



Generative Artificial Intelligence-Driven Image Semantic Segmentation and Visual Scene Understanding for English Language Teaching Environments

Yun Li^{ID}, Lan Zhao^{ID}, Yan Zhao^{*ID}

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

Corresponding Author Email: lcc_zhaoyan@bdlg.edu.cn

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.430424>

ABSTRACT

Received: 26 April 2026

Revised: 20 August 2026

Accepted: 27 August 2026

Available online: 31 August 2026

Keywords:

image semantic segmentation, visual scene understanding, diffusion models, attention mechanism, teaching scene parsing

While diffusion models have established a new paradigm for open-world image segmentation and scene understanding through their inherent vision-language joint representations, existing methods face notable limitations when applied to the text-dense and semantically hierarchical domain of English language teaching (ELT). These limitations include semantic misalignment with generic category taxonomies, coarse multi-scale attention fusion, and the disconnection between pixel-level perception and scene-level reasoning. To address these challenges, this paper proposes a Generative Scene Parsing Network (GSP-Net), an end-to-end intelligent parsing framework tailored for pedagogical scenarios. First, we construct a domain-specific taxonomy encompassing three hierarchical levels: physical environment, instructional media, and semantic content. A scene-type-aware dynamic prompt encoding strategy is introduced to steer the cross-attention of the diffusion model, enabling it to accurately anchor teaching-relevant semantic regions conditioned on the current lesson type. Second, a differentiable attention gating network is proposed to adaptively fuse multi-level self-attention features based on spatial location and local text density, significantly refining the boundary clarity of densely textured areas such as blackboard writing and projection zones. Furthermore, a differentiable scene graph propagation module is designed to transform segmentation masks into structured scene representations, which subsequently drive a large language model (LLM) for semantic reasoning over teaching themes and core knowledge points. Notably, a closed-loop feedback mechanism based on image reconstruction loss is established to back-propagate generative errors to the attention fusion module, enabling self-supervised continual evolution of the segmentation model under annotation-free scenarios. This work provides both a feasible technical pathway and theoretical support for the in-depth application of generative AI in educational visual understanding.

1. INTRODUCTION

Generative artificial intelligence (AI) represented by diffusion models has triggered a profound paradigm shift in the field of computer vision [1, 2]. With their superior generation quality and stable training characteristics, diffusion models have become the mainstream technical route for image generation tasks. However, in recent years, researchers have gradually realized that the value of pre-trained diffusion models goes far beyond generation itself; the attention maps produced during the denoising process contain rich vision-semantic prior knowledge [3, 4]. Specifically, the self-attention maps in models such as Stable Diffusion can capture internal spatial structures and object contours of images, while the cross-attention maps establish dense correspondence between textual semantics and image regions. This discovery has given rise to a series of zero-shot image segmentation methods such as Diffusion-based Training-free Segmenter (DiffSegmenter), Seeded + Diffusion (SeeDiff), and Fast and Accurate Segmentation (FA-Seg). These methods can extract semantic segmentation masks from the internal

representations of diffusion models without any labeled data, demonstrating the great potential of generative models in visual understanding tasks [5, 6].

The deep integration of image segmentation and visual scene understanding is an inevitable trend for computer vision to advance toward high-level semantic reasoning [7, 8]. Breakthroughs in multimodal large models have made end-to-end processing from pixel-level perception to scene-level cognition possible; vision-language models are capable of simultaneously performing composite tasks such as object detection, relation extraction, and semantic reasoning [9, 10]. However, existing research mostly targets general scenarios such as autonomous driving, remote sensing imagery, or natural images. Systematic research focusing on the educational domain—a vertical scenario characterized by a highly structured semantic system and clear application orientation—remains notably absent [11, 12]. The visual intelligent parsing of English teaching scenes possesses unique academic value and application demands. Against the backdrop of the accelerated deployment of smart education, automated understanding of classroom teaching images—

including the recognition of physical environment elements such as blackboards, podiums, and projection screens; the localization of instructional media such as writing areas, projection clips, and word cards; and the comprehension of semantic content such as English vocabulary, sentence structures, and grammatical annotations—constitutes the key technical foundation for applications including intelligent teaching assistance, classroom behavior analysis, and automatic instructional resource generation [13, 14]. The prevalent dense text arrangement, interleaved semantic hierarchies, and lesson-type-driven scene dynamics in teaching scenarios impose stringent requirements on segmentation accuracy and semantic understanding capability that far exceed those of general scenarios [15-18].

