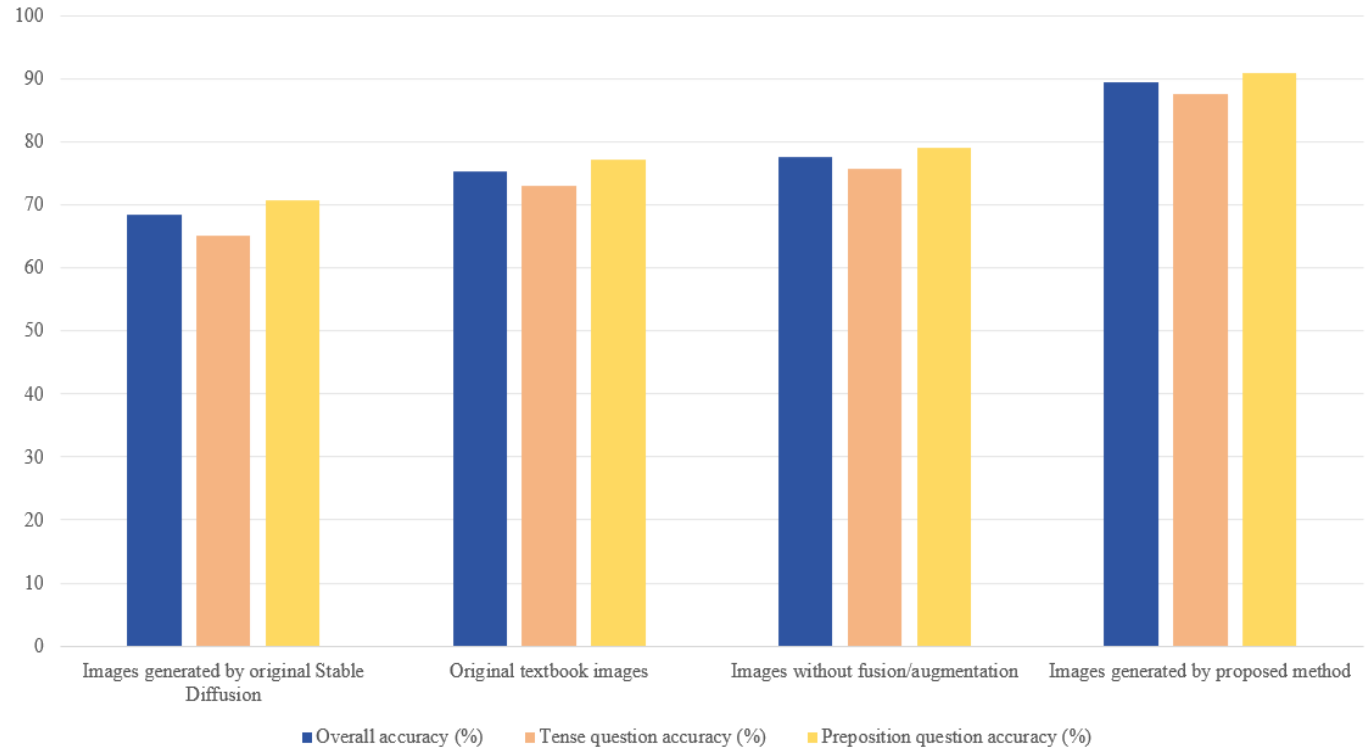

Figure 7. Accuracy results of user evaluation in teaching scenarios

Table 5. User evaluation results in additional teaching scenarios

Image Type	Eye-Tracking Path Entropy	Time to First Fixation (ms)	Core Region Fixation Proportion (%)	Semantic Accuracy	Visual Quality	Pedagogical Suitability	Overall Satisfaction
Original Stable Diffusion-generated images	1.86	1286	42.3	3.0	3.3	3.1	3.2
Original textbook images	1.52	987	58.7	4.0	4.2	4.1	4.1
Images without fusion/augmentation	1.48	924	61.2	3.7	3.9	3.8	3.8
Images generated using the proposed method	1.13	652	76.5	4.6	4.4	4.5	4.5

Note: Time to first fixation refers to the duration from when learners first view the image until their first fixation on the core instructional region; lower values indicate higher cognitive guidance efficiency. Core region fixation proportion denotes the ratio of fixation duration on core knowledge point regions to total fixation duration.

