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

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 785-797
Journal homepage: http://iieta.org/journals/ts

Modeling and Deep Optimization of an English Learning Assistance System via Multimodal Fusion of Image Captioning and Visual Semantic Embedding

Weifeng Deng[1], Lin Wang[1*], Xue Deng[2], Fan Gu[3]

[1] Hainan Vocational University of Science and Technology, Haikou 571127, China
[2] Department of Physical Education, Hebei Vocational University of Technology and Engineering, Xingtai 054000, China
[3] School of Education and Music, Hainan Vocational University of Science and Technology, Haikou 571127, China

Corresponding Author Email: 249550389@qq.com

Copyright: ©2026 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

https://doi.org/10.18280/ts.430219

Received: 19 October 2025
Revised: 22 February 2026
Accepted: 12 March 2026
Available online: 30 April 2026

Keywords:
image processing, visual semantic embedding, image captioning, cross-modal fusion, English learning assistance, multi-granularity feature fusion, gated attention, adversarial optimization

ABSTRACT

With the rapid penetration of cross-modal fusion technologies in education, image-assisted learning has emerged as a vital approach to enhancing English learning efficiency. However, existing methods of image captioning and visual semantic embedding exhibit significant limitations in educational scenarios, struggling to simultaneously satisfy requirements for image processing precision, fine-grained semantic alignment, and personalized learning needs. To address these challenges, this paper investigates the system modeling and deep optimization strategies for an English learning assistance system that integrates image captioning with visual semantic embedding. First, a multi-granularity cross-layer fusion visual encoder is designed, leveraging the complementary strengths of ConvNeXt and Swin Transformer. An adaptive fusion mechanism based on learnable gating units is introduced to effectively integrate local texture features with global semantic representations, thereby improving the accuracy of visual feature extraction. Second, a hierarchical visual semantic embedding method is proposed, constructing dual-granularity embedding spaces at both word and phrase levels. By introducing the structural Wasserstein distance, the method enhances multi-modal similarity measurement and achieves fine-grained alignment between visual features and English semantics. On this basis, a visual-semantic gated attention generator is developed, employing dynamic gating mechanisms to adaptively regulate the balance between visual and semantic attention, which significantly improves the accuracy and robustness of caption generation. Simultaneously, an error-driven adversarial optimization strategy is introduced; an error-pattern discriminator is constructed using prior distributions of learners' linguistic errors, ensuring that generated descriptions are not only visually faithful but also pedagogically appropriate. Furthermore, a cross-modal adaptive curriculum learning method is proposed, defining a joint difficulty metric based on visual entropy and linguistic complexity to dynamically adjust training strategies. A visual-semantic alignment heatmap is designed to facilitate deep integration of learning feedback with image processing techniques. This study not only overcomes the adaptation bottlenecks of existing multimodal fusion technologies in educational contexts, but also extends the application boundaries of image processing in education, offering a novel research paradigm and technical support for the interdisciplinary field of computer vision and language learning.

1. INTRODUCTION

The rapid development of cross-modal fusion technology has provided significant support for interdisciplinary innovation [1]. In particular, the deep integration of image processing and natural language processing has demonstrated broad application prospects in the field of intelligent education [2, 3]. As one of the most widely adopted second language learning scenarios globally [4], English learning benefits from image-assisted instruction, which leverages intuitive and concrete advantages to effectively reduce learners' vocabulary memorization difficulty and enhance sentence application ability. This approach serves as a critical pathway to resolving the pain points of "abstractness and contextual disconnect" in traditional English learning [5]. As a core medium connecting visual information and linguistic expression [6], image captioning technology determines the effectiveness of image-assisted English learning; furthermore, core modules within image processing technology, such as feature extraction and visual semantic alignment, form the foundation for ensuring the accuracy and granularity of image captioning [7]. Currently, existing image captioning and visual semantic embedding technologies are primarily designed for general-purpose scenarios [8, 9]. They lack targeted adaptation to the personalized needs of English learning, struggle to simultaneously capture precise local image details and express complete global semantics, and fail to meet learners' core demands for fine-grained object description and contextualized sentence expression. Moreover, the insufficient integration of image processing technology with English learning feedback [10, 11] limits the application effectiveness of relevant systems in practical educational scenarios. This status quo presents an urgent demand for cross-disciplinary research combining image processing and education. Conducting research on the modeling and deep optimization of English learning assistance systems that integrate image captioning and visual semantic embedding can not only specifically address the adaptation bottlenecks of existing technologies in educational scenarios—providing personalized and precise assistance for English learners—but also expand the application boundaries of image processing technology. It promotes the deep integration of cross-modal fusion technology and intelligent education, offering new research perspectives and technical paradigms for the intersection of image processing and education [12], thus holding significant theoretical value and practical application significance.

Despite significant progress in cross-modal fusion and image captioning technologies, and the application of related research findings across various fields, specialized research targeting English learning assistance scenarios still faces numerous urgent deficiencies. These shortcomings severely restrict the practical implementation of these technologies in educational settings and constitute the core gap in current research. At the level of visual feature extraction, existing methods mostly adopt single Convolutional Neural Network or Transformer architectures [13], either focusing on extracting local texture details while neglecting the correlation of global semantic structures, or concentrating on capturing global semantics while losing fine-grained local features. The cross-layer feature fusion process [14] lacks effective adaptive adjustment mechanisms, making it impossible to dynamically balance the weights of local and global features according to image content complexity, and thus struggling to meet the dual requirements of English learning, which necessitate both precise descriptions of individual object details and complete expressions of overall scene meanings. At the level of visual semantic embedding, most existing multimodal alignment methods employ single-granularity embedding strategies [15], failing to fully account for the hierarchical semantic requirements of words, phrases, and scenes in English learning, resulting in insufficient alignment precision between visual features and English semantics [16]; meanwhile, similarity measurement methods mostly rely on traditional cosine similarity [17], which cannot effectively capture distributional structural differences between visual features and textual semantics, further reducing the reliability of multimodal alignment. At the level of caption generation, attention mechanisms in existing models mostly adopt fixed-weight allocation methods [18], lacking dynamic selection capabilities, and cannot adaptively switch the focus between visual and semantic attention based on image clarity, semantic complexity, and learner proficiency differences. This results in generated descriptions that are either overly verbose—exceeding the learner's comprehension range—or too brief—lacking key semantic information—making it difficult to adapt to the personalized needs of English learners at different proficiency levels. At the level of optimization strategies, existing model optimizations mostly focus on improving the general performance of caption generation [19], lacking targeted optimization based on common error patterns of English learners and lacking adaptive learning mechanisms for images of varying difficulties, leading to insufficient model generalization. More critically, existing research generally lacks the deep integration of image processing technology and learning feedback [20]; it cannot transform image feature analysis and semantic alignment results into effective learning feedback information, failing to meet the core requirements of image processing journals regarding technical practicality and implementability, nor can it truly leverage the assistive role of technology in English learning.

Addressing the aforementioned research gaps, this paper conducts research on the modeling and optimization of English learning assistance systems, with core contributions including five aspects: designing a multi-granularity cross-layer fusion visual encoder that combines the advantages of ConvNeXt and Swin Transformer, utilizing learnable fusion gating to dynamically balance local and global features, thereby enhancing image feature representation precision; proposing a hierarchical visual semantic embedding method that constructs dual-granularity embedding spaces at the word and phrase levels, introducing structural Wasserstein distance to optimize similarity measurement and achieving fine-grained alignment between visual and English semantics; designing a visual-semantic gated attention generator that regulates the ratio of visual to semantic attention through dynamic gating, improving the accuracy and robustness of caption generation; introducing an error-driven adversarial optimization strategy that constructs a discriminator based on priors of English learning error distributions and incorporates consistency constraints, ensuring that descriptions adhere to image content while adapting to learning needs; and proposing a cross-modal adaptive curriculum learning method that defines difficulty metrics based on visual entropy and linguistic complexity, dynamically adjusting training weights, and designing visual-semantic alignment heatmaps to achieve deep integration of image processing and learning feedback.

The subsequent structure of this paper is as follows: Chapter 2 provides an overview of related work, focusing on the status quo and gaps in core research directions, and clarifies the distinctions between this paper and existing studies. Chapter 3 elaborates on the system framework, technical details of each core module, and core formulas. Chapter 4 validates the effectiveness of the proposed method through multiple sets of experiments and quantitatively analyzes the contribution of each module. Chapter 5 analyzes the advantages of the proposed method based on experimental results, dissects its limitations, and suggests future research directions. Chapter 6 summarizes the core work and innovations and reaffirms the research value and application prospects.

