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

Traitement du Signal
Vol. 43, No. 3, June, 2026, pp. 1093-1104
Journal homepage: http://iieta.org/journals/ts

Image Understanding and Information Extraction for English Education Scenarios via Multi-Scale Visual Feature Enhancement and Cross-Modal Semantic Constraints

Can Zhao

Department of Public Course Teaching, Hebei Chemical & Pharmaceutical College, Shijiazhuang 050026, China

Corresponding Author Email: ZhaoCan202310@163.com

https://doi.org/10.18280/ts.430302

Received: 5 January 2026
Revised: 22 May 2026
Accepted: 30 May 2026
Available online: 30 June 2026

Keywords:
multi-scale feature fusion, dynamic gating mechanism, cross-modal semantic alignment, educational knowledge graph, dual-task collaborative optimization, English education image understanding

ABSTRACT

In the digital transformation of smart education, intelligent semantic understanding and structured information extraction of English teaching imagery are essential for automated assignment grading, resource retrieval, and personalized learning. However, generic image understanding methods struggle with domain-specific challenges, including large object-scale variations, weak alignment between visual and pedagogical semantics, and inconsistent optimization objectives across educational tasks. To address these limitations, an integrated framework for image understanding and information extraction specific to English education scenarios was proposed by incorporating multi-scale visual feature enhancement, cross-modal semantic alignment, and multi-task collaborative optimization mechanisms. A global scene semantics-guided dynamic gating fusion strategy was employed to adaptively adjust the fusion weights of visual features across different hierarchical levels, thereby mitigating representational deficiencies caused by the substantial scale disparity between panoramic scene layouts and minute textual symbols. A hierarchical English education knowledge graph was constructed to establish a dual-dimensional cross-modal contrastive constraint system, enabling precise alignment between global scene semantics and localized knowledge-point regions and facilitating the convergence of visual representations toward the specialized pedagogical semantic space. By leveraging the topological priors of scene-knowledge point associations embedded within the knowledge graph, a multi-task collaborative distillation strategy coupled with feature consistency regularization was designed. This established a closed-loop mechanism for the bidirectional optimization of visual perception and instructional objectives, effectively resolving feature conflicts and semantic deviations inherent in multi-task learning. Systematic experiments were conducted on a self-constructed English education image dataset. The results demonstrated that the proposed method significantly outperformed state-of-the-art baseline approaches in both scene classification and knowledge-point recognition tasks, exhibiting superior scale robustness and model interpretability. This study provides a specialized technical paradigm for vision-based intelligent analysis within the educational domain, offering effective support for large-scale intelligent applications in smart English education scenarios.

1. INTRODUCTION

The continued deepening of the digital transformation in education has driven the rapid proliferation of smart English teaching models [1-3], with the volume of image-based instructional resources—centered on textbook illustrations, classroom blackboard content, handwritten assignments, and teaching slides—experiencing sustained growth. The automated semantic parsing and structured information extraction from multi-source English teaching images constitute a fundamental enabling technology for the deployment of smart education applications, including intelligent educational resource retrieval, personalized instructional content delivery, and automated academic assessment [4-6]. Imagery from English education contexts differs fundamentally from generic natural images [7-9], exhibiting strong domain specificity. The visual elements within such imagery span an extremely wide range, simultaneously carrying both macroscopic teaching scenarios and micro-level textual symbol-based knowledge points. Moreover, all visual content is deeply coupled with standardized pedagogical knowledge systems, necessitating that the information parsing process strictly adhere to subject-specific teaching logic and cognitive principles. General-purpose image understanding algorithms, designed primarily for natural scenes, lack domain-adaptive mechanisms and are consequently unable to simultaneously accommodate the scale diversity and semantic specialization characteristic of educational imagery [10-12], thus failing to meet the stringent requirements of fine-grained instructional information extraction. Dedicated research on image understanding techniques tailored to the vertical domain of English education

is therefore essential. Such research not only represents a significant extension of domain adaptation studies within computer vision, contributing to the theoretical refinement of domain-adaptive cross-modal perception, but also serves to consolidate the technological foundation for the digital and intelligent advancement of smart English education, offering substantial theoretical novelty and practical industrial value.

At present, image understanding and cross-modal learning techniques have been preliminarily applied to educational visual analysis scenarios. However, critical deficiencies remain within the existing technical framework, rendering it incapable of meeting the fine-grained parsing demands of English teaching imagery. With respect to multi-scale feature representation [13-15], mainstream feature fusion methodologies rely on fixed topological structures and static fusion weights for cross-scale feature interaction, lacking scene-adaptive adjustment capabilities. Given the highly dynamic scale distribution of English teaching imagery—where primary and secondary visual targets differ markedly across instructional scenarios—static fusion mechanisms are unable to selectively enhance feature representations of fine-grained knowledge points, such as small textual characters and word symbols. This inherently leads to weakened small-target features, missed detections, and semantic deviations, resulting in demonstrably deficient scale-adaptation performance. Regarding cross-modal semantic association [16, 17], generic image-text contrastive learning models, trained on general-purpose commonsense datasets without the incorporation of education-domain-specific knowledge constraints, can only achieve coarse-grained scene-topic matching via visual encoding, falling short of identifying the fine-grained teaching knowledge points corresponding to localized image regions. A pronounced heterogeneous semantic gap persists between low-level visual perceptual signals and high-level specialized pedagogical semantics, preventing the model from achieving pixel-level semantic localization of knowledge points and thereby precluding genuine educationally meaningful semantic understanding. With respect to task optimization mechanisms [18-20], existing educational visual processing solutions predominantly adopt a single-task independent training paradigm, neglecting the intrinsic logical associations between scene types and core knowledge points within the pedagogical system. Multi-task joint learning processes are pervasively plagued by feature competition and semantic-logical conflicts. The task optimization objectives, being detached from practical instructional application requirements, fail to establish a closed-loop mechanism for the mutual optimization of visual perception capabilities and educational cognitive goals, thereby substantially constraining the practical deployment performance of the models.

To address the core bottlenecks of the existing technologies described above, an end-to-end integrated image understanding and information extraction framework tailored for English education scenarios is constructed, providing a full-chain specialized visual parsing solution. A holistic architecture featuring deep coupling among feature enhancement, semantic alignment, and task collaboration is established, transcending the performance limitations of traditional serial concatenation models. A scene-adaptive dynamic gating fusion mechanism is designed to resolve the representation challenge posed by drastically varying target scales in teaching imagery. A hierarchical cross-modal constraint system is introduced via the incorporation of an educational knowledge graph, bridging the heterogeneous semantic gap between visual perception and pedagogical semantics. A multi-task collaborative optimization strategy is developed based on knowledge topology priors, enabling bidirectional empowerment between perception tasks and educational objectives. Furthermore, a dedicated English education image dataset is constructed, and comprehensive performance and application advantages of the proposed methodology are thoroughly validated through multi-dimensional experiments.

The remainder of this study is organized as follows. The technical principles and implementation details of the proposed tripartite framework and its constituent core modules are elaborated in Chapter 2. Multiple sets of comparative experiments, ablation studies, and robustness evaluations are configured in Chapter 3, with which performance verification and result analysis are conducted. The intrinsic operational mechanisms of the methodology are examined in Chapter 4, where existing limitations are objectively analyzed and future research directions are outlined. The research outcomes are synthesized in Chapter 5, with which the core innovations and application values are summarized.

2. METHODOLOGY

2.1 Overall framework overview

An end-to-end integrated image understanding and information extraction framework tailored for English education scenarios is constructed, systematically incorporating the core functionalities of multi-scale visual feature enhancement, cross-modal semantic alignment, and multi-task collaborative optimization. The overall inference process of the framework follows a progressive optimization logic. Input English teaching images are first processed via an improved backbone network for multi-level foundational visual feature extraction. The acquired multi-scale features are subsequently fed into an adaptive gating fusion module for dynamic enhancement, thereby accommodating the inherent characteristic of substantial target-scale variation within teaching imagery. The scale-optimized visual features are then mapped into a unified semantic embedding space through a hierarchical cross-modal constraint mechanism, where dual-layer semantic alignment between global scenes and localized knowledge points is achieved by leveraging the domain prior knowledge embedded in the English education knowledge graph. Finally, teaching scene classification and knowledge-point recognition tasks are performed in parallel via the collaborative optimization module.