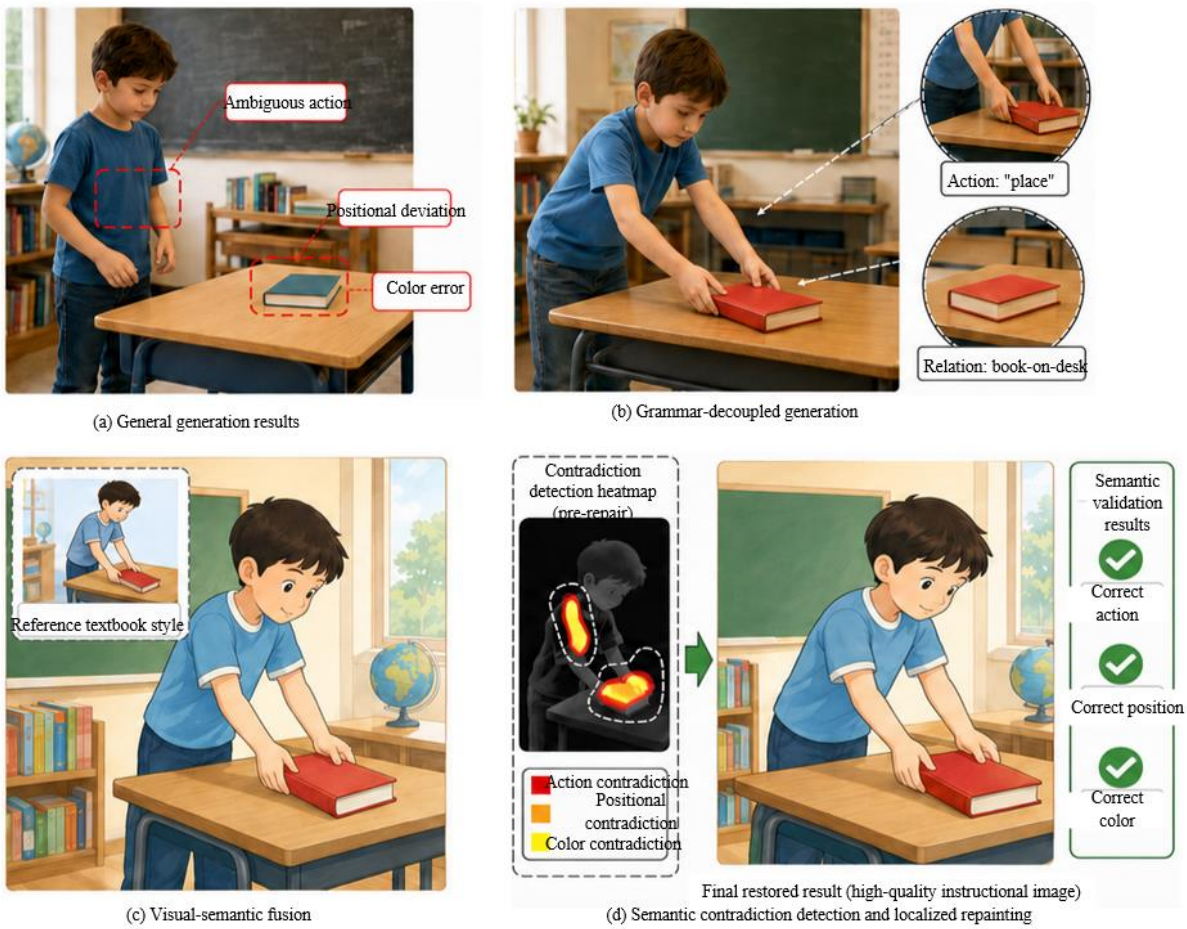

Figure 8. Implementation effect of cross-modal image enhancement and visual-semantic fusion

The user experimental data demonstrate that the instructional images generated by the proposed method achieve optimal pedagogical application effectiveness. Compared with conventional generated images, the proposed images improve learner preposition grammar learning accuracy by 20.1 percentage points and tense knowledge point accuracy by 22.4 percentage points, significantly enhancing learners' mastery of grammatical knowledge points. In terms of eye-tracking metrics, the images generated by the proposed method yield the lowest visual path entropy among learners, with a time to first fixation on the core region of only 652 ms and a core region fixation proportion reaching 76.5%. This indicates that learner visual attention is more concentrated, enabling rapid focus on core grammatical knowledge points, with ineffective visual search behaviors substantially reduced. At the subjective evaluation level, the proposed images outperform all comparative materials in semantic accuracy, pedagogical suitability, and overall satisfaction, with an overall satisfaction score of 4.5, demonstrating higher learner recognition of the visual quality and pedagogical adaptability of the proposed images.