2. METHOD

2.1 Overall system framework

The English learning assistance system integrating image captioning and visual semantic embedding, constructed in this paper, centers on high-precision image processing and deep multimodal semantic fusion. The overall architecture is shown in Figure 1, clearly presenting the logical connections and input-output closed loop of the five core modules. The system takes raw images and English learning corpora as initial inputs; first, it completes image feature extraction through a multi-granularity cross-layer fusion visual encoder, outputting high-quality visual features that combine local texture details and global semantic structures, providing core support for subsequent multimodal alignment. The extracted visual features are fed into the hierarchical visual semantic embedding module to complete fine-grained alignment with English word-level and phrase-level semantics, generating unified-dimensional visual-semantic embedding features. Subsequently, a visual-semantic gated attention generator generates precise image captions adapted to English learning scenarios based on these embedding features. To further enhance the accuracy of captions and learning adaptability, an error-driven adversarial optimization module is introduced to iteratively optimize the generator, combining priors of English learning error distribution to ensure the rationality and practicality of the descriptions. Finally, a cross-modal adaptive curriculum learning module dynamically adjusts training sample weights, and simultaneously generates visual-semantic alignment heatmaps, transforming image processing results into intuitive learning feedback to achieve deep integration of image processing technology and English learning assistance. The collaborative effort of all modules forms an end-to-end intelligent assistance system, ensuring the precision of visual feature extraction, the fineness of semantic alignment, and the adaptability of caption generation.

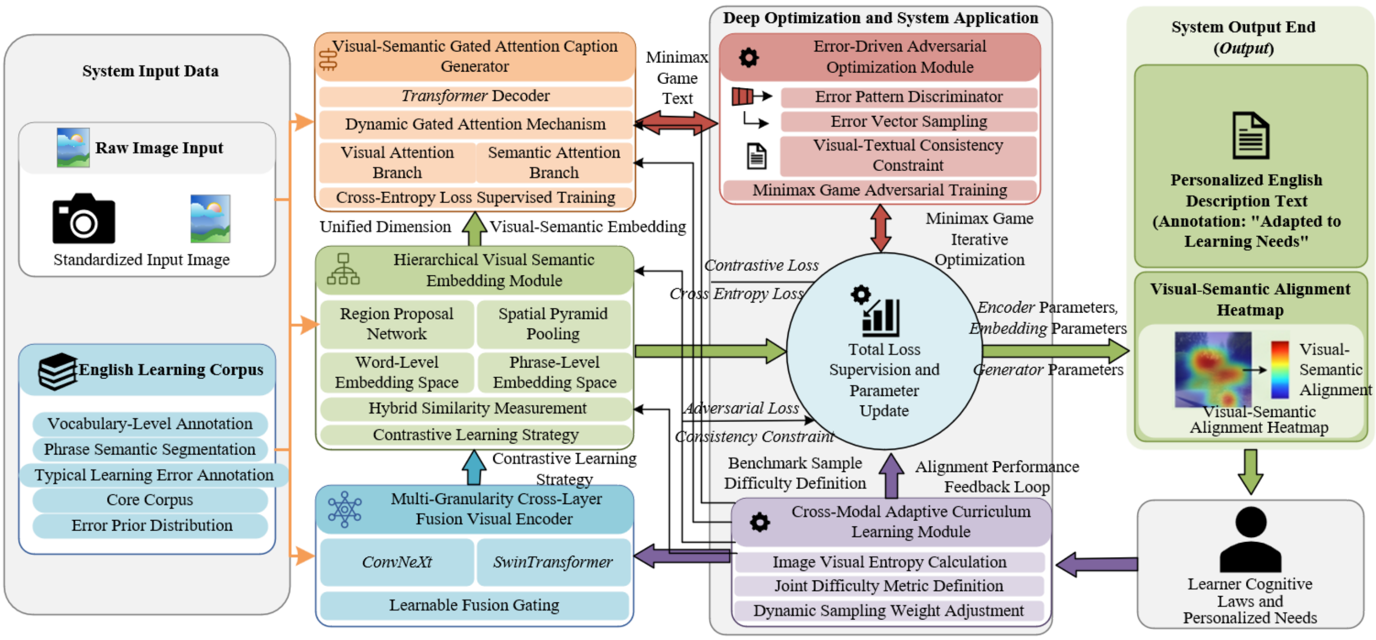

Figure 1. Overall architecture of the English learning assistance system integrating image captioning and visual semantic embedding

2.2 Multi-granularity visual feature extraction and cross-layer fusion

The multi-scale representation capability of visual features is the foundation for achieving fine-grained cross-modal understanding. To balance spatial detail preservation and high-level semantic modeling, this section adopts a dual-branch collaborative encoding structure to complete image feature parsing, and the network structure diagram is shown in Figure 2. Defining the standardized input image as $I \in \mathrm{R}^{3 \times H_0 \times W_0}$, relying on a lightweight convolutional encoding network to complete hierarchical feature iterative extraction, and passing through multi-level downsampling and feature transformation, four groups of feature maps with different scales F_1, F_2, F_3, F_4 are sequentially output. Shallow network features retain rich edge textures and spatial details, while deep features complete high-level semantic abstraction and target semantic condensation. The hierarchical output of multi-stage features covers visual expression requirements of different granularities, providing diverse visual representations for subsequent cross-modal semantic matching. Although convolutional encoding has efficient computational advantages in local spatial feature capture, its inherent local receptive field constraint makes it difficult to model long-range scene associations. To compensate for global semantic modeling capabilities, this paper adds a global semantic enhancement branch, inputting the middle-level feature F_3 into the Swin Transformer structure. This level of feature balances computational overhead and modeling performance by retaining basic spatial structure while fusing primary semantic information. Relying on the shifted window self-attention mechanism, the module can establish long-distance dependencies between image regions and mine global contextual information such as scene layout and target interaction, ultimately generating global semantic features F_{swin}, thereby constructing a dual-path feature expression system parallelizing local details and global semantics. To achieve the organic integration of the two heterogeneous features, this paper constructs an adaptive gated fusion module to dynamically complete the weighted aggregation of deep convolutional features and global semantic features. First, an upsampling operation $\mathrm{U_P}(\ )$ is performed on the deep feature F_4 to complete feature scale unification and matching. Global average pooling operators are adopted to compress the spatial dimension and reduce redundant information interference. The pooling operation can be defined as:

$$\mathrm{AvgPool}(F) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F(i,j) \tag{1}$$

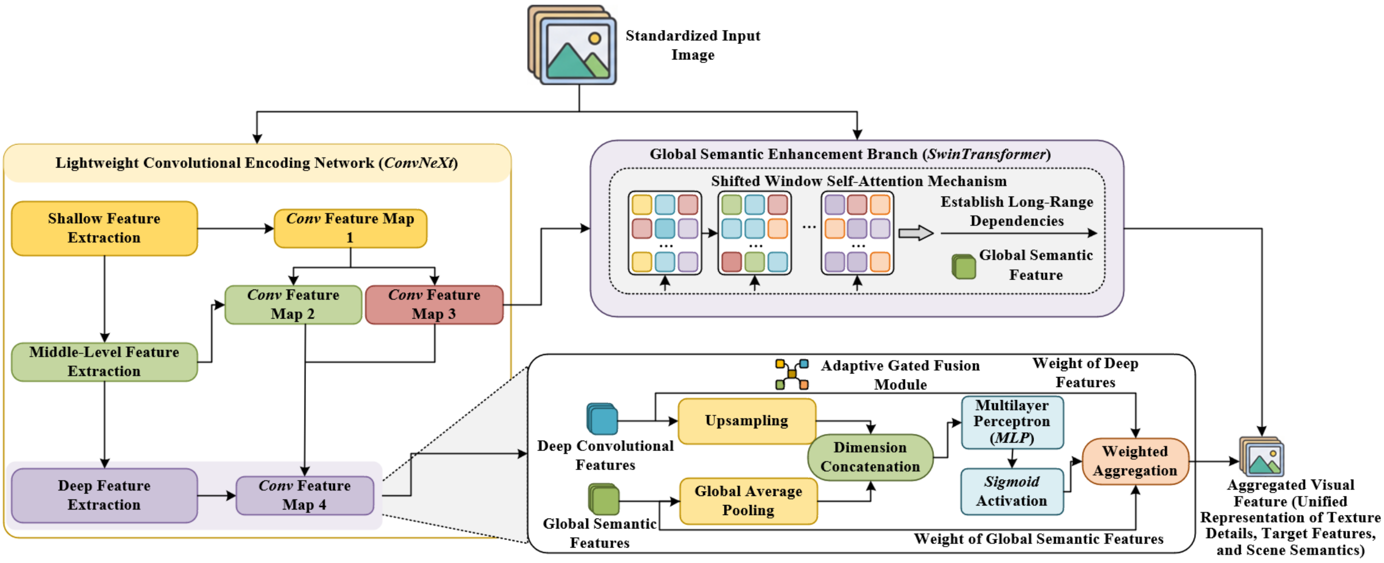

Figure 2. Network structure diagram of multi-granularity visual feature extraction and adaptive cross-layer fusion