The proposed framework departs from the conventional paradigm of serial modular stacking in traditional vision models, enabling full-module feature sharing and gradient interplay through a unified semantic embedding space. During model training, the semantic alignment loss and multi-task supervision signals are back-propagated throughout the entire process to the feature extraction and feature fusion modules, establishing a closed-loop optimization mechanism characterized by positive reinforcement of visual features and corrective feedback from semantic constraints. Concurrently, the structured prior knowledge of the educational knowledge graph permeates the entire model pipeline. This prior knowledge can guide the dynamic weight allocation of multi-scale features, reinforcing the representational capacity of critical knowledge-point regions, while also regulating the

output logic of dual tasks and constraining the model's semantic reasoning process. Ultimately, a deep integration of domain knowledge and visual perception is achieved, ensuring the specialized parsing performance of the model within English education scenarios.

2.2 Scene-adaptive dynamic multi-scale gating fusion module

To effectively accommodate the extreme scale distribution disparities inherent in English education imagery, a scene-adaptive dynamic multi-scale gating fusion module is designed, in which the fusion weights of visual features across different hierarchical levels are dynamically regulated by the global semantic information of the scene, thereby achieving refined enhancement of multi-scale features. Foundational feature extraction is performed based on the ConvNeXt-V2 network, with targeted optimizations applied to the network architecture in accordance with the visual characteristics of educational imagery. The network architecture of the scene-adaptive dynamic multi-scale gating fusion module is illustrated in Figure 1. Three-channel color images with an input size of $H \times W \times 3$ are fed into the model, from which feature maps with four distinct receptive fields—corresponding to downsampling factors of 4×, 8×, 16×, and 32×—are obtained via the backbone network's downsampling operations. These feature maps correspond respectively to fine-grained textural details and coarse-grained global semantic information within the imagery. To strengthen the model's capacity for capturing fine-grained knowledge points such as minute textual characters and word symbols, the convolutional kernel size in the network's lower-level convolutional modules is adjusted to 5 × 5, thereby preserving detail perception capability while reducing computational overhead. Concurrently, the global attention structures in the higher network layers are removed and replaced by a pooled global embedding generation mechanism, further improving inference efficiency while maintaining the quality of global semantic representation. Finally, feature maps from all hierarchical levels are uniformly mapped to a 256-dimensional channel space via 1 × 1 convolution operators, eliminating dimensional discrepancies and providing a unified representational foundation for cross-scale feature fusion.

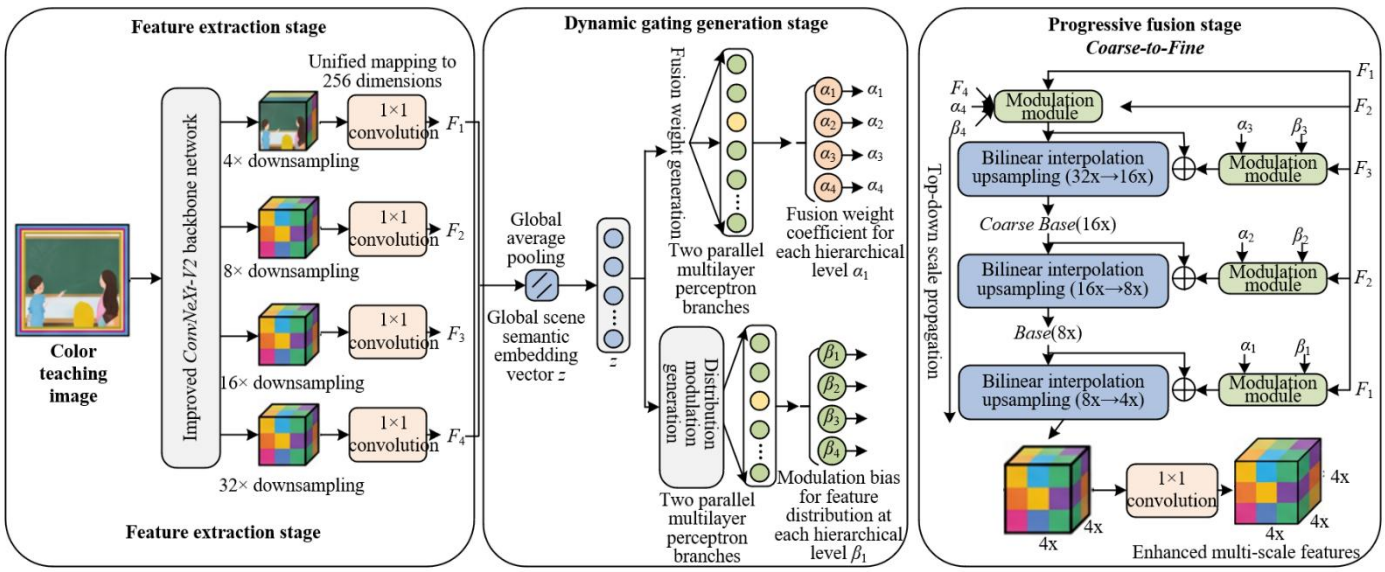

Figure 1. Network architecture of the scene-adaptive dynamic multi-scale gating fusion module

To achieve scene-adaptive dynamic regulation of feature fusion weights, a scene semantic embedding representation is constructed based on high-level global features, with which the adaptive weighting of multi-scale features is guided by global contextual information. Global average pooling is first performed on the highest-level feature map to aggregate global semantic information across the entire image, followed by linear transformation to generate a fixed-dimensional scene semantic embedding vector, computed as follows:

$$z=Linear(GAP(F_4)) \tag{1}$$

where, z denotes the global scene semantic embedding vector, GAP denotes the global average pooling operation, and F_4 denotes the highest-level semantic feature map. Based on this embedding vector, two parallel lightweight multilayer perceptron branches are constructed, with which the fusion weights and distribution modulation parameters corresponding to each hierarchical level are respectively generated:

$$\alpha_i,\beta_i=Sigmoid(MLP_i(z)),\ i\in\{1,2,3,4\} \tag{2}$$

where, α_i denotes the fusion weight coefficient for the i-th level features, used to regulate the overall contribution proportion of features at different scales, and β_i denotes the feature distribution modulation bias, used to correct feature offset errors across different teaching scenarios. The parameter values are adaptively adjusted by the model according to the scene attributes of the input image: low-level detail feature weights are strengthened for text-dense scenes, whereas high-level semantic feature representations are enhanced for full-scene illustration scenes. Moreover, only minimal inference overhead is introduced by the lightweight branch structure, achieving a balance between accuracy and efficiency.

A coarse-to-fine progressive feature fusion strategy is adopted in this work, replacing the simple element-wise summation paradigm of traditional feature pyramids, thereby effectively avoiding semantic drift during multi-scale fusion.

A stable coarse-grained scene representation base is first constructed based on high-level semantic features. The highest-level features are upsampled via bilinear interpolation to match the dimensions of the second-highest-level features, followed by residual enhancement to obtain the base scene features:

$$F_{coarse}=F_3+Conv(Up(F_4)) \tag{3}$$

where, F_{coarse} denotes the coarse-grained fused features, and $Up(\ )$ denotes the bilinear interpolation upsampling operation. Based on the coarse-grained features, all hierarchical features are uniformly upsampled to the smallest feature resolution, and weighted fusion is performed in conjunction with dynamic gating parameters. Finally, the enhanced multi-scale features are obtained via residual connection, which fuses the detailed features with the global base features:

$$F_{enh}=Conv_{1\times1}\left(\sum_{i=1}^{4}\left(\alpha_i\cdot Up_i(F_i)+\beta_i\right)\right)+F_{coarse} \tag{4}$$

where, F_{enh} denotes the final enhanced features. The summation operation achieves adaptive fusion of multi-scale features. The modulation biases correct the distribution deviations of each hierarchical level in real time. In addition, the 1 × 1 convolution is used to integrate the fused feature information and compress the feature dimensions.