Although generative AI has achieved substantial results in image segmentation and scene understanding, research oriented toward English teaching scenarios still faces several key bottlenecks [19, 20]. First, a fundamental semantic mismatch exists between the general semantic category systems relied upon by existing zero-shot segmentation methods and the semantic elements of teaching scenes [21, 22]. Whether ADE20K or Cityscapes, their category definitions revolve around daily objects or traffic elements; core semantic concepts in teaching scenes—such as writing text areas, word cards, and sentence structures—are entirely absent from general taxonomies, causing segmentation models to fail to effectively encode and recognize teaching-specific objects [11, 12]. Second, the fusion strategies for multi-level attention maps in diffusion models lack fine-grained design oriented toward scene characteristics [3, 4]. Existing methods typically apply simple averaging to self-attention maps at various layers or adopt globally fixed fusion weights. Such practices overlook the inherent differences in spatial topological properties and semantic granularity across attention maps at different levels, and fail to adaptively adjust the reliance on spatial details versus global context according to scene type [5, 6]. This problem is particularly acute in text-dense regions of teaching scenes; blurred boundaries between writing areas and projection zones lead to content adhesion or jagged edges in segmentation results, yet existing methods offer no specialized enhancement mechanism for such regions [17, 18]. Third, there is a significant task disconnection among image segmentation, scene understanding, and content generation [23, 24]. Existing work typically treats the three as independent pipeline stages; the quality of segmentation output directly determines the upper bound of understanding, while understanding results can neither retroactively correct errors in the segmentation stage nor leverage the reconstruction capability of generative models to impose self-supervised constraints on segmentation features [25, 26]. This unidirectional information flow leads to error accumulation across stages and forfeits the potential for continuous optimization brought by closed-loop feedback.

To address the aforementioned problems, this paper proposes the Generative Scene Parsing Network (GSP-Net). The core contributions of this framework are reflected in three aspects. First, we construct a hierarchical semantic category system for English teaching scenes, named ETST, covering 21 categories across three tiers: physical environment layer, instructional media layer, and semantic content layer. We also design a scene-type-driven dynamic prompt encoding strategy to enable the cross-attention of the diffusion model to precisely anchor teaching-semantic regions relevant to the current lesson type. Second, a differentiable attention gating network

is proposed to achieve spatially adaptive fusion of multi-level self-attention maps, while a structure-tensor-based text density enhancement factor is introduced to specifically optimize the edge-blurring problem in writing and projection areas. Third, we build an integrated closed-loop architecture of segmentation-understanding-generation. A differentiable scene graph propagation network transforms segmentation results into structured scene representations to drive a large language model (LLM) in performing teaching semantic reasoning; furthermore, image reconstruction loss is utilized to construct a feedback loop that back-propagates error signals to the attention fusion module, realizing self-supervised continual evolution of the segmentation model in annotation-free scenarios. Section 2 elaborates on the overall framework and technical details of GSP-Net; Section 3 reports the experimental setup, dataset construction, comparative evaluation, and ablation analysis results; Section 4 concludes the paper and discusses future directions.

2. METHODOLOGY

2.1 Overall framework overview

GSP-Net adopts pre-trained Stable Diffusion 2.1 as the backbone to construct an end-to-end image semantic segmentation and visual scene understanding framework oriented toward English teaching scenarios. Let the input teaching scene image be $I \in \mathbb{R}^{H \times W \times 3}$; the network outputs a pixel-level semantic segmentation mask M , a structured scene graph G_{scene} , and a teaching-semantic-enhanced image I_{gen} . The overall framework consists of three collaborative modules: a teaching-context-driven dynamic prompt encoding and hierarchical attention anchoring module, which injects scene semantic priors into the diffusion process; a spatial-density-driven multi-scale attention topology fusion module, responsible for generating refined segmentation masks; and a differentiable scene graph propagation and generative closed-loop feedback module, which completes the cross-modal mapping from segmentation to semantic reasoning and enables self-supervised evolution. The three modules share a computational graph during back-propagation, forming a unified system that can be optimized end-to-end.