The comprehensive subjective and objective experimental results confirm that the proposed framework not only achieves quantitative improvements in image semantic precision and visual quality, but also tangibly optimizes contextual English learning effectiveness, demonstrating substantial practical value for real teaching applications.

3.7 Visualization experimental results

To intuitively validate the overall processing capability of the proposed method in English language teaching image generation, semantic preservation, and fine-grained error correction, the visual results of the same instructional sentence at different processing stages are comprehensively demonstrated in this experiment. As shown in Figure 8, the general generation results exhibit issues such as ambiguous action state representation, color deviation of the book, and inaccurate object spatial relationships, indicating that reliance solely on textual statistical associations is insufficient for stable mapping of English tenses, attributes, and prepositional relations. Following the introduction of the grammar-decoupled mechanism, the human hand action, the red book, and its spatial topology on the desktop are simultaneously corrected, demonstrating that tense feature modulation and syntactic structure constraints effectively enhance the accuracy of transforming linguistic rules into visual structures. Subsequent visual-semantic fusion, while preserving the core entity relationships, unifies the composition and stylistic presentation of the textbook images. Pedagogical saliency augmentation further concentrates the visual response on key knowledge regions such as human actions, the book, and the desktop, thereby reducing interference from background information on learner attention. The semantic contradiction detection results are capable of precisely localizing local mismatches at the action, positional, and attribute levels, and performing targeted restoration through constrained localized repainting while preserving the original scene structure and non-defective regions. In summary, the proposed method forms a closed-loop optimization pipeline encompassing grammar-controllable generation, cross-modal semantic fusion, cognitively oriented augmentation, and fine-grained semantic correction. This pipeline significantly enhances image-text semantic consistency, spatial relationship accuracy, and pedagogical visual adaptability while improving image naturalness, thereby providing effective technical support for the construction of high-quality English language teaching visual resources driven by generative artificial intelligence.

4. DISCUSSION

The syntactic graph bias attention mechanism introduced in this work exhibits excellent interpretability, effectively ameliorating the uncontrollability and implicit black-box characteristics inherent in the attention matching processes of conventional text-to-image models. This mechanism constructs a structured prior matrix based on textual syntactic dependency structures, transforming abstract prepositional spatial topological relations into learnable attention bias constraints, and precisely modulating the feature association strength between image visual regions and textual semantic units. Compared with cross-modal matching approaches that rely solely on statistical data-driven learning, explicit syntactic constraints are capable of guiding the model to establish stable object spatial correspondences during the scene layout stage, fundamentally reducing generation defects arising from semantic confusion across various spatial prepositions. Simultaneously, the pedagogical saliency augmentation mechanism exhibits differentiated optimization effects across different preposition knowledge points. While the improvement in visual discriminability for basic spatial relations is steady and significant, the cognitively oriented non-uniform enhancement strategy, in complex prepositional scenarios involving occlusion and nesting logic, is capable of prominently highlighting the core regions of entity interactions while suppressing interference from irrelevant background information. The auxiliary gain for challenging grammar learning is particularly pronounced, thereby validating the necessity of task-specific image enhancement strategies.