Through this progressive fusion architecture—in which a stable global semantic base is first constructed, followed by the injection of local detailed features—the interference of low-level image noise with global scene semantics is effectively suppressed, thereby resolving the representation challenge of simultaneously accommodating large-scale scenes and minute knowledge-point targets in educational imagery. Furthermore, the residual connection design stabilizes gradient propagation during model training and reduces the convergence difficulty of the dynamic gating mechanism, enabling the model to adaptively accommodate the scale feature distributions of various English teaching scenarios and significantly enhancing the feature representation capacity for fine-grained knowledge points.

2.3 Hierarchical instructional semantic-guided cross-modal contrastive constraint module

To eliminate the cross-modal heterogeneous bias between visual representations and pedagogical semantics, a dual-layer cross-modal contrastive constraint system is constructed based on a structured English education knowledge graph, through which precise convergence of visual features and domain semantics is achieved. The schematic diagram of the hierarchical instructional semantic-guided cross-modal contrastive constraint is illustrated in Figure 2. A hierarchical knowledge graph structure is established, with a multi-level entity system defined according to instructional application logic. This system encompasses three entity types—teaching scenes, thematic units, and fine-grained knowledge points—and defines the inclusion, prerequisite association, and scene-adaptation relationships among entities. Based on the English curriculum standards for basic education and textbook statistical data, the association probability distributions between scenes and knowledge points are quantitatively derived, providing structured prior support for the model's semantic reasoning. For standardized textual semantic representation, an education-domain pre-trained encoder, edu-BERT, is employed to encode the graph entities and output fixed-dimensional textual semantic vectors. To unify the representation spaces of visual and textual modalities, a structurally symmetric dual-projection mapping network is designed. Both projection modules adopt a two-layer fully connected structure with the rectified linear unit activation function, uniformly outputting 512-dimensional feature vectors consistent with the text encoder's output dimension. The unified cross-modal projection process can be defined as:

$$f_v=\Phi_v(GAP(F_{enh})),f_t=\Phi_t(edu\text{-}BERT(caption)) \tag{5}$$

where, Φ_v and Φ_t denote the visual projection network and the textual projection network, respectively; F_{enh} denotes the enhanced visual features input to the module; and f_v and f_t denote the visual representation and textual semantic representation in the joint space, respectively. Strict alignment of modal dimensions ensures the stability and effectiveness of cross-modal similarity computation.

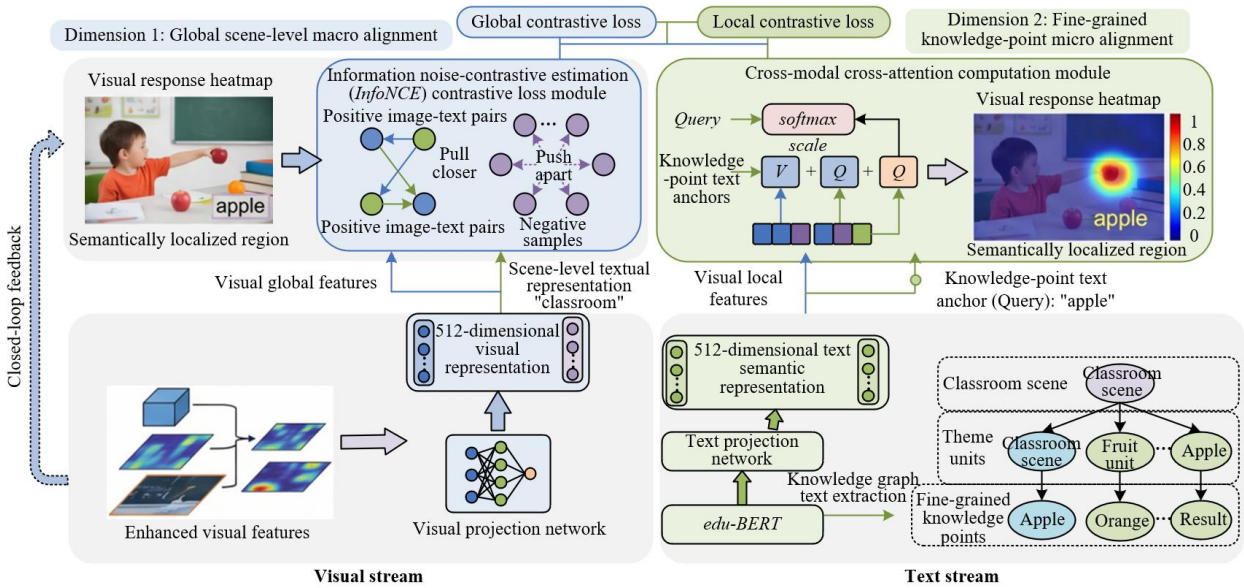

Figure 2. Schematic diagram of the hierarchical instructional semantic-guided cross-modal contrastive constraint

Global scene-level contrastive loss is employed to achieve macro alignment between holistic image content and teaching scene semantics, constraining the model to perceive the instructional attributes of images from a global perspective. Cross-modal contrastive learning constraints are constructed based on the normalized temperature-scaled information noise-contrastive estimation loss function. By differentiating positive and negative samples within each batch, matched image-text sample pairs are pulled closer in the embedding space, while feature intervals between heterogeneous samples are enlarged. The specific computation is formulated as:

$$L_{global}=-\log\frac{exp(cos(f_v,f_t^+)/\tau)}{\sum_{j=1}^{B}exp\,(cos(f_v,f_t^{(j)})/\tau)} \tag{6}$$

where, τ denotes the temperature hyperparameter, used to adjust the dispersion of feature similarity distributions and enhance the discriminative capability of contrastive learning; f_t^+ denotes the positive sample text representation that precisely matches the input image; $f_t^{(j)}$ denotes the text representation of any negative sample within the batch; and B denotes the total batch size. Through this loss function, the learning direction of visual representations is regulated at the global semantic level, effectively avoiding the issues of ambiguous scene classification and global semantic matching misalignment commonly observed in general-purpose models, thereby achieving precise adaptation between global image content and teaching scenes.

To overcome the limitation that global constraints cannot achieve fine-grained knowledge-point localization, a local semantic alignment strategy based on cross-attention mechanism is introduced, with which semantic matching at the pixel-level visual region is accomplished through knowledge-point text anchors, enabling fine-grained semantic association without reliance on manually annotated bounding boxes. The knowledge-point text vectors encoded by the knowledge graph are adopted as query vectors, while the enhanced visual features serve as key-value vectors. Semantic association weights are computed via cross-modal cross-attention, generating knowledge-point-aware visual response heatmaps. The specific computation formula is as follows:

$$Attn=Softmax\left(\frac{(t_vW_Q)(F_{enh}W_K)^T}{\sqrt{d_k}}\right)(F_{enh}W_V) \tag{7}$$

where, $t_v$denotes the text embedding vector of the knowledge-point entity, W_Q, W_K, and W_V denote learnable modality-adaptive weight matrices, and d_k denotes the feature dimension of the query vectors. The magnitude of attention weights directly corresponds to the semantic association strength between image regions and target knowledge points. Based on this, a local contrastive loss function L_{local} is constructed, which drives the model to enhance feature representations of highly associated visual regions while attenuating interference from irrelevant background areas. Through this mechanism, pixel-level semantic alignment of knowledge points is achieved, effectively resolving the technical challenge of ambiguous fine-grained knowledge-point localization.