The U-Net denoising sub-network of the diffusion model produces two complementary types of attention maps during generation. Let the number of U-Net decoder layers be L , at denoising time step t , the cross-attention map of the l -th layer is $A_l^{cross,t} \in \mathbb{R}^{H_l \times W_l \times L_p}$, where $H_l \times W_l$ is the spatial resolution and L_p is the text prompt length. The self-attention map is $A_l^{self,t} \in \mathbb{R}^{H_l \times W_l \times D_l}$, where D_l is the number of attention heads at that layer. The cross-attention maps characterize the response intensity between each semantic category in the text condition and the image spatial positions, while the self-attention maps encode the long-range dependencies among pixels within the image and the shape-topological information of objects. The core design of GSP-Net lies in the scene-adaptive extraction, fusion, and enhancement of these two types of attention maps, rather than adopting the simple averaging or fixed-weight combination strategies used in existing methods. A differentiable attention gating network dynamically adjusts the contribution ratio of self-attention at each layer based on spatial location and local image characteristics, allowing the fusion process to adaptively adjust with scene content. A text density enhancement factor

further focuses on text-dense regions such as writing areas and projection zones, locating high-texture directional regions via structure tensor analysis and amplifying their attention responses, effectively alleviating semantic adhesion at dense text boundaries.

The inference stage adopts a single-pass denoising strategy to balance computational efficiency and segmentation quality. Given the input image I and the scene-aware condition C_i , the model executes only $T = 20$ denoising steps instead of the conventional 50-step full sampling of diffusion models. This design is based on the following observation: attention maps have already converged to a stable semantic response distribution in the early-to-mid denoising steps; further increasing the number of steps yields only marginal improvement in segmentation accuracy while significantly increasing computational burden. To eliminate random fluctuations in single-step attention maps, attention maps across all time steps are aggregated to obtain stable semantic responses:

$$\bar{A}_l^{cross} = \frac{1}{T} \sum_{t=1}^T A_l^{cross,t}, \bar{A}_l^{self} = \frac{1}{T} \sum_{t=1}^T A_l^{self,t}, \quad (1)$$

where $\bar{A}_l^{cross}$ and $\bar{A}_l^{self}$ are the time-step-aggregated cross-attention maps and self-attention maps, respectively, serving as inputs for subsequent fusion and enhancement processing. The total computational complexity of GSP-Net is $O(T \cdot L \cdot H \cdot W \cdot d)$, where d is the feature dimension of attention heads. Compared with baseline methods, the additionally introduced computational overhead mainly comes from the forward propagation of the differentiable gating network ($O(L \cdot H \cdot W)$) and the computation of the structure-tensor-based text density factor ($O(H \cdot W \cdot k^2)$), where k is the Gaussian kernel size; the combined portion accounts for approximately 12% to 15% of total inference time, which is within an acceptable range for practical deployment.

2.2 Teaching-context-driven dynamic prompt encoding and hierarchical attention anchoring

Semantic objects in teaching scenarios possess a distinct hierarchical structure, progressing from physical facilities to instructional media and further to abstract semantic content, with a clear concrete-to-abstract progressive relationship. The general category systems adopted by existing zero-shot segmentation methods struggle to characterize this hierarchical semantic organization, resulting in non-targeted responses in cross-attention maps over semantic categories. To address this, we construct the English Teaching Scene Hierarchical Semantic Category System, named it as the English Teaching Scene Tree (ETST), dividing all semantic concepts into three tiers: the physical environment layer covers eight categories of basic physical facilities including blackboard, podium, desks, projection screen, wall, and window; the instructional media layer contains seven categories of carriers bearing teaching information, such as writing text areas, projection content areas, word cards, textbook pages, exercise books, and electronic screens; the semantic content layer includes six categories of semantic units directly related to core teaching content, including English vocabulary, sentence structures, dialogue bubbles, grammatical annotations, and phonetic symbol annotations. Let the complete category set be $C = C_{L1} \cup C_{L2} \cup C_{L3}$, with a total of 21 categories. The semantic focus involved in different

lesson types varies significantly: vocabulary lessons emphasize vocabulary and phonetic symbol regions, while reading lessons additionally focus on sentence structures and dialogue bubbles. Therefore, the activation status of each category in the model changes dynamically with the scene type rather than being fully enabled. Figure 1 illustrates the teaching-context-driven dynamic prompt encoding and hierarchical attention anchoring mechanism.