The proposed framework demonstrates stable performance advantages in routine English language teaching image-text processing tasks; however, certain performance limitations persist in complex semantic scenarios. When confronted with overly lengthy teaching texts involving multiple entities and nested grammatical structures, the corresponding syntactic graph topologies become more intricate, rendering the graph convolutional encoding process susceptible to feature coupling and relational redundancy, which in turn leads to a degradation in the model's parsing accuracy for multi-level spatial semantics. Constrained by the sample distribution of the dataset, the model's generalization capability for modeling low-frequency, rare prepositional relations remains relatively weak, rendering the precise visual expression of various niche spatial semantics difficult to achieve. Furthermore, the performance of the semantic contradiction detection module is significantly influenced by the threshold parameter. When the threshold value is set too high, numerous fine-grained attribute defects are filtered out, resulting in missed detections of semantic errors. Conversely, when the threshold is set too low, redundant contradiction masks are generated, triggering unnecessary localized repainting perturbations that disrupt the already optimized visual and semantic structures of the image. This parameter sensitivity imposes certain constraints on the general deployment of the model.

The core innovations and academic contributions of this work are concentrated in the fields of cross-modal image processing and visual generation, with English Language Teaching scenarios serving merely as a highly constrained and standardized validation vehicle and application context. All functional modules throughout the paper are centered around core computer vision technologies, with conditional diffusion generation, multi-scale cross-modal fusion, cognitive saliency modeling, and semantically constrained image inpainting serving as the technical kernel. Linguistic rules are employed solely as computable priors for achieving precisely controllable regulation of visual tasks. This work does not focus on research into language teaching theories or pedagogical strategies. Rather, through the construction of a mapping pathway from linguistic semantics to visual features, the industry-wide problems of insufficient semantic controllability and poor task adaptability in general-purpose image generation and enhancement methods are addressed. A novel technical paradigm and feasible optimization directions are thereby provided for general computer vision tasks such as semantically driven controllable image generation, task-oriented visual enhancement, and fine-grained image-text alignment correction.

5. CONCLUSION

A four-stage closed-loop cross-modal image generation and enhancement framework was constructed in this work, sequentially encompassing grammar-decoupled conditional

diffusion generation, multi-granularity visual-semantic fusion, pedagogical saliency-based adversarial augmentation, and fine-grained semantic contradiction automatic correction. The core research rationale lies in the transformation of discrete English grammatical rules into continuous, differentiable visual modulation signals. Through multiple original mechanisms—including tense feature modulation, syntactic graph attention bias, text-driven dynamic fusion, cognitively oriented differentiated augmentation, and image-text alignment-guided localized repainting—the inherent deficiencies of general-purpose text-to-image models were systematically addressed, including tense expression distortion, spatial prepositional relation misalignment, imbalanced visual attention distribution, and the inability to automatically correct fine-grained semantic deviations. Leveraging a self-constructed, comprehensively annotated English language teaching image dataset, multiple sets of quantitative comparative experiments, module ablation studies, and real-learner evaluation experiments were conducted. All evaluation metrics demonstrated that the proposed framework simultaneously achieved compatibility among image visual quality, image-text semantic matching precision, and practical pedagogical effectiveness, forming a general-purpose technical solution for image enhancement and editing in finely constrained semantic scenarios.

The validity of the proposed method was verified through English language teaching scenarios, and the entire visual processing pipeline exhibited substantial potential for transferability and extensibility. In future work, the technical framework can be extended to the automated production of educational visual materials in other disciplines such as mathematics, science, and humanities. Furthermore, the core modules—including explicit structured semantic prior injection, cognitive saliency modeling, and fine-grained image-text alignment correction—are also transferable to the field of embodied intelligence, serving tasks such as robotic environmental perception and multi-modal visual reasoning, thereby providing a feasible technical reference direction for the design of semantically controllable cross-modal vision systems.

REFERENCES

[1] Croitoru, F.A., Hondru, V., Ionescu, R.T., Shah, M. (2023). Diffusion models in vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(9): 10850-10869. https://doi.org/10.1109/TPAMI.2023.3261988

[2] Cao, H., Tan, C., Gao, Z., et al. (2024). A survey on generative diffusion models. IEEE Transactions on Knowledge and Data Engineering, 36(7): 2814-2830. https://doi.org/10.1109/TKDE.2024.3361474

[3] Noetel, M., Griffith, S., Delaney, O., et al. (2021). Video improves learning in higher education: A systematic review. Review of Educational Research, 91(2): 204-236. https://doi.org/10.3102/0034654321990713

[4] Çeken, B., Taşkın, N. (2022). Multimedia learning principles in different learning environments: A systematic review. Smart Learning Environments, 9(1): 19. https://doi.org/10.1186/s40561-022-00200-2