The global and local dual-layer contrastive losses form a collaborative optimization system, establishing a full-dimensional semantic constraint mechanism spanning from macroscopic scenes to microscopic knowledge points. During model back-propagation, gradient information from the local semantic alignment loss is directly transmitted to the preceding multi-scale fusion module, dynamically adjusting the gating weights and feature distribution parameters of the scene-adaptive dynamic multi-scale gating fusion module. Through this collaborative optimization mechanism, the representational capacity of visual features at corresponding scales is adaptively strengthened according to the semantic localization results of knowledge points, establishing a bidirectional closed loop in which visual feature enhancement empowers semantic alignment, and semantic constraints recursively optimize feature extraction. This mechanism thoroughly bridges the modal gap between visual perception and education-specific pedagogical semantics, significantly enhancing the model's fine-grained semantic understanding capability in English education scenarios.

2.4 Dual-task collaborative optimization module based on semantic topology priors

A parallel dual-task inference structure is constructed based on the optimized joint semantic features, through which teaching scene classification and knowledge-point recognition are simultaneously accomplished, breaking the limitation of single-task independent optimization in traditional educational vision models. Two task-specific output branches with identical architectural configurations are deployed at the top of the model, corresponding respectively to the scene classification task and the knowledge-point recognition task. Each branch is composed of a global average pooling layer and a two-layer fully connected network, capable of mapping the unified semantic features to task-specific probability spaces. The output dimension of the scene classification branch is matched to the preset total number of scene categories, while the output dimension of the knowledge-point recognition branch corresponds to the total number of knowledge points annotated in the dataset, ultimately yielding scene category probability distributions and knowledge-point label probability distributions. Through the parallel dual-branch structure, the representational advantages of the preceding feature enhancement and semantic alignment modules are fully shared, avoiding redundant feature extraction during multi-task training and maintaining the model's lightweight characteristics while preserving inference accuracy. The flowchart of the dual-task collaborative optimization based on semantic topology priors is illustrated in Figure 3.

To constrain the model prediction process through the inherent logic of the teaching domain, a bidirectional collaborative distillation loss function is constructed in conjunction with the topological associations of the English education knowledge graph, through which mutual correction between the scene task and the knowledge-point task is achieved. The conditional probability distributions between scenes and knowledge points are quantitatively derived from the knowledge graph, characterizing the occurrence patterns of various knowledge points across different teaching scenarios. These distributions are normalized via the Softmax function to obtain standardized prior probability distributions. Domain prior information is incorporated into the task supervision process, and the fitting error between dual-task outputs and prior knowledge is measured via Kullback–Leibler divergence. The collaborative distillation loss is computed as follows:

$$L_{co}=KL\left(Softmax(\log P_k+\log P_{prior})\,\|\,P_s\right) \tag{8}$$

where, P_k denotes the model's knowledge-point prediction probability distribution, P_s denotes the scene classification prediction probability distribution, and P_{prior} denotes the scene-knowledge-point prior probability distribution derived from the knowledge graph. A bidirectional constraint mechanism is established through this loss function. On one hand, the long-tail errors in knowledge-point recognition can be corrected using scene classification results and domain priors, compensating for recognition biases in less frequent knowledge points. On the other hand, the plausibility of scene classification results can be reversely validated through the distribution characteristics of knowledge points, ensuring that the dual-task outputs conform to the inherent logic of English instruction.

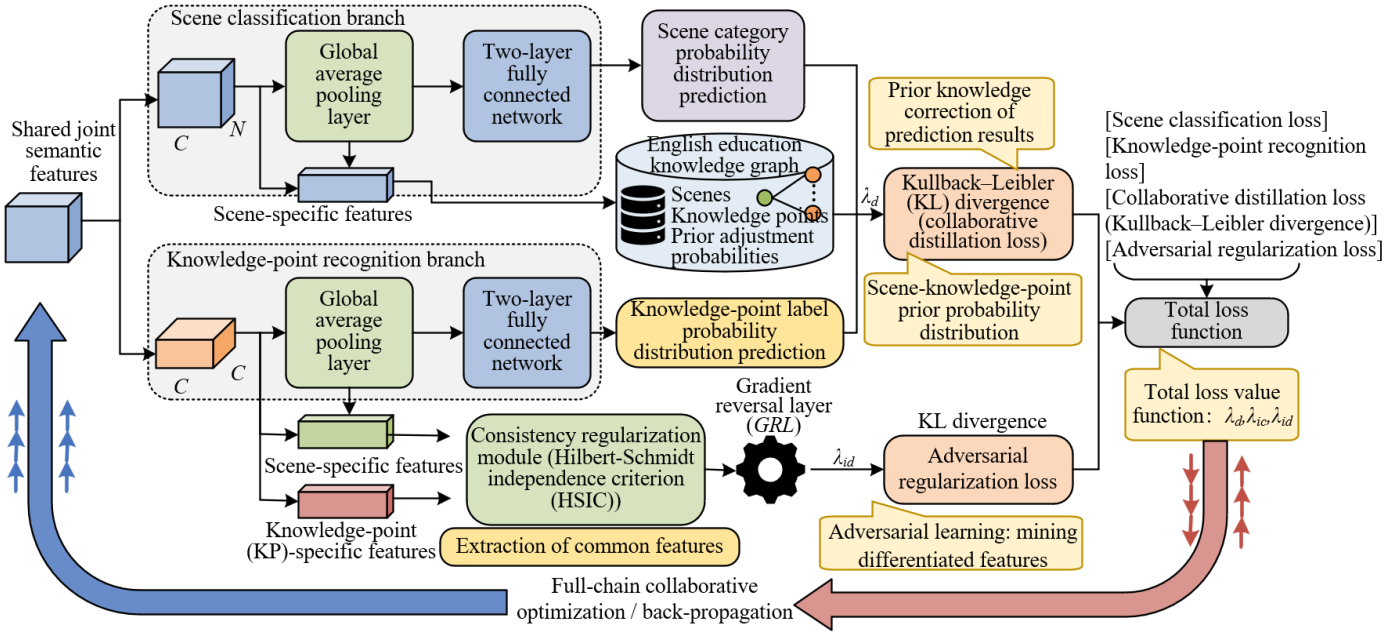

Figure 3. Flowchart of dual-task collaborative optimization based on semantic topology priors

To address the pervasive issues of feature sharing conflicts and representational homogenization in multi-task learning, a feature consistency regularization strategy integrating independence measurement and adversarial constraints is designed, achieving the optimization objective of preserving commonality while maintaining distinction among multi-task features. The task-specific features from the dual-task branches are extracted, and a regularization loss function is constructed by combining the Hilbert-Schmidt independence criterion with Kullback–Leibler divergence-based adversarial constraints:

$$L_{cons}=HSIC(H_s,H_k)-\lambda\cdot KL(P_s\|P_k) \tag{9}$$

where, H_s and H_k denote the task-specific features of the scene task and the knowledge-point task, respectively; $HSIC(\,)$ is used to measure the correlation between the two sets of features; and λ denotes the adversarial constraint balancing coefficient. Maximization of the Hilbert-Schmidt independence criterion drives the dual tasks to share common educational semantic features, reducing redundant learning overhead. Concurrently, a gradient reversal mechanism is introduced into the Kullback–Leibler divergence constraint term, through which the feature values are kept constant during forward propagation, while the gradients are reversely weighted during back-propagation. In an adversarial learning manner, the dual-task branches are forced to mine differentiated feature information, effectively preventing performance degradation caused by multi-task training. This mechanism requires no additional network parameters and achieves task feature balancing solely through gradient modulation, offering high training efficiency.