To achieve scene-adaptive semantic anchoring, lesson-type discrimination is first performed on the input image. A lightweight EfficientNet-B0 is adopted as the scene classifier f_{scene} to determine the lesson-type category to which the current teaching scene belongs:

$$\hat{t} = \arg \max_{t_k \in T} P(t_k | I), \quad (2)$$

$$T = \{\text{vocabulary, reading, listening \& speaking, grammar}\},$$

where $P(t_k | I)$ is the class probability distribution output by the classifier. This classifier achieves a classification accuracy of 98.7% on the self-built dataset, providing a reliable basis for subsequent dynamic semantic screening. Based on the classification result $\hat{t}$, a semantically relevant subset $C_i^{select} \subseteq C$ is selected from the complete category set, so that the cross-attention of the diffusion model only needs to perform matching within the semantic space corresponding to relevant categories, effectively suppressing response interference from irrelevant semantic regions. Unlike directly using manually set fixed text prompts, this method maps the selected semantic categories into a continuous space via a lightweight adapter to obtain conditional embeddings:

$$C_i = \text{Adapter}(\text{Embed}(C_i^{select})) \in \mathbb{R}^d, \quad (3)$$

where, Embed is a fixed word embedding layer, Adapter consists of a two-layer multilayer perceptron (MLP) with dimension mapping from d to $4d$ and back to d , the value of d takes 768, and the introduced parameter count is approximately 2.4 million. This continuous prompt embedding replaces traditional text descriptions as the cross-modal conditional input to the U-Net of the diffusion model, preserving original category semantic information while gaining tunability in the continuous space.

During the decoding process of the U-Net, the computation method of cross-attention maps at each layer directly affects the spatial localization accuracy of semantic categories. The cross-attention map of the l -th layer decoder is given by:

$$A_l^{cross}(i,j) = \text{Softmax}\left(\frac{Q_l(i,j) \cdot K_l^T(C_i)}{\sqrt{d_k}}\right), \quad (4)$$

where, $Q_l \in \mathbb{R}^{H_l \times W_l \times d_k}$ is the query matrix obtained by linearly projecting image features, $K_l \in \mathbb{R}^{l_{select} \times d_k}$ is the key matrix obtained by key mapping of the conditional embedding C_i , and $d_k = 64$ is the dimension of the key space. The underlying logic of this formulation is to match semantic hierarchies with spatial receptive fields: categories at the L3 semantic content layer, such as English vocabulary and sentence structures, cooperate with self-attention in deeper decoders, whose larger receptive fields can capture semantic associations among word regions scattered across the image; categories at the L1 physical environment layer, such as blackboard edges and podium contours, combine with self-attention in shallower decoders, whose smaller receptive fields benefit the accurate

localization of geometric boundaries of physical facilities. This hierarchical anchoring design allows cross-attention maps to specialize at different scales, avoiding mutual interference between high-level semantics and low-level details under a single scale.

In teaching scenes, certain categories tend to co-occur; for example, blackboards and writing text areas exhibit high co-occurrence, as do projection screens and projection content areas. Let the co-occurrence frequency of category p and category q in the training set be f_{pq} ; then the calibrated cross-attention response is:

$$A_l^{cross,cal}(i,j,p) = A_l^{cross}(i,j,p) \cdot \left(1 + \lambda_{co} \sum_{q \neq p} f_{pq} \cdot I(\hat{y}_q(i,j) > \tau_{conf})\right), \quad (5)$$

where, $\lambda_{co} = 0.3$ is the co-occurrence enhancement coefficient, $\tau_{conf} = 0.5$ is the confidence threshold, $I(\cdot)$ is the indicator function, and $\hat{y}_q(i,j)$ is the current predicted confidence of category q at pixel (i,j) . This calibration takes effect at the pixel level: when a pixel shows a medium-confidence response to category p and a high-confidence response of a highly co-occurring category q exists in its neighborhood, the response of category p is enhanced; conversely, if a region isolatesly responds to a category rarely co-occurring with others in teaching scenes, its confidence is suppressed. This mechanism utilizes the inherent structure of scene semantics as a regularizer, effectively reducing isolated false segmentation caused by local texture ambiguity.