[5] Kasneci, E., Sessler, K., Küchemann, S., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103: 102274. https://doi.org/10.1016/j.lindif.2023.102274

[6] Crompton, H., Edmett, A., Ichaporia, N., Burke, D. (2024). AI and English language teaching: Affordances and challenges. British Journal of Educational Technology, 55(6): 2503-2529. https://doi.org/10.1111/bjet.13460

[7] Jiang, R., Zheng, G.C., Li, T., Yang, T.R., Wang, J.D., Li, X. (2024). A survey of multimodal controllable diffusion models. Journal of Computer Science and Technology, 39(3): 509-541.

[8] Bie, F.X., Yang, Y.B., Zhou, Z.Z., et al. (2025). RenAIssance: A survey into AI text-to-image generation in the era of large model. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3): 2212-2231. https://doi.org/10.1109/TPAMI.2024.3522305

[9] Zhang, J., Huang, J., Jin, S., Lu, S. (2024). Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(8): 5625-5644. https://doi.org/10.48550/arXiv.2304.00685

[10] Chen, F.L., Zhang, D.Z., Han, M.L., et al. (2023). VLP: A survey on vision-language pre-training. Machine Intelligence Research, 20(1): 38-56. https://doi.org/10.48550/arXiv.2202.09061

[11] Wang, X., Chen, G.Y., Qian, G.W., et al. (2023). Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research, 20(4): 447-482.

[12] Gao, J., Li, P., Chen, Z., Zhang, J. (2020). A survey on deep learning for multimodal data fusion. Neural Computation, 32(5): 829-864. https://doi.org/10.1162/neco_a_01273

[13] Zhang, X., Demiris, Y., Guo, Y. (2023). Visible and infrared image fusion using deep learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(8): 10535-10554.

[14] Bayoudh, K., Knani, R., Hamdaoui, F., Mtibaa, A. (2022). A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. The Visual Computer, 38(8): 2939-2970. https://doi.org/10.1007/s00371-021-02166-7

[15] Li, Y., Daho, M.E.H., Conze, P.H., et al. (2024). A review of deep learning-based information fusion techniques for multimodal medical image classification. Computers in Biology and Medicine, 177: 108635. https://doi.org/10.1016/j.compbiomed.2024.108635

[16] Zhao, F., Zhang, C., Geng, B. (2024). Deep multimodal data fusion. ACM Computing Surveys, 56(9): 1-36. https://doi.org/10.1145/3649447

[17] Zhai, G., Min, X. (2020). Perceptual image quality assessment: A survey. Science China Information Sciences, 63(11): 211301. https://doi.org/10.1007/s11432-019-2757-1

[18] Li, C.Y., Zhang, Z.C., Wu, H.N., et al. (2024). AGIQA-3K: An open database for AI-generated image quality assessment. IEEE Transactions on Circuits and Systems for Video Technology, 34(8): 6833-6846. https://doi.org/10.48550/arXiv.2306.04717

[19] Borji, A. (2021). Saliency prediction in the deep learning era: Successes and limitations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 679-700. https://doi.org/10.1109/TPAMI.2019.2935715

[20] Yan, F., Chen, C., Xiao, P., Qi, S., Wang, Z., Xiao, R. (2022). Review of visual saliency prediction:

Development process from neurobiological basis to deep models. Applied Sciences, 12(1): 309. https://doi.org/10.3390/app12010309
[21] Huang, Y., Huang, J.C., Liu, Y.F., et al. (2025). Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(6): 4409-4437. https://doi.org/10.1109/TPAMI.2025.3541625
[22] Xu, Z.S., Zhang, X.F., Chen, W., et al. (2023). A review of image inpainting methods based on deep learning. Applied Sciences, 13(20): 11189. https://doi.org/10.3390/app132011189
[23] Qin, Z., Zeng, Q., Zong, Y., Xu, F. (2021). Image inpainting based on deep learning: A review. Displays, 69: 102028. https://doi.org/10.1016/j.displa.2021.102028
[24] He, C.M., Shen, Y.Q., Fang, C.Y., et al. (2025). Diffusion models in low-level vision: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(6): 4630-4651. https://doi.org/10.1109/TPAMI.2025.3545047