To integrate multi-dimensional supervision signals and achieve full-chain collaborative optimization, a holistic loss function incorporating multiple constraints is constructed, through which the parameter updates of all modules are uniformly regulated, forming a multi-level closed-loop optimization system. The complete expression of the holistic loss function is:

$$L_{total}=L_{cls}^{s}+L_{cls}^{k}+\gamma_1(L_{global}+L_{local})+\gamma_2 L_{co}+\gamma_3 L_{cons} \tag{10}$$

where, L_{cls}^{s} and L_{cls}^{k} denote the cross-entropy classification losses for scene classification and knowledge-point recognition, respectively, which are responsible for ensuring the discriminative performance of the foundational tasks; γ_1, γ_2, and γ_3 denote hyperparameters used to balance the optimization weights of each loss term. A progressive weight adjustment strategy is adopted to align with the model training dynamics: smaller semantic constraint weights are set in the early training stages, allowing the model to prioritize learning basic visual discriminative features, while the weights of cross-modal contrastive losses are gradually increased in the middle and late training stages to achieve refined convergence in the semantic space. The collaborative distillation and regularization loss weights are maintained at relatively low values to prevent excessive intervention of domain constraints and multi-task adversarial mechanisms in foundational task learning. All hyperparameters are determined through grid

search on the validation set, ensuring training stability and experimental result reproducibility.

Through the complete dual-task collaborative optimization mechanism, deep coupling among visual features, cross-modal semantics, and educational tasks is achieved. The front-end feature enhancement and semantic alignment modules provide high-quality representational support for dual-task inference, while the back-end multi-task supervision signals and domain prior constraints recursively optimize the feature extraction and semantic alignment processes. Through the full-chain gradient back-propagation and bidirectional empowerment mechanism, the disconnection between perceptual tasks and instructional objectives in traditional educational vision models is thoroughly resolved. The image understanding logic of the model is thereby rendered highly consistent with the pedagogical cognitive principles of English instruction, significantly enhancing the model's adaptability and parsing accuracy in real-world smart education scenarios.

3. EXPERIMENTS AND RESULT ANALYSIS

3.1 Experimental setup

A dedicated English education scene image dataset, designated EngEdu-Img, was constructed for the tasks of English teaching image understanding and knowledge-point extraction, filling the gap of specialized visual datasets in the vertical education domain. The image sources in the dataset covered three mainstream English teaching scenarios—static textbook illustrations, real-time classroom blackboard content, and student handwritten assignments—comprising 12 fine-grained teaching scene labels and 87 subject core knowledge-point labels, with a cumulative total of 12,860 valid annotated images. The dataset was randomly partitioned into training, validation, and test sets at a ratio of 7:1.5:1.5, with the partitioning process strictly ensuring balanced distribution of samples across all scene categories and knowledge points. All image annotation work was completed by professional English teaching and research personnel, with scene classification, knowledge-point category, and fine-grained region annotations performed simultaneously, effectively ensuring the domain specificity and annotation accuracy of the dataset and providing reliable data support for model training and performance validation.

All models were constructed, trained, and tested based on the PyTorch deep learning framework, with the hardware platform equipped with a single NVIDIA A100 80GB graphics processing unit. The backbone network was initialized with ConvNeXt-V2-Base weights pre-trained on the ImageNet dataset. Input images were uniformly resized to 512 × 512 pixels, and the training batch size was set to 32. The AdamW optimizer was adopted for model optimization, with an initial learning rate of 3e-5, combined with a cosine annealing strategy for dynamic learning rate decay. The total number of training epochs was set to 60, with a learning rate warm-up operation implemented during the first 5 epochs to ensure training stability in the early stages. The temperature coefficient for cross-modal contrastive learning was fixed at 0.07, and all loss-balancing hyperparameters were determined through grid search on the validation set to ensure experimental reproducibility.

3.2 Baseline comparison experiments

To comprehensively validate the overall performance advantages of the proposed framework, comparative experiments were conducted between the proposed method and mainstream algorithms in the fields of multi-scale fusion, cross-modal semantic understanding, and educational visual processing. For multi-scale fusion methods, the feature pyramid network, the path aggregation network, and the bidirectional feature pyramid network were selected as baselines to verify the superiority of the dynamic feature fusion mechanism. For cross-modal methods, three classical pre-trained models—contrastive language–image pre-training, align before fuse, and bootstrapping language–image pre-training—were adopted to validate the enabling effects of domain semantic constraints. For educational vision methods, the existing specialized educational vision models EduVision and EduFormer were selected as comparison targets. With scene classification Top-1 and Top-3 accuracy, as well as knowledge-point recognition mean average precision and Macro-F1, adopted as the core evaluation metrics, quantitative performance comparisons of all algorithms were completed. Table 1 presents the quantitative comparison results of different methods on the EngEdu-Img test set.

Table 1. Overall performance comparison results of different methods

Method Category	Model	Top-1 Accuracy (%)	Top-3 Accuracy (%)	Mean Average Precision (%)	Macro-F1 (%)
Multi-scale fusion	Feature Pyramid Network	82.36	93.15	71.28	70.92
	Path Aggregation Network	83.52	94.02	72.65	72.15
	Bidirectional Feature Pyramid Network	84.18	94.56	73.81	73.06
Cross-modal understanding	Contrastive Language–Image Pre-training	85.62	95.18	75.23	74.85
	Align Before Fuse	86.35	95.72	76.19	75.62
	Bootstrapping Language–Image Pre-training	87.01	96.05	77.05	76.38
Education-specific model	EduVision	87.68	96.32	78.26	77.51
	EduFormer	88.25	96.78	79.12	78.29
Proposed method	An integrated framework for image understanding and information extraction specific to English education scenarios	92.17	98.64	85.37	84.92

From the experimental results, it is observed that the proposed method significantly outperforms all comparison algorithms across all evaluation metrics. Compared to the optimal baseline model, EduFormer, the Top-1 accuracy of the proposed method is improved by 3.92%, the Top-3 accuracy by 1.86%, the mean average precision metric by 6.25%, and the Macro-F1 metric by 6.63%. Traditional multi-scale fusion methods can only perform simple stacking of visual features, lacking adaptability to the extreme scale disparities in educational imagery, with obvious deficiencies in fine-grained

knowledge-point recognition performance. Generic cross-modal pre-trained models possess basic image-text alignment capabilities but lack educational domain prior knowledge constraints, enabling only coarse-grained scene matching without achieving knowledge-point-level fine-grained semantic parsing. Existing education-specific models, although optimized for educational scenarios, have not achieved full-chain coupling of feature enhancement, semantic alignment, and task collaboration, suffering from disconnection between visual representations and pedagogical semantics as well as multi-task optimization conflicts. Through the adaptive feature enhancement enabled by the dynamic multi-scale gating fusion, the semantic gap bridged by the knowledge graph-driven dual-layer cross-modal constraints, and the task conflicts resolved by the topology prior-based dual-task collaborative mechanism, a complete optimization system tailored for educational scenarios is formed by the proposed method. Consequently, comprehensive improvements in both scene classification and knowledge-point recognition performance are achieved, fully validating the effectiveness and superiority of the framework in English education image understanding tasks.

3.3 Module ablation experiments

To validate the effectiveness of each core module and its subcomponents, systematic ablation experiments were conducted in this section. A ConvNeXt-V2 backbone network equipped with a dual-task classification head was adopted as the baseline model, onto which the scene-adaptive dynamic multi-scale gating fusion module, the hierarchical semantic-guided cross-modal contrastive constraint module, and the dual-task collaborative optimization module were successively added to verify the coupling gains of the overall framework. Simultaneously, component-wise ablation experiments were performed for the core subcomponents of each module, validating the independent contributions of the dynamic gating mechanism, the local contrastive loss, the collaborative distillation loss, and the feature consistency regularization term. The overall module ablation experimental results are presented in Table 2, and the core subcomponent ablation experimental results are illustrated in Figure 4.