The two sets of pooled feature vectors are concatenated and fed into a multilayer perceptron (MLP). After learning feature correlations through nonlinear transformation, continuous fusion weight γ is generated via a Sigmoid activation function. The overall fusion process follows the calculation form below:

$$F_v=\gamma\cdot \mathrm{Up}(F_4)+(1-\gamma)\cdot F_{swin},\\ \gamma=\sigma\left(\mathrm{MLP}\left([\mathrm{AvgPool}(F_4);\mathrm{AvgPool}(F_{swin})]\right)\right) \tag{2}$$

The weight coefficients can be autonomously updated iteratively based on image content complexity and target density, dynamically adjusting the contribution ratio of the two types of features to avoid the problem of singular representation caused by fixed fusion weights.

The aggregated visual feature F_v, outputted by adaptive fusion, integrates the fine-grained perception advantages of convolutional structures with the global modeling capabilities of Transformers, realizing the unified representation of texture details, target features, and scene semantics. The fusion strategy weakens the feature distribution differences brought by different network architectures and strengthens the robustness and integrity of visual representation. This optimized visual feature can adapt to hierarchical semantic parsing tasks; it can support the precise description of local objects while satisfying the complete semantic expression of the overall scene, providing stable and reliable front-end visual support for subsequent visual semantic embedding alignment and English text generation tasks.

2.3 Hierarchical visual semantic embedding and contrastive learning

Figure 3 shows the interaction diagram of hierarchical visual semantic embedding and visual-semantic gated attention mechanism. Among them, the core of visual semantic embedding is to establish precise mapping between visual features and language semantics; however, there exist hierarchical semantic requirements of words, phrases, and scenes in English learning scenarios, and single-granularity embedding makes it difficult to achieve precise alignment of multi-dimensional semantics. To this end, this paper constructs a hierarchical visual semantic embedding method, synchronously generating dual-granularity embedding features at the word and phrase levels, realizing hierarchical matching between visual information and English semantics, and providing refined multimodal support for subsequent scenario-based description generation. Both dual-granularity embedding spaces adopt a unified dimension d, where the word-level embedding matrix $E_{word}\in \mathrm{R}^{N_w\times d}$corresponds to the semantics of basic English vocabulary, and the phrase-level embedding matrix $E_{phrase}\in \mathrm{R}^{N_p\times d}$ corresponds to the semantics of common English phrases, ensuring the interoperability of semantic features at different levels.

Word-level embedding is constructed based on an English learning core corpus, performing semantic encoding on vocabulary through a pre-trained language model to retain contextual association information and realize fine-grained semantic representation. Phrase-level embedding combines visual region features and linguistic phrase semantics; first, object regions are extracted from the multi-granularity visual feature F_v via a Region Proposal Network to screen out semantically meaningful image regions, and then a Spatial Pyramid Pooling operation is performed on each region to capture regional feature information at different scales. Spatial Pyramid Pooling can be expressed as:

$$\mathrm{SPP}(F)=\\ [\mathrm{MaxPool}(F,k_1);\mathrm{MaxPool}(F,k_2);\mathrm{MaxPool}(F,k_3)] \tag{3}$$

where, k_1,k_2,k_3 are pooling windows of different scales. Fixed-dimensional regional feature vectors are obtained through feature concatenation, which are subsequently fused with corresponding English phrase semantic encodings to generate phrase-level visual semantic embeddings, realizing precise correspondence between visual regions and phrase semantics.

To improve the alignment precision of the dual-granularity embedding space, this paper designs a hybrid similarity measurement function, combining the advantages of cosine similarity and structural Wasserstein distance, balancing the capture of feature vector correlation and distribution differences. Cosine similarity is used to measure the linear correlation degree between visual features and semantic features, while structural Wasserstein distance quantifies the distribution difference of the two types of features in the embedding space, effectively compensating for the defect that traditional similarity measures cannot capture distributional structural differences. The similarity calculation is as follows:

$$s(v,c)=\lambda\cdot\cos(v,c)+(1-\lambda)\cdot exp\left(-\frac{W_2(p_v,p_c)}{\tau_w}\right) \quad (4)$$

where, $\lambda\in[0,1]$ is the adaptive weight, τ_w is the temperature coefficient, p_v and p_c represent the distribution probabilities of visual features and semantic features respectively, and $W_2()$ is the second-order Wasserstein distance. Its simplified calculation form is:

$$W_2(p_v,p_c)=\inf_{\gamma\in\Pi(p_v,p_c)}\sqrt{\int\|x-y\|^2 d\gamma(x,y)} \quad (5)$$

$\Pi(p_v,p_c)$ is the set of joint distributions of p_v and p_c. A contrastive learning strategy is adopted to jointly optimize the dual-granularity embedding space, constructing visual-semantic positive and negative sample pairs to guide the alignment of embedding features. Positive sample pairs consist of matched visual features and semantic features, while negative sample pairs are non-matched feature combinations. Matching feature pairs are clustered and non-matching feature pairs are separated by minimizing the contrastive loss. The loss function is defined as:

$$L_{\text{HVSE}}=-\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{exp(s(v_i,c_i^+)/\tau)}{\sum_{j=1}^{B}exp(s(v_i,c_j)/\tau)}+\log\frac{exp(s(c_i,v_i^+)/\tau)}{\sum_{j=1}^{B}exp(s(c_i,v_j)/\tau)}\right] \quad (6)$$

where, B is the batch size, τ is the temperature coefficient of the contrastive loss, v_i^+ and c_i^+ are the positive samples of v_j and c_i, respectively. During the contrastive learning process, word-level and phrase-level embedding spaces are optimized synchronously to ensure that both fine-grained vocabulary semantics and mid-grained phrase semantics align precisely with visual features, providing a high-quality multimodal feature basis for subsequent visual-semantic gated attention generation and adapting to the dual requirements of vocabulary application and phrase expression in English learning.

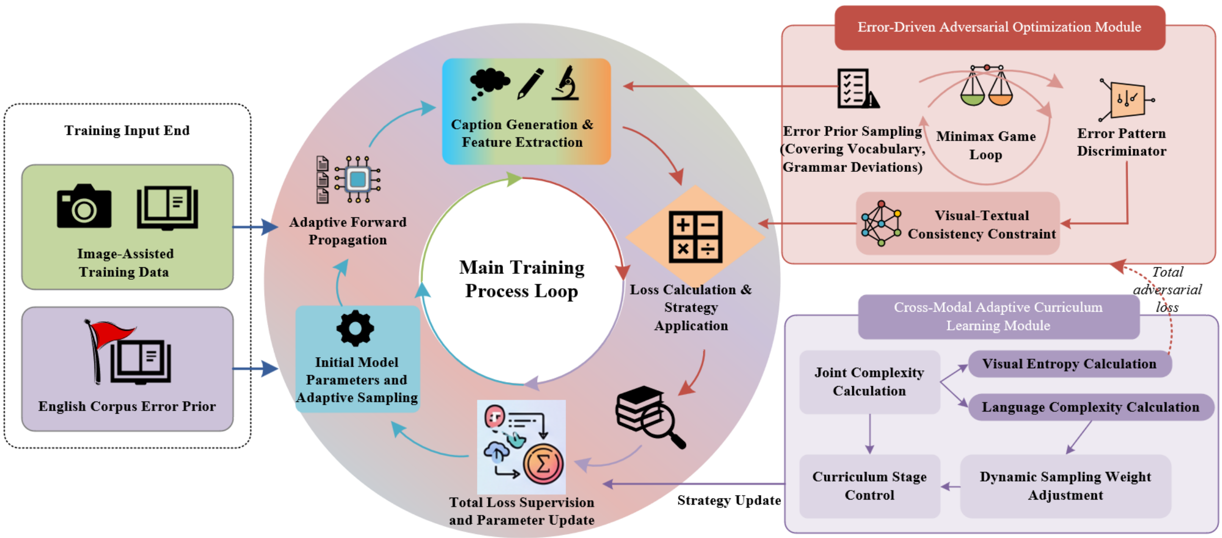

Figure 3. Interaction diagram of hierarchical visual semantic embedding and visual-semantic gated attention mechanism

2.4 Visual-semantic gated attention caption generator

The core of image captioning lies in generating coherent, accurate text expressions adapted to English learning scenarios based on visual features and semantic information. This paper constructs a caption generation framework based on a Transformer decoder, focusing on achieving dynamic fusion of visual information and semantic information through a gated attention mechanism to adapt to caption generation tasks under different image complexities and English learning requirements. The decoder takes the features output by the hierarchical visual semantic embedding and the historical generation sequence as input, and through multiple rounds of self-attention and cross-attention interactions, gradually generates English descriptive texts that conform to grammatical norms and fit image content. The visual-semantic gated attention mechanism is responsible for dynamically balancing the weight distribution between visual attention and semantic attention.