Table 2. Overall module ablation experimental results

Model Configuration	Top-1 Accuracy (%)	Top-3 Accuracy (%)	Mean Average Precision (%)	Macro-F1 (%)
Baseline model	84.18	94.56	73.81	73.06
Baseline + the scene-adaptive dynamic multi-scale gating fusion module	87.53	96.12	78.65	77.93
Baseline + the scene-adaptive dynamic multi-scale gating fusion module + the hierarchical semantic-guided cross-modal contrastive constraint module	90.26	97.58	82.14	81.65
Baseline + the scene-adaptive dynamic multi-scale gating fusion module + the hierarchical semantic-guided cross-modal contrastive constraint module + the dual-task collaborative optimization module (full module)	92.17	98.64	85.37	84.92

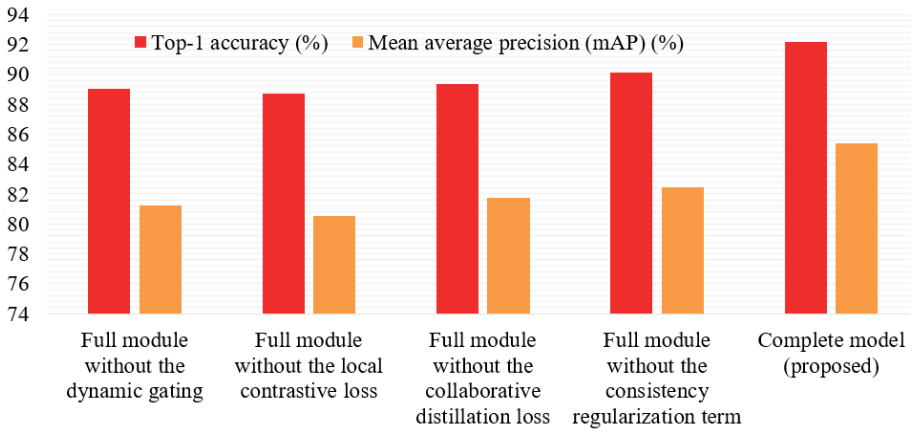

Figure 4. Core subcomponent ablation experimental results

From the overall module ablation results, it is observed that each core module contributes positive gains to model performance, with significant collaborative optimization effects exhibited among the three modules. After the addition of only the scene-adaptive dynamic multi-scale gating fusion module, the model's scale adaptation capability is substantially enhanced, with the mean average precision improved by 4.84% compared to the baseline model, demonstrating that the dynamic gating fusion strategy can effectively address the representation challenge caused by uneven scale distribution in educational imagery. Following the addition of the hierarchical semantic-guided cross-modal contrastive constraint module, the model's semantic alignment capability is markedly strengthened, with further improvements achieved in both scene classification and knowledge-point recognition performance, indicating that hierarchical cross-modal constraints can effectively bridge the heterogeneous semantic gap between visual perception and pedagogical semantics. With all modules combined, optimal model performance is attained, validating the coupling value of the tripartite architecture comprising feature enhancement, semantic alignment, and task collaboration. The modules are not simply stacked but achieve bidirectional empowerment through full-chain gradient back-propagation. The subcomponent ablation experimental results indicate that all core innovative subcomponents are necessary conditions for high model performance. The dynamic gating mechanism exerts the greatest influence on fine-grained recognition performance,

with a notable decline in knowledge-point recognition accuracy observed after its removal. The local contrastive loss directly determines the model's pixel-level knowledge-point localization capability, serving as the core component for achieving fine-grained semantic alignment. The collaborative distillation and consistency regularization terms effectively optimize the multi-task learning logic, avoiding task conflicts and feature homogenization while ensuring synchronous and efficient optimization of both tasks. The synergistic effects of all components collectively support the optimal performance of the model.

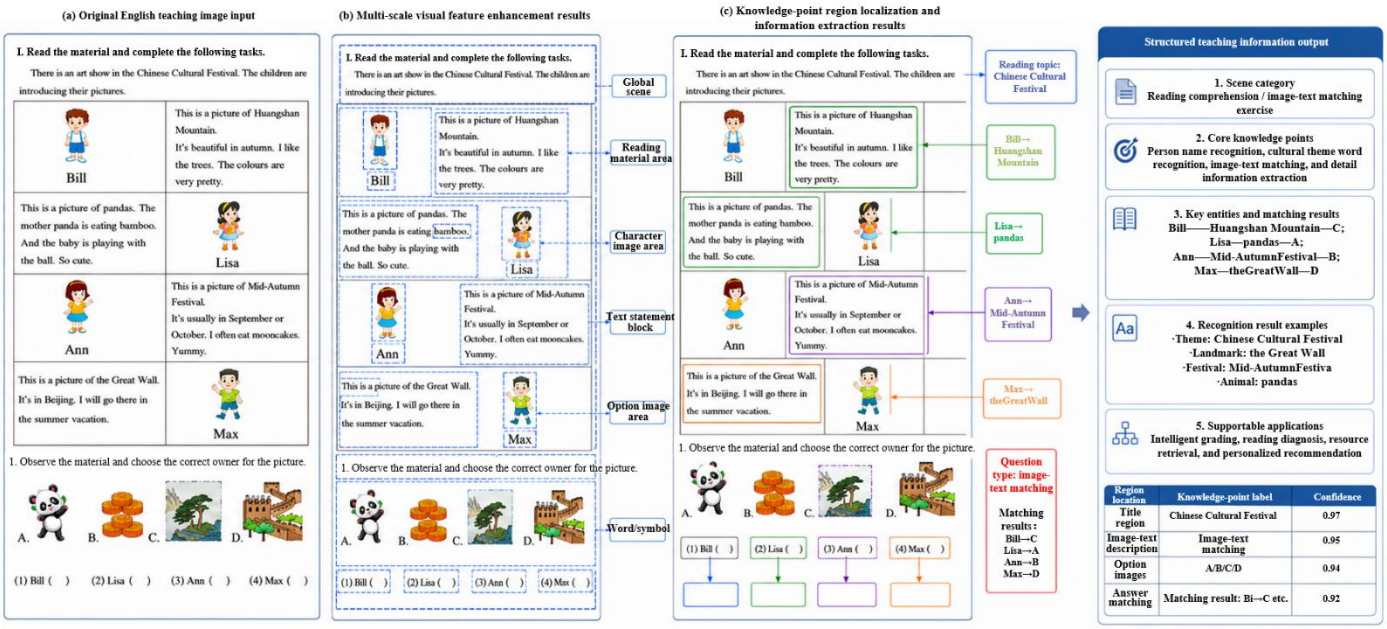

Figure 5. Implementation effect diagram of English education image understanding

To validate the complete effectiveness of the proposed method—from image perception to semantic understanding and then to structured information output—in real-world English education scenarios, visualization experiments on method implementation effects are conducted. As illustrated in Figure 5, after the input of the original English reading exercise image, the model is able to first stably cover targets at different hierarchical levels—including the full-page layout, character illustrations, text statement blocks, option image areas, and fine-grained word symbols—through the multi-scale visual feature enhancement mechanism, demonstrating that the proposed method possesses favorable scale adaptation capability for the complex characteristic of coexisting macroscopic layouts and microscopic knowledge units in educational imagery. On this basis, the knowledge-point region localization results further indicate that the model is capable not only of accurately establishing correspondences among character names, text descriptions, and candidate images but also of effectively recognizing reading topics, cultural entities, and key information items, satisfactorily accomplishing image-text matching, detail extraction, and semantic association tasks. These results reflect the significant role of cross-modal semantic constraints in eliminating the misalignment between visual representations and pedagogical semantics. Particularly in the matching results for Bill, Lisa, Ann, and Max, as well as the extraction of key entities such as Chinese Cultural Festival, Huangshan Mountain, Mid-Autumn Festival, the Great Wall, and pandas, the model exhibits clear semantic orientation and strong pedagogical knowledge interpretability. Finally, the structured teaching information output module is able to uniformly organize scene categories, core knowledge points, key entity matching relationships, recognition result examples, and region-level confidence information, demonstrating that the proposed method has transcended mere image recognition and achieved high-quality information extraction and usable knowledge generation oriented toward smart English education applications.

3.4 Scale robustness validation experiments

To specifically validate the adaptation capability of the scene-adaptive dynamic multi-scale gating fusion module for multi-scale targets, the knowledge-point targets in the test set were divided into three scale levels—large, medium, and small—according to pixel occupancy ratio in this section. The recognition accuracies of the optimal baseline model, EduFormer, and the proposed method on targets at different scales were respectively calculated, with quantitative comparisons of the scale robustness of the two model types conducted. The perceptual advantage of the proposed method for small-scale knowledge points was particularly verified. Figure 6 presents the comparison results of multi-scale target recognition performance.

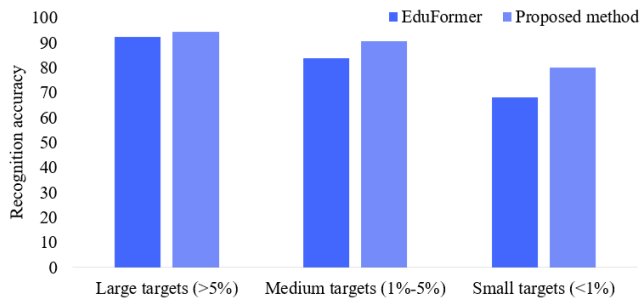

Figure 6. Recognition accuracy comparison for knowledge-point targets at different scales (%)

From the experimental data, it is observed that the recognition accuracy gap between the two model types for large-scale teaching targets is relatively small, with high-accuracy recognition achieved by both, demonstrating that visual perception of conventional-scale targets is of low difficulty. In the medium-scale target recognition task, the accuracy of the proposed method is improved by 6.96% over the comparison model, with the scale adaptation advantage preliminarily demonstrated. For the most challenging small-scale knowledge-point targets in educational scenarios—such as individual words and characters—the recognition accuracy of the baseline model is substantially degraded, while the proposed method is still able to maintain relatively high recognition accuracy, with a performance improvement of up to 11.68%. These results fully confirm the core advantage of the scene-adaptive dynamic gating fusion mechanism. Through this mechanism, the multi-scale feature weights are dynamically adjusted by the model according to the scene attributes of the input image, with low-level fine-grained features automatically strengthened in text-dense scenes. Consequently, the deficiency of insufficient small-target representation capability in traditional static fusion methods is effectively compensated, thoroughly resolving the recognition challenge caused by the large scale span in English teaching imagery and demonstrating strong scale robustness.

3.5 Generalization and robustness validation experiments

To validate the cross-scenario generalization capability of the model, a cross-data-source test set was constructed, containing three types of out-of-domain samples: English textbook images from different versions, classroom images captured by mobile devices, and teaching slide images captured from screens. The performance of the model on non-training data sources was tested. The experimental results are presented in Table 3.

From the data, it is observed that the performance degradation of the model across various cross-domain data sources is relatively small, with Top-1 accuracy maintained above 88% and the mean average precision metric stabilized above 81% in all cases. Without specific fine-tuning for out-of-domain scenarios, the model is still able to maintain excellent recognition performance, demonstrating that the proposed framework learns general English teaching semantic features rather than dataset-specific features, possessing favorable cross-scenario generalization capability and adaptability to diverse smart teaching application scenarios. To validate the anti-interference capability of the model in complex real-world environments, three types of common image degradations—Gaussian noise, motion blur, and illumination shift—are respectively added to the test set images, with the stability of the model tested. The experimental results are illustrated in Figure 7.

Table 3. Cross-data-source generalization performance test results

Test Data Source	Top-1 Accuracy (%)	Mean Average Precision (%)
Training set same-source data	92.17	85.37
Different-version textbook images	90.35	83.26
Classroom captured images	89.62	82.15
Screen-captured slide images	88.95	81.63

From the experimental results, it is observed that only slight performance degradation is exhibited by the model under various interference conditions, with overall accuracy still maintained at a relatively high level. Among these conditions, the smallest impact on the model is caused by illumination shift, while the most significant interference effect is exhibited by motion blur; however, the overall degradation is controlled within 4% in all cases. Benefiting from the multi-scale feature enhancement and semantic prior constraints, the model does not rely solely on low-level visual texture features but primarily performs inference and judgment based on high-level pedagogical semantics. Consequently, strong anti-interference capability is possessed by the model against visual disturbances such as image noise, blur, and illumination variations, enabling adaptability to complex image acquisition environments in real teaching scenarios and demonstrating favorable potential for practical deployment.

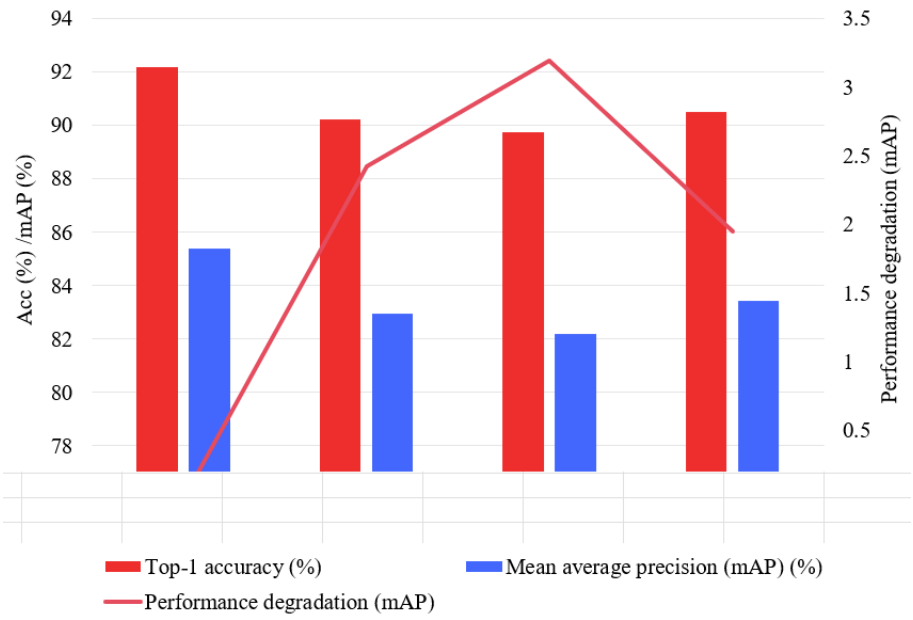

Figure 7. Model robustness test results under interference conditions

4. DISCUSSION

In conjunction with the quantitative experimental data and visualization results presented above, the intrinsic operational mechanisms and module collaborative logic of the proposed integrated framework are further analyzed in this chapter, with the core principles underlying the model's performance improvements elucidated. The dynamic multi-scale gating fusion mechanism designed in this work abandons the traditional fixed-weight feature stacking paradigm, relying on global scene semantic information to drive adaptive weighted regulation of multi-scale features, whereby detail features or global semantic features can be dynamically strengthened according to the visual distribution characteristics of English teaching scenarios. In conjunction with the coarse-to-fine progressive fusion paradigm, the model is able to construct a stable scene semantic base using high-level features, thereby avoiding semantic drift caused by low-level texture noise and fundamentally resolving the challenges of large scale span and weak small-target representation in teaching imagery. On this basis, the hierarchical cross-modal contrastive constraint system establishes a dual-layer semantic convergence mechanism from global scenes to local knowledge points, using the domain knowledge graph as a semantic anchor to guide visual features toward fine-grained alignment. Moreover, the semantic supervision signals can recursively optimize the front-end feature fusion process, forming a closed-loop empowerment relationship between feature enhancement and semantic matching. The dual-task collaborative optimization strategy, based on the inherent topological regularities of teaching scenarios, unifies the semantic logic of both tasks through collaborative distillation constraints. Combined with the feature regularization mechanism, this strategy balances task feature sharing and differentiated learning, effectively resolving the industry pain points of disconnection between perceptual tasks and educational objectives in traditional vision models, as well as feature conflicts in multi-task learning.