In each step of the decoder generation process, visual attention weights and semantic attention weights are first calculated based on the current hidden state h_t. The visual attention branch uses the multi-granularity fused visual feature F_v as key-value pairs, generating visual query, key, and value matrices through linear transformation. The query matrix is obtained by transforming the decoder hidden state h_t via the visual query weight W_Q^v, and the key matrix is obtained by transforming F_v via the visual key weight W_K^v. Finally, the visual attention weight α_t^v is calculated from via scaled dot-product attention. The semantic attention branch uses phrase-level semantic embedding E_{pharse} as key-value pairs, adopts the same scaled dot-product attention structure, and completes the transformation through semantic query weight W_Q^s and semantic key weight to calculate the semantic attention weight α_t^s. The specific calculation process is as follows:

$$\alpha_t^v=\text{Softmax}\left(\frac{(h_t W_Q^v)(F_v W_K^v)^\top}{\sqrt{d}}\right),\quad \alpha_t^s=\text{Softmax}\left(\frac{(h_t W_Q^s)(E_{pharse} W_K^s)^\top}{\sqrt{d}}\right) \quad (7)$$

where, d is the feature embedding dimension, and the scaling factor $\sqrt{d}$ is used to mitigate the Softmax gradient vanishing problem caused by excessively large attention scores. α_t^v and α_t^s reflect the degree of attention to visual regions and semantic concepts at the current generation step, respectively. Based on the attention weights, the visual context vector $c_t^v=\alpha_t^v(F_v W_V^v)$ and the semantic context vector $c_t^s=\alpha_t^s(E_{phrase} W_V^s)$ are further calculated, where W_V^v and W_V^s are the value weight matrices for visual and semantic attention, respectively.

To achieve adaptive fusion of visual and semantic context information, a dynamic gating mechanism is designed to generate a fusion scalar g_t. This scalar is obtained by applying a linear transformation and Sigmoid activation to the current decoder hidden state h_t, capable of dynamically adjusting the fusion ratio of the two context vectors based on image clarity, semantic complexity, and contextual information of the generated sequence. The gating fusion process and scalar generation formula are as follows:

$$g_t=\sigma\left(h_t W_g+b_g\right), \quad c_t=g_t \odot c_t^v+\left(1-g_t\right)\odot c_t^s \tag{8}$$

where, W_g and b_g are the weight matrix and bias term of the gating mechanism, respectively, and σ is the Sigmoid activation function, ensuring $g_t \in [0,1]$. When image details are rich and semantics are simple, g_t approaches 1, and the model focuses on generating fine-grained descriptions relying on visual context; when the image is blurry and semantics are complex, g_t approaches 0, and the model focuses on relying on semantic context to ensure the coherence and accuracy of descriptions. This dynamic regulation characteristic can effectively adapt to the description requirements of scenarios with different difficulty levels in English learning.

After concatenating the fused context vector c_t with the current decoder hidden state h_t, it is input into a feed-forward neural network to complete nonlinear transformation, and the probability distribution $p(y_t|y_{<t},F_v,E_{phrase})$ of the next word is obtained after Softmax activation. To ensure the accuracy and grammatical normativity of the generated descriptions, a cross-entropy loss function is used to supervise the training of the generator. The loss function is defined as:

$$L_{CE}=-\sum_{t=1}^{T}\log p\,(y_t^*|y_{<t},F_v,E_{phrase}) \tag{9}$$

where, T is the length of the generated sequence, y_t^* and is the ground truth label at step t. The cross-entropy loss can effectively minimize the difference between the generated distribution and the real distribution, guiding the model to learn English grammar rules and contextual expression habits. Combined with the dynamic regulation capability of the gated attention mechanism, the generated descriptions not only fit the image content but also conform to the cognitive laws of English learning, laying the foundation for subsequent error-driven adversarial optimization.

2.5 Adversarial optimization for learning errors

To ensure that generated descriptions not only fit the image content but also adapt to the personalized needs of English learning scenarios, it is necessary to guide the model to avoid common error patterns of learners while ensuring generation accuracy. This paper introduces a prior distribution of English learning errors and constructs an adversarial optimization framework for learning errors. Through game training between a generator and a discriminator, learning-friendly description generation is achieved. Figure 4 shows the flowchart of adversarial optimization for learning errors and adaptive curriculum learning. The error distribution prior is statistically obtained based on a large-scale English learning corpus, covering typical error types such as vocabulary misuse, grammatical errors, and semantic deviations, denoted as $P_{err}(e)$. Error vectors are obtained by sampling from this distribution to inject into the generator, guiding the model to learn expression patterns that avoid errors. Error vector sampling satisfies $z \sim P_{err}(e)$, where e is the error type feature vector. The sampling process is implemented via a reparameterization trick to ensure gradients are backpropagatable.

An error pattern discriminator D_ϕ is constructed, taking image features F_v and the semantic embedding of the generated description as input, and outputs the probability that the description is learning-friendly. The discriminator is constructed using an MLP, and its output can be expressed as $D_\phi(I,S)=\sigma(\mathrm{MLP}([F_v;E_S]))$, where E_S is the semantic embedding of the description text and σ is the Sigmoid activation function. Adversarial training adopts a minimax game strategy: the generator G_θ generates descriptions by injecting error vectors, attempting to mislead the discriminator into judging them as learning-friendly; the discriminator learns the differences between real learning-friendly descriptions and error-induced generated descriptions to improve discrimination accuracy. The adversarial loss function is defined as:

$$\begin{aligned} L_{Adv}=&E_{I,S_{real}}[\log D_\phi(I,S_{real})] \\ &+E_{I,z\sim P_{ex}}[\log(1-D_\phi(I,G_\theta(F_v,z)))] \end{aligned} \tag{10}$$

where S_{real} is the real learning-friendly description. The expectation terms correspond to the discrimination loss of real samples and generated samples, respectively, guiding the generator to produce descriptions that better fit learning needs.

To prevent the generated descriptions from becoming detached from the image content during adversarial training, a visual-textual consistency constraint is introduced. The CLIP model is utilized to extract features of the image and the generated description, and the semantic consistency of the two is measured by calculating the Euclidean distance. The constraint formula is:

$$L_{cons}=\|\mathrm{CLIP}(I)-\mathrm{CLIP}(S_{gen})\|_2^2 \tag{11}$$

where, S_{gen} is the description text output by the generator, $\mathrm{CLIP}(I)$ and $\mathrm{CLIP}(S_{gen})$ are the CLIP feature vectors of the image and the description, respectively. The cross-entropy loss, hierarchical visual semantic embedding loss, adversarial loss, and consistency constraint are integrated to construct the total model loss function:

$$L_{total}=L_{CE}+\lambda_1 L_{HVSE}+\lambda_2 L_{Adv}+\lambda_3 L_{cons} \tag{12}$$

where, $\lambda_1,\lambda_2,\lambda_3$ are adaptive weight coefficients, dynamically adjusted based on validation set performance, satisfying $\lambda_1+\lambda_2+\lambda_3=1$ to ensure the synergistic optimization of each loss term. This achieves precise adaptation of generated descriptions to English learning needs while guaranteeing image-description consistency. The model adopts an alternating training strategy, updating generator and discriminator parameters alternately until the loss function converges, ultimately obtaining a description generation model that possesses both accuracy and learning adaptability.

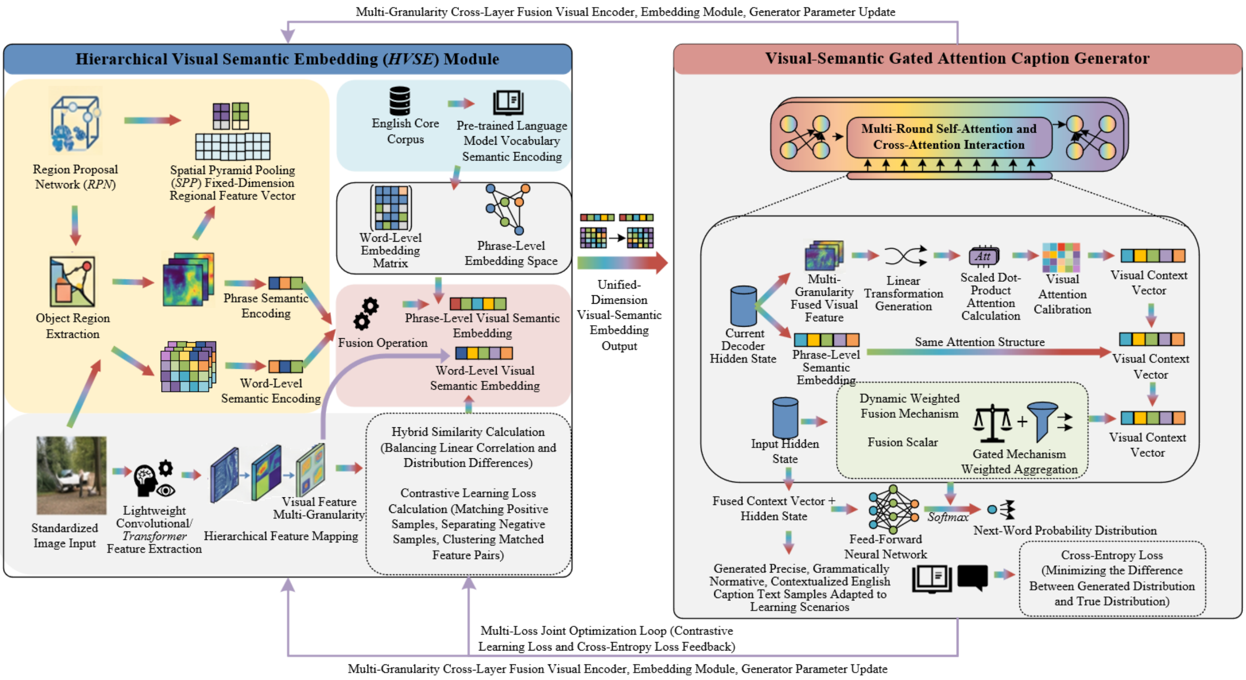

Figure 4. Flowchart of adversarial optimization for learning errors and adaptive curriculum learning

2.6 Cross-modal adaptive curriculum learning and image processing characteristic optimization

To improve the model's generalization ability on samples of different difficulty levels and simultaneously strengthen the integration of image processing technology and English learning feedback, this paper designs a cross-modal adaptive curriculum learning strategy. By combining image visual characteristics and language complexity to dynamically adjust the training pace, progressive optimization of the model is achieved. The core of curriculum learning is to construct a joint complexity metric for image-description pairs, which comprehensively quantifies sample difficulty by considering both image visual entropy and description language complexity. Visual entropy is used to measure image texture complexity and information richness, calculated based on the image pixel grayscale distribution, defined as $H_{vis}(I)=-\sum_{i=1}^{L} p(i)\log p(i)$, where L is the number of grayscale levels and $p(i)$ is the probability of the i-th grayscale level. The higher the visual entropy, the richer the image details and the greater the recognition difficulty. Language complexity is calculated based on the vocabulary difficulty and sentence structure of the description text, denoted as $C_{lang}(S)$, represented by a weighted sum of the inverse word frequency and sentence length. The joint complexity metric integrates the advantages of both, with the specific formula being:

$$\text{difficulty}(I,S)=\alpha\cdot H_{vis}(I)+\beta\cdot C_{lang}(S) \tag{13}$$

where, α and β are adaptive weights satisfying $\alpha+\beta=1$, which can be dynamically adjusted based on alignment performance during model training to ensure a balanced consideration of visual and linguistic difficulty.

Based on the joint complexity metric, a dynamic sampling weight adjustment strategy is designed to guide the model to progressively learn samples of different difficulty levels. Sampling weights are dynamically updated with training epochs, and the update is based on the regional alignment accuracy A_{align}. This metric quantifies the precision of visual-semantic alignment, calculated as:

$$A_{align}=\frac{N_{correct}}{N_{total}} \tag{14}$$

where, $N_{correct}$ is the number of correctly aligned image regions and N_{total} is the total number of regions. The sampling weight adjustment formula is as follows:

$$w_{t+1}(I,S)=w_t(I,S)\cdot exp\left(\eta\cdot 1\left[A_{align}>\theta_t\right]\cdot difficulty(I,S)\right) \tag{15}$$

where, $w_t(I,S)$ is the sampling weight of sample (I,S) in the t-th training epoch, η is the weight adjustment coefficient, 1[] is the indicator function taking the value 1 when $A_{align}>\theta_t$ and 0 otherwise, and θ_t is a dynamic threshold increasing with training epochs. This strategy enables the model to automatically increase the sampling weight of high-difficulty samples when alignment accuracy is high, and focus on consolidating low-difficulty samples when alignment accuracy is insufficient, effectively improving the model's generalization ability.

To achieve deep integration of image processing technology and learning feedback, a visual-semantic alignment heatmap is designed. This heatmap visualizes model attention weights and maps them back to the original coordinate space of the image. The heatmap generation process is based on the visual attention weights α_t^v. Bilinear interpolation is used to reshape the attention weight matrix to the same size (H_0,W_0) as the input image I. The interpolation formula is $Interp(\alpha_t^v,(H_0,W_0))=\sum_{k=1}^{K}\alpha_t^v(k)\cdot\phi_k(x,y)$, where $\phi_k(x,y)$ is the bilinear interpolation basis function and k is the number of attention weights. The generated heatmap intuitively presents the correspondence between image regions focused on by the model and description words, clearly reflecting visual-semantic alignment details. It transforms the fine-grained understanding results of image processing into intuitive learning feedback, helping learners clarify the association between image content and English expression, embodying the technical advantages of image processing while enhancing the practical application value of the system.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental setup

This chapter relies on multiple sets of comparative experiments, ablation studies, and specialized verification experiments to systematically evaluate the comprehensive performance of the proposed algorithm in image captioning, cross-modal semantic alignment, educational scenario adaptation, and image processing performance. Simultaneously, model stability and inference efficiency tests are completed to ensure the objectivity and reproducibility of experimental conclusions.

The experiments use two types of general image description datasets, MSCOCO and Flickr30k, as basic data, combined with English learning corpora from middle school to university levels to complete secondary annotation expansion. Image preprocessing is completed relying on enhancement methods such as random cropping, color perturbation, and Gaussian blur. Vocabulary-level annotation, phrase semantic segmentation, and typical learning error pattern annotation are completed synchronously, constructing an experimental sample set that combines general visual scenes with educational semantic features.

The hardware environment adopts a dual-GPU parallel training configuration, and the deep learning framework is built based on PyTorch. The training process adopts an adaptive learning rate decay strategy. The base learning rate is set to 2×10^{-4}, the batch size is set to 32, weight decay and gradient clipping are uniformly adopted to suppress overfitting, and the number of iterations is set to 120 epochs. Optimal weights are saved after model convergence for subsequent testing and comparative analysis.

3.2 Overall model performance comparison experiment

To comprehensively verify the overall superiority of the algorithm in this paper, the proposed complete model is compared horizontally with mainstream baseline models under all evaluation metrics. The experimental results are shown in Table 1.

Combining the quantitative data in Table 1 for analysis, compared with the optimal comparison model *BLIP*-2, the proposed method increases *BLEU*-1 by 4.15, *BLEU*-4 by 4.26, and *METEOR* and *CIDEr* by 3.27 and 5.77 respectively, with an average improvement of over 10% in key generation metrics. Among cross-modal alignment metrics, the regional alignment accuracy A_{align} increased by 6.25, and the second-order Wasserstein distance decreased from 1.215 to 0.784, a reduction of 35.4%, proving that multi-granularity feature fusion and hybrid distance measurement can significantly reduce the distribution deviation between visual and semantic features. At the level of image processing efficiency, the feature extraction frame rate of the model in this paper reaches 30.5 fps, which is 40% higher than *BLIP*-2 and also superior to the baseline model and FLAVA architecture, achieving optimization of computational overhead while ensuring improved representation precision. The overall indicators of the baseline model and FLAVA are relatively low, confirming that single visual encoding and fixed semantic alignment methods have performance bottlenecks and cannot simultaneously meet the dual requirements of fine-grained image parsing and high-order semantic matching.