The proposed framework exhibits excellent recognition accuracy, scale robustness, and generalization capability in conventional English teaching scenarios; however, clear performance boundaries and application limitations remain. For extremely small-scale characters with very low pixel occupancy and severely distorted handwritten English content, effective visual feature information is scarce and morphological irregularities are present, making it difficult for the model to achieve precise feature matching and knowledge-point recognition, with obvious bottlenecks existing in fine-grained parsing performance. Meanwhile, the semantic reasoning capability of the model is highly dependent on the coverage of the structured educational knowledge graph. The current knowledge base, with its limited entity count and topological associations, only covers conventional teaching scenarios and core knowledge points, rendering it incapable of adapting to extended teaching content and niche teaching scenarios. Consequently, the breadth of the model's semantic understanding is constrained by the completeness of the domain priors. Furthermore, under extremely complex acquisition conditions such as large-area occlusion and high-intensity imaging noise, the fundamental visual features of the images are severely degraded, with slight degradation observed in the model's cross-modal semantic alignment accuracy and task recognition performance.

In view of the limitations of the current research, future work can be continuously deepened in three directions: scenario expansion, capability upgrade, and deployment optimization. Subsequently, the educational knowledge system can be extended to cover multiple grade levels and multiple languages, with the topological association relationships between scenes and knowledge points refined, thereby breaking through the application constraints of a single English scenario and constructing a more generalizable educational visual semantic parsing framework. Furthermore, the open-ended semantic generation capability of large language models can be integrated to break through the inherent paradigm of fixed-label classification, achieving open-ended knowledge-point mining, semantic summarization, and structured information extraction from teaching imagery, with the higher-order semantic understanding capability of the model further enhanced. Finally, lightweight optimization of the overall framework can be performed for terminal deployment requirements, with computational overhead reduced through model distillation, structural pruning, and quantization inference, thereby adapting to the real-time inference requirements of various teaching terminals and mobile devices and promoting the large-scale deployment and engineering application of intelligent educational image understanding technology in smart teaching scenarios.

5. CONCLUSION

In response to the three core challenges in intelligent parsing of English teaching imagery—scale representation imbalance, semantic heterogeneity between visual perception and pedagogical semantics, and the disconnection between perceptual tasks and teaching cognitive objectives—an end-to-end integrated visual understanding framework was constructed in this study, forming a complete technical system characterized by deep coupling of feature enhancement, semantic alignment, and task collaboration. Through the scene-adaptive dynamic gating fusion mechanism, adaptive modulation of multi-scale features was accomplished, effectively overcoming the representation deficiencies caused by scale disparities between panoramic illustrations and minute textual symbols. By leveraging the hierarchical English education knowledge graph, a dual-layer cross-modal contrastive constraint was established, guiding visual features to converge toward the professional pedagogical semantic space and systematically eliminating the inter-modal semantic gap. In conjunction with the topological priors of scenes and knowledge points, a dual-task collaborative optimization mechanism was constructed, achieving bidirectional iterative optimization of visual perception and educational objectives while mitigating the pervasive issues of feature competition and logical bias in multi-task learning.

Based on the self-constructed EngEdu-Img dataset, multiple sets of validation experiments were conducted, including baseline comparisons, module ablation studies, scale robustness evaluations, generalization and anti-interference tests, and visualizations. Both quantitative metrics and qualitative visualization results demonstrated that the proposed framework significantly outperformed existing mainstream algorithms in scene classification and knowledge-point recognition tasks, while also exhibiting stable scale adaptation capability and cross-domain generalization

performance. The complete specialized image parsing scheme proposed in this study not only refines the implementation pathway for cross-modal visual perception in the vertical education domain but also provides a valuable technical paradigm for the deployment of general-purpose computer vision technologies to specialized industry scenarios.

REFERENCES

[1] Zhai, X., Chu, X., Chai, C.S., et al. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. Complexity, 2021: 8812542. https://doi.org/10.1155/2021/8812542

[2] Nai, R. (2022). The design of smart classroom for modern college English teaching under Internet of Things. PLOS ONE, 17(2): e0264176. https://doi.org/10.1371/journal.pone.0264176

[3] 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

[4] Chu, L., Liu, Y., Zhai, Y., Wang, D., Wu, Y. (2024). The use of deep learning integrating image recognition in language analysis technology in secondary school education. Scientific Reports, 14(1): 1-11. https://doi.org/10.1038/s41598-024-52592-5

[5] Memon, J., Sami, M., Khan, R.A., Uddin, M. (2020). Handwritten optical character recognition (OCR): A comprehensive systematic literature review (SLR). IEEE Access, 8: 142642-142668. https://doi.org/10.1109/access.2020.3012542

[6] Zhang, Q., Liu, F., Song, W. (2025). IMTLM-Net: Improved multi-task transformer based on localization mechanism network for handwritten English text recognition. Complex & Intelligent Systems, 11(1): 1-18. https://doi.org/10.1007/s40747-024-01713-8

[7] Liu, L., Ouyang, W., Wang, X., et al. (2019). Deep learning for generic object detection: A survey. International Journal of Computer Vision, 128(2): 261-318. https://doi.org/10.1007/s11263-019-01247-4

[8] Cheng, G., Yuan, X., Yao, X., et al. (2023). Towards large-scale small object detection: Survey and benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11), 13467-13488. https://doi.org/10.1109/tpami.2023.3290594

[9] Abdallah, A., Eberharter, D., Pfister, Z., Jatowt, A. (2024). A survey of recent approaches to form understanding in scanned documents. Artificial Intelligence Review, 57(12): 1-38. https://doi.org/10.1007/s10462-024-11000-0

[10] Chernyshova, Y.S., Sheshkus, A.V., Arlazarov, V.V. (2020). Two-step CNN framework for text line recognition in camera-captured images. IEEE Access, 8: 32587-32600. https://doi.org/10.1109/access.2020.2974051

[11] Ma, W., Gong, C., Xu, S., Zhang, X. (2020). Multi-scale spatial context-based semantic edge detection. Information Fusion, 64: 238-251. https://doi.org/10.1016/j.inffus.2020.08.014

[12] Qian, L., Yu, T., Yang, J. (2023). Multi-scale feature fusion of covariance pooling networks for fine-grained visual recognition. Sensors, 23(8): 3970. https://doi.org/10.3390/s23083970

[13] Li, X., Ding, H., Yuan, H., et al. (2024). Transformer-based visual segmentation: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12): 10138-10163. https://doi.org/10.1109/tpami.2024.3434373

[14] 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/10.1109/tpami.2024.3369699

[15] Chen, F., Zhang, D., Han, M., et al. (2023). VLP: A survey on vision-language pre-training. Machine Intelligence Research, 20(1): 38-56. https://doi.org/10.1007/s11633-022-1369-5

[16] Wang, X., Chen, G., Qian, G., et al. (2023). Large-scale multi-modal pre-trained models: A comprehensive survey. Machine Intelligence Research, 20(4): 447-482. https://doi.org/10.1007/s11633-022-1410-8

[17] Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C. (2023). Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4): 4396-4415. https://doi.org/10.1109/TPAMI.2022.3195549

[18] Zhang, Y., Yang, Q. (2022). A survey on multi-task learning. IEEE Transactions on Knowledge and Data Engineering, 34(12): 5586-5609. https://doi.org/10.1109/TKDE.2021.3070203

[19] Vandenhende, S., Georgoulis, S., Van Gansbeke, W., Proesmans, M., Dai, D., Van Gool, L. (2021). Multi-task learning for dense prediction tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(7): 3614-3633. https://doi.org/10.1109/tpami.2021.3054719

[20] Han, J., Zhong, W., Ding, C., Wang, C. (2025). A multi-task learning framework with prompt-based hybrid alignment for vision-language model. Neurocomputing, 638: 130121. https://doi.org/10.1016/j.neucom.2025.130121