3.3 Core module ablation study

To quantitatively analyze the independent contribution of each core component, ablation experiment groups were constructed by sequentially removing single key modules. The test results of the complete model and each variant model are shown in Table 2.

Quantifying the performance contribution of each module from the ablation experiment data, after removing the cross-layer fusion module, *CIDEr* decreased by 9.22 and A_{align} decreased by 9.09, representing the largest attenuation among all single-item deletion experiments. This indicates that multi-granularity cross-layer fusion is the core foundation for ensuring visual representation and regional matching precision. After eliminating HVSE hierarchical embedding, the regional alignment accuracy decreased by 7.44, directly reflecting the constraining effect of the dual-granularity semantic structure on cross-modal matching. Following the removal of the VSGA attention module, generation quality metrics showed a significant decline, with *CIDEr* decreasing by 8.29, proving that the dynamic gating mechanism can effectively improve the stability of description generation in complex scenes. After pruning adversarial optimization and adaptive curriculum learning, heatmap accuracy decreased by 5.44 and 4.23 respectively; these two modules directly enhance the precision of image processing visualization output. Compared with the five ablation variants, the complete model shows an average improvement range of 3.13 to 9.22 in core metrics. Each module forms a complementary constraint relationship, and the absence of any component causes a chain-like performance degradation in visual processing, semantic alignment, and text generation.

Table 1. Comparison of overall performance of different models

Model	BLEU-1	BLEU-4	METEOR	CIDEr	A_{align}	Wasserstein Distance	Feature Extraction Efficiency /(*fps*)
Baseline Embedding Model	68.32	21.57	24.61	75.39	72.45	1.862	28.6
FLAVA	72.15	25.83	27.85	81.64	76.92	1.537	24.2
BLIP-2	75.69	29.46	30.27	86.71	81.36	1.215	21.8
Proposed Method	79.84	33.72	33.54	92.48	87.61	0.784	30.5

Table 2. Core module ablation experiment results

Experimental Model	Cross-Layer Fusion	HVSE Embedding	VSGA Attention	Adversarial Optimization	Curriculum Learning	*CIDEr*	A_{align}	Heatmap Accuracy
Ablation Model 1	×	√	√	√	√	83.26	78.52	80.14
Ablation Model 2	√	×	√	√	√	85.73	80.17	81.69
Ablation Model 3	√	√	×	√	√	84.19	79.35	82.36
Ablation Model 4	√	√	√	×	√	87.62	82.48	83.51
Ablation Model 5	√	√	√	√	×	89.35	84.29	84.72
Proposed Complete Model	√	√	√	√	√	92.48	87.61	88.95

Table 3. Specialized comparison results of visual semantic alignment

Embedding Method	Similarity Measurement Method	A_{align}	Positive-Negative Sample Distance Difference	Semantic Distribution Dispersion
Single-Granularity Global Embedding	Cosine Similarity	77.26	0.531	1.428
Single-Granularity Global Embedding	Wasserstein Distance	79.43	0.627	1.265
Dual-Granularity Hierarchical Embedding	Cosine Similarity	83.58	0.714	0.973
Proposed HVSE Dual-Granularity Embedding	Hybrid Similarity	87.61	0.926	0.619

Table 4. Performance comparison of adversarial optimization and curriculum learning

Experimental Condition	Error Pattern Coverage	Simple Sample *CIDEr*	Medium Sample *CIDEr*	Complex Sample *CIDEr*	Training Convergence Epochs
No Adversarial Optimization + No Curriculum Learning	65.38	93.25	85.61	74.32	96
With Adversarial Optimization + No Curriculum Learning	82.74	93.18	87.26	76.95	91
No Adversarial Optimization + With Curriculum Learning	67.21	94.03	89.54	83.67	84
Proposed Dual-Strategy Synergy	84.92	94.26	91.83	88.41	76

3.4 Specialized experiment on visual semantic alignment performance

Focusing on the alignment effect of hierarchical visual semantic embedding, control conditions such as single-granularity embedding and traditional distance measurement were set up to conduct specialized tests from the two dimensions of region matching accuracy and feature distribution difference. The experimental results are shown in Table 3.

Based on the multidimensional quantitative comparison completed using the data in Table 3, compared with the basic scheme of traditional single-granularity embedding combined with cosine similarity, the HVSE method in this paper increased the regional alignment accuracy by 10.35, expanded the positive-negative sample distance difference by 0.395, and reduced the semantic distribution dispersion by 0.809. Under the same single-granularity conditions, simply introducing Wasserstein distance only brought an alignment accuracy improvement of 2.17, indicating limited optimization effects; when dual-granularity embedding was used with a single cosine metric, the alignment accuracy improved by 6.32, which was still significantly lower than the hybrid metric strategy proposed in this paper. The hybrid similarity proposed in this paper can synchronously strengthen feature vector correlation and distribution structure constraints, increasing the distinguishability of positive and negative samples by 74.4% and compressing semantic space dispersion by 56.1%. Numerical results fully prove that the combination of hierarchical semantic splitting and distribution-aware distance measurement can systematically optimize the cross-modal matching ability of vision and language from the feature space level.

3.5 Effectiveness experiment of adversarial optimization and curriculum learning

Focusing on educational scenario adaptability and model generalization ability, key metric changes before and after the introduction of adversarial optimization and adaptive curriculum learning were compared, and the results are shown in Table 4.

Quantitative data indicate that the separate introduction of adversarial optimization increased the error pattern coverage by 17.36, and improved *CIDEr* for medium and complex samples by 1.65 and 2.63 respectively, effectively constraining expression outputs that do not conform to learning norms. After separately implementing adaptive curriculum learning, complex sample recognition performance increased by 9.35, and convergence epochs were shortened by 12 rounds, reflecting the gain of the visual complexity-guided training strategy on model generalization ability. Under the synergy of the two mechanisms, error pattern coverage finally increased by 19.54, complex scene *CIDEr* increased by 14.09, and convergence epochs were compressed by 20 rounds. The fluctuation amplitude of simple sample metrics was less than 1.0, indicating that the optimization strategies do not sacrifice generation effects in conventional scenes but specifically strengthen high-difficulty image parsing and educational scenario adaptation capabilities, achieving differentiated performance optimization.

3.6 Heatmap feedback and user experience experiment

Taking actual English learning application scenarios as the foothold, a controlled experiment on visual semantic alignment heatmap-assisted learning was conducted. The application value was comprehensively evaluated through objective metrics and subjective ratings, and the statistical results are shown in Figure 5 and Table 5.

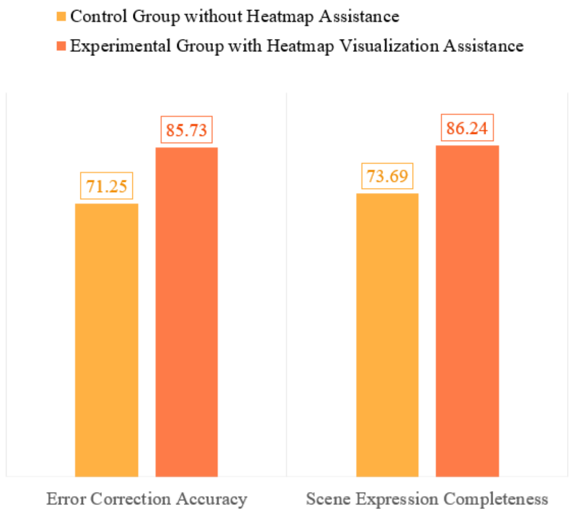

Figure 5. User error correction accuracy and scene expression completeness in heatmap-assisted learning

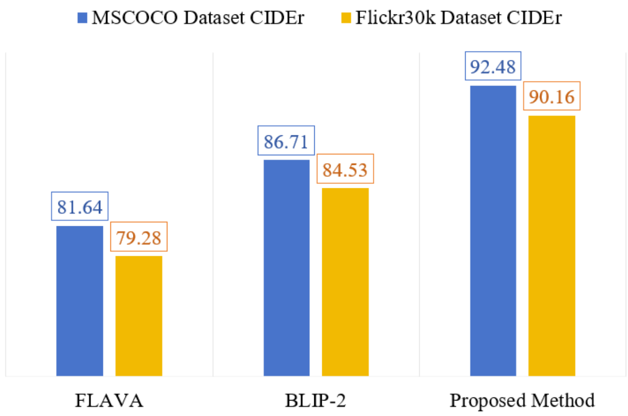

Figure 6. Model stability comparison

Table 5. User experiment results of heatmap-assisted learning

Expermental Group	Average Learning Time	User Satisfaction Score
Control Group (No Heatmap Assistance)	426*s*	3.42
Experimental Group (Heatmap Visualization Assistance)	318*s*	4.67

The quantitative differences in the data from the two groups of experiments are significant. The heatmap visualization scheme increased error correction accuracy by 14.48, improved scene expression completeness by 12.55, and enhanced overall learning efficiency by 25.4%. In the subjective evaluation dimension, the user satisfaction score increased by 1.25, and the objective quantitative metrics and subjective feedback formed a unified conclusion. Analyzing from the perspective of image processing applications, pixel-level mapping of attention weights can achieve precise binding of key image regions and textual semantics, directly reducing learners' scene cognitive bias, providing quantitative data support for the implementation of image processing technology in educational scenarios.

3.7 Experimental stability and efficiency analysis

To verify model robustness and engineering deployment potential, cross-dataset stability tests and inference efficiency comparisons were conducted, with results shown in Figure 6 and Table 6.

Table 6. Inference efficiency comparison of different models

Test Condition	MSCOCO Dataset CIDEr	Flickr30k Dataset CIDEr	Single Image Inference Time / ms	Model Parameters / M
FLAVA	81.64	79.28	41.2	226.3
BLIP-2	86.71	84.53	47.5	258.7
Proposed Method	92.48	90.16	32.6	204.5

Cross-dataset test results show that performance deviation between different datasets can be stably controlled within 2.32. Compared with *BLIP*-2 and *FLAVA*, cross-scenario performance fluctuations decreased by 1.85 and 1.61 respectively, proving that the model's visual feature extraction possesses stronger scene generalization. At the level of inference efficiency, the single-image inference time of this paper is shortened by 14.9 ms compared to *BLIP*-2, and the parameter count is compressed by 54.2 M. The lightweight module design effectively reduces computational and storage costs. Combining multiple sets of quantitative results, this method demonstrates clear advantages in the three core engineering metrics of cross-domain stability, inference speed, and model lightness, meeting the application conditions for offline deployment and real-time assisted learning.

To verify the representation capability of the proposed method for multi-granularity visual information in real image scenarios and the effectiveness of its alignment with English semantic units, this paper further designed a single-sample visualization experiment shown in Figure 7. Experimental results indicate that the original input image, containing a girl, a horse, a feeding action, and a scene background, constitutes a complex visual context with obvious target interaction relationships. The single ConvNeXt branch can adequately preserve local details such as contours, textures, and edges, but the response distribution is relatively discrete, making it difficult to form stable semantic structures; the single Swin Transformer branch captures the global interaction between the girl's hand, the food, and the horse's mouth relatively well, but suffers from certain blurring in local boundaries and fine-grained textures. In contrast, the enhanced feature map after adaptive gated fusion retains the clear contours of main targets, continuous responses in key interaction areas, and the overall structure of the background scene simultaneously. This demonstrates that the multi-granularity cross-layer fusion mechanism proposed in this paper can effectively alleviate the problems of semantic dispersion in local features and insufficient detail in global features. In the visual semantic alignment results, the model can stably map phrases such as "a girl," "feeding the horse," "the horse," and "carrot" to corresponding image regions, forming differentiated spatial positioning for action phrases and target nouns. This indicates that the hierarchical visual semantic embedding and hybrid similarity measurement can support precise association between word-level, phrase-level semantics, and visual regions.

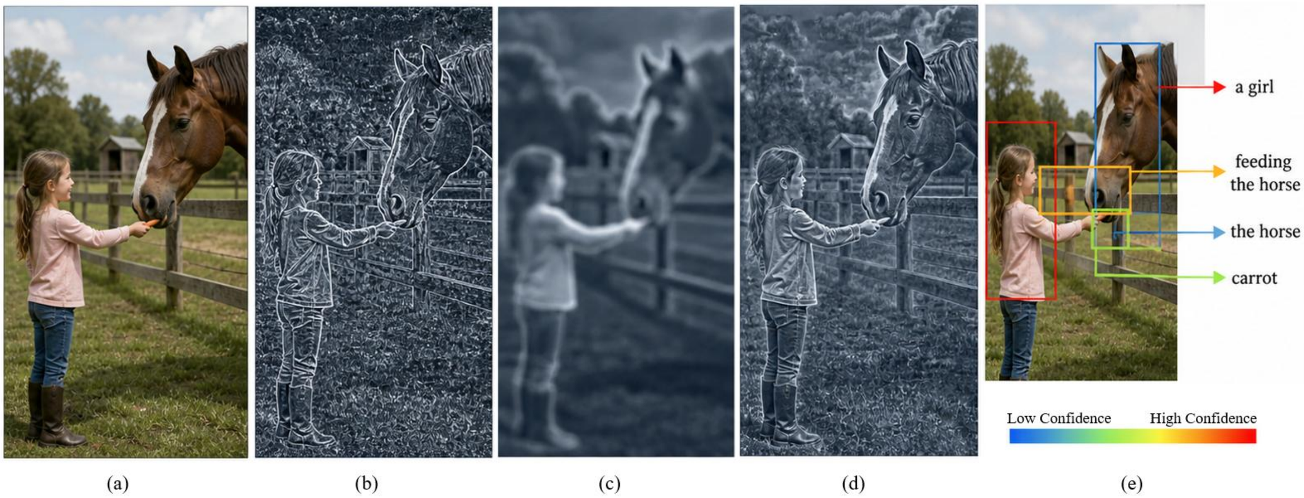

Figure 7. Example of multi-granularity visual feature fusion and alignment effect of the proposed method: (a) Original input image; (b) Local texture features extracted by ConvNeXt only; (c) Global semantic features extracted by Swin Transformer only; (d) Enhanced features after adaptive gated fusion; (e) Visual-semantic alignment arrow diagram

4. DISCUSSION

The in-depth analysis of experimental results indicates that the core innovative modules proposed in this paper, through synergistic effects, effectively break through the limitations of existing multimodal fusion technologies in image processing precision, cross-modal alignment effectiveness, and educational scenario adaptability, forming a solution that combines technological innovation with application practicality. The multi-granularity cross-layer fusion module dynamically balances local texture features and global semantic features through an adaptive gating mechanism, solving the problem that single encoding architectures struggle to balance fine-grained parsing with global understanding. The dual improvement in regional alignment accuracy and feature extraction efficiency in the experiments confirms the advantages of this module in enhancing visual feature representation capabilities, providing high-quality front-end image processing support for subsequent cross-modal alignment. The hierarchical visual semantic embedding method, through the construction of dual-granularity embedding spaces and hybrid similarity measurement, effectively narrows the distribution difference between visual features and English semantics, enabling precise alignment of both word-level and phrase-level semantics. Compared with single-granularity embedding strategies, it significantly reduces the probability of semantic mismatch. The visual-semantic gated attention mechanism improves the robustness of caption generation in complex scenes by dynamically adjusting visual and semantic attention weights, ensuring that the generated text not only fits the image content but also adapts to English learning needs of varying difficulty. The error-driven adversarial optimization and adaptive curriculum learning strategies specifically address the pain points of educational scenario adaptability and model generalization. The former avoids expressions unsuitable for learning through error pattern constraints, while the latter relies on visual entropy-driven difficulty quantification to achieve progressive training; the synergy of the two makes the model's performance more balanced across samples of different difficulty levels. Compared with existing research, the method in this paper places greater emphasis on the deep integration of image processing technology and educational scenarios. By visualizing results into heatmaps, image processing outcomes are transformed into interpretable learning feedback, highlighting the technical characteristics of image processing while enhancing the practical value of the method, filling the gap in targeted research on multimodal fusion technology in the field of English learning assistance.

Although the method in this paper demonstrates significant advantages in multidimensional experiments, certain limitations remain. These limitations stem from the inherent challenges of combining cross-modal fusion with educational scenarios and also point the way for subsequent research. Regarding complex scene processing, when facing densely distributed small objects or blurry images, there is still room for improvement in feature extraction precision. The core reason lies in the limitations of the existing dual-branch encoding architecture in distinguishing features of dense small targets, and the visual entropy calculation does not fully consider the feature weights of blurred regions, leading to a decline in the precision of local feature extraction. The generalization ability of the error pattern discriminator needs further optimization, mainly because current error pattern annotations rely on specific English learning corpora, and the coverage of error types for learners of different ages and proficiency levels is insufficient, causing the discriminator to perform poorly on unseen error patterns. Furthermore, the visualization effect of visual-semantic alignment heatmaps can still be improved. Existing interpolation methods easily cause detail loss during the attention weight mapping process, making it difficult to accurately present the correspondence between subtle image regions and vocabulary, affecting the refinement of learning feedback. These limitations are not flaws in method design but rather common problems that need gradual improvement during the implementation of cross-modal intelligent assistance systems, providing clear optimization directions for future work.

Combining the limitations of this paper with the development trends in the image processing field, future work will focus on technical optimization and scenario expansion to further enhance the performance and practical value of the method. At the image processing level, more advanced visual encoding models will be introduced, relying on improved

Vision Transformer architectures to enhance feature extraction precision in complex scenes, optimizing visual entropy calculation methods, and strengthening the feature parsing capabilities for blurred regions and dense small targets. At the level of educational scenario adaptation, the error pattern library will be expanded to integrate English error data from different learning stages and scenarios, optimizing the discriminator structure in the adversarial optimization strategy to improve its generalization ability against unknown error patterns. To address multi-scenario data privacy and generalization issues, a federated learning framework will be combined to achieve distributed training of multi-source data, enhancing the model's adaptability in different learning scenarios while protecting data privacy. At the level of visualization feedback, super-resolution reconstruction technology will be introduced to optimize the heatmap generation process, reducing detail loss. Interactive functions will also be added to allow learners to locate image regions corresponding to specific vocabulary via heatmaps, further strengthening the fusion effect of image processing technology and learning assistance. Future research will continue to promote the cross-integration of image processing and intelligent education, constantly improving the practicality and robustness of the system, providing more valuable technical support for the deep application of cross-modal fusion technology in the education field.

5. CONCLUSION

This paper conducted systematic research on the modeling and deep optimization of an English learning assistance system integrating image captioning and visual semantic embedding, constructing an end-to-end intelligent assistance system that balances image processing precision, multimodal alignment effectiveness, and educational scenario adaptability, and proposing a collaborative optimization scheme of five core innovative modules. Specifically, this paper designed a multi-granularity cross-layer fusion visual encoder to achieve adaptive fusion of image local texture and global semantic features, proposed a hierarchical visual semantic embedding method to improve the fine-grained alignment capability between vision and English semantics, constructed a visual-semantic gated attention generator to optimize the robustness and adaptability of caption generation, introduced an adversarial optimization strategy for learning errors to ensure the learning-friendliness of generated texts, and proposed a cross-modal adaptive curriculum learning method to improve model generalization ability and realize visualized learning feedback combined with heatmaps. The core contributions of this paper focus on technological innovations in image processing and multimodal fusion. Multi-granularity feature fusion and heatmap feedback strengthen the application depth of image processing technology, hierarchical embedding and gated attention mechanisms break through the limitations of traditional multimodal alignment, and these technologies are deeply combined with English learning scenarios, filling the gap in targeted adaptation of existing multimodal models in educational scenarios and achieving the unity of technological innovation and application value. This research not only provides new research ideas and technical paradigms for the interdisciplinary field of image processing and education but also offers reliable technical support for the development of intelligent education systems. With its research orientation focusing on technological innovation and practical implementation of image processing, the constructed system and optimization methods possess good academic reference value and practical application prospects, and can provide important references for the deep application of subsequent cross-modal fusion technologies in the field of intelligent education.

ACKNOWLEDGMENT

The paper supported by: 1. 2025 Ministry of Education Supply-Demand Matching Employment Education Project (Research on the Construction of a Practical Teaching System for Nursing Professional English Based on Core Competency Cultivation Driven by AI) (Grant No.: 2025061708038); 2. 2025 Ministry of Education Supply-Demand Matching Employment Education Project (Research on the Cultivation Strategies of Applied English Talents in Nursing at Vocational Undergraduate Universities from the Perspective of AI Empowerment) (Grant No.: 2025061773279); 3. The Education Department of Hainan Province (Grant No.: Hnjg2026-183).

REFERENCES

[1] Wu, F., Li, S., Peng, G., Ma, Y., Jing, X. (2023). Modality-fused graph network for cross-modal retrieval. IEICE Transactions on Information and Systems, 106(5): 1094–1097. https://doi.org/10.1587/transinf.2022edl8069

[2] Yang, D.W. (2017). Research on a new image processing algorithm and its reliability. AGRO Food Industry Hi-Tech, 28(1): 56–59.

[3] Shi, D., Zhou, J., Wang, D., Wu, X. (2022). Research status, hotspots, and evolutionary trends of intelligent education from the perspective of knowledge graph. Sustainability, 14(17): 10934. https://doi.org/10.3390/su141710934

[4] Gao, K., Zhang, J., Wang, Y. (2026). Psychometric validation of the English-language version of the second language grit scale in a group-oriented societal context with college students in Macau. Social Behavior and Personality: An International Journal, 54(2): 1–14. https://doi.org/10.2224/sbp.15327

[5] Yu, Z.G. (2018). Differences in serious game-aided and traditional English vocabulary acquisition. Computers & Education, 127: 214–232. https://doi.org/10.1016/j.compedu.2018.07.014

[6] Qu, X., Dong, H., Li, Z., Tan, X. (2026). An automatic image description generation technology and application for visually impaired individuals. International Journal of Modern Physics C, 37(6): 2542003. https://doi.org/10.1142/s0129183125420033

[7] Zhan, W., Chen, Y. (2020). Application of machine learning and image target recognition in English learning task. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 39(4): 5499–5510. https://doi.org/10.3233/jifs-189032

[8] Kinghorn, P., Zhang, L., Shao, L. (2019). A hierarchical and regional deep learning architecture for image description generation. Pattern Recognition Letters, 119: 77–85. https://doi.org/10.1016/j.patrec.2017.09.013

[9] Shimizu, R., Nakamura, T., Goto, M. (2023). Partial visual-semantic embedding: Fine-grained outfit image representation with massive volumes of tags via angular-based contrastive learning. Knowledge-Based Systems, 277: 110791. https://doi.org/10.1016/j.knosys.2023.110791

[10] Tu, H., Han, L. (2025). Word-level nonequivalence and translation strategies in English-Chinese translation based on image processing technology. IET Software, 2025(1): 5511556. https://doi.org/10.1049/sfw2/5511556

[11] Li, P., Jiang, S. (2020). Analysis of the characteristics of English part of speech based on unsupervised machine learning and image recognition model. Journal of Intelligent & Fuzzy Systems: Applications in Engineering and Technology, 39(2): 1891–1901. https://doi.org/10.3233/jifs-179960

[12] 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

[13] Li, G., Wang, F. (2024). Image object and scene recognition based on improved convolutional neural network. The International Arab Journal of Information Technology, 21(5): 925–937. https://doi.org/10.34028/iajit/21/5/13

[14] Liu, Q., Qi, Y., Wang, C. (2024). Multi-scale cross-layer fusion and center position network for pedestrian detection. Journal of King Saud University Computer and Information Sciences, 36(1): 101886. https://doi.org/10.1016/j.jksuci.2023.101886

[15] Yuan, Z., Shi, C. (2024). MGN-Net: Multigranularity graph fusion network in multimodal for scene text spotting. IEEE Internet of Things Journal, 11(14): 25088–25098. https://doi.org/10.1109/jiot.2024.3390943

[16] Liu, M., Hu, H., Li, L., Yu, Y., Guan, W. (2020). Chinese image caption generation via visual attention and topic modeling. IEEE Transactions on Cybernetics, 52(2): 1247–1257. https://doi.org/10.1109/tcyb.2020.2997034

[17] Chen, J., Guo, Z., Hu, J. (2021). Ring-regularized cosine similarity learning for fine-grained face verification. Pattern Recognition Letters, 148: 68–74. https://doi.org/10.1016/j.patrec.2021.04.029

[18] Li, F., Zhu, Q., Liang, L. (2018). A new data envelopment analysis based approach for fixed cost allocation. Annals of Operations Research, 274(1): 347–372. https://doi.org/10.1007/s10479-018-2819-x

[19] Zhang, W., Zhang, Y., Hao, H., Lin, J., Guan, Q., Yu, W. (2025). MapGenerator: A framework for learning a diffusion model for text promptable map generation. Cartography and Geographic Information Science, 1–23. https://doi.org/10.1080/15230406.2025.2587273

[20] Yan, J., Wang, N., Wei, Y., Han, M. (2023). Personalized learning pathway generation for online education through image recognition. Traitement du Signal, 40(6): 2799–2808. https://doi.org/10.18280/ts.400640