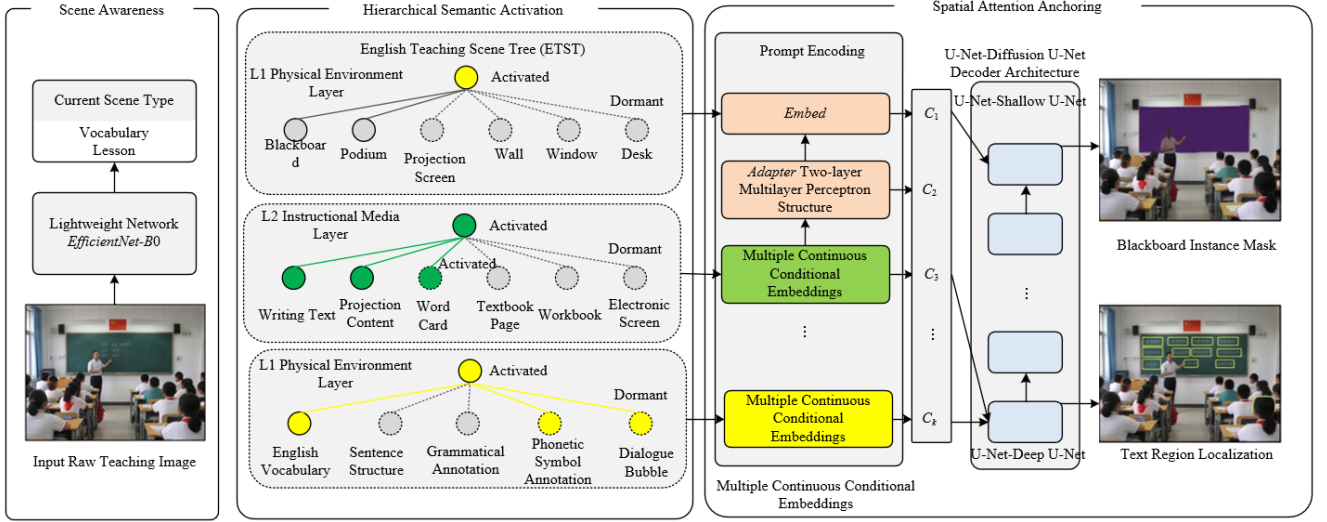


Figure 1. Teaching-context-driven dynamic prompt encoding and hierarchical attention anchoring mechanism

2.3 Spatial-density-driven multi-scale attention topology fusion

Self-attention maps output by different layers of the diffusion U-Net decoder exhibit essential differences in semantic granularity: self-attention maps of shallow decoders have higher spatial resolution and are sensitive to local edges and texture details; self-attention maps of deep decoders undergo multiple downsampling steps, with significantly enlarged receptive fields, capturing overall object shapes and global spatial layouts. Existing methods apply simple averaging or fixed weighted combinations to self-attention maps at all layers, overlooking the spatial-position-dependent variation in the need for detailed versus global information. In English teaching scenarios, character-level boundaries in writing areas require precise localization by shallow high-resolution features, while the distinction between projection content areas and their surrounding environment relies more on semantic context provided by deep features. To this end, self-attention maps $A_l^{self,t_l} \in \mathbb{R}^{H_l \times W_l \times D_l}$ are extracted from all $L=4$ decoder layers of the U-Net. The spatial resolutions of each layer are 64×64 , 32×32 , 16×16 , and 8×8 , with corresponding attention head counts D_l being 8, 8, 16, and 16 in layer order. All self-attention maps are upsampled to a unified resolution $H \times W = 64 \times 64$ via bilinear interpolation:

$$\hat{A}_l = \text{BilinearUpsample}(A_l^{self}) \in \mathbb{R}^{H \times W \times D_l}, \quad (6)$$

So that attention responses at different scales can be compared and fused pixel-wise on the same spatial grid. Figure 2 illustrates the architecture of the spatial-density-driven multi-scale attention topology fusion network, a Differentiable Attention Gating (DAG).

After obtaining multi-level self-attention maps at a unified spatial resolution, the key lies in determining the contribution weight of each layer at every spatial position. Unlike global fixed weights or simple averaging strategies adopted by existing methods, a DAG network is designed to achieve spatial-location-aware dynamic fusion. The network computes fusion weights $\alpha_l(i,j)$ for each decoder layer at each spatial position, whose values are jointly determined by three factors: the intensity of the self-attention response at that position, whether the position belongs to a text-dense region, and the structural similarity of attention maps between adjacent layers. The weights adopt a Softmax-normalized form:

$$\alpha_l(i,j) = \frac{\exp\left(\frac{\phi_l \cdot \bar{A}_l(i,j) + \beta_l \cdot I(D(i,j) > \tau)}{+\gamma_l \cdot \text{Sim}(\bar{A}_l, \bar{A}_{l-1})}\right)}{\sum_{m=1}^L \exp\left(\frac{\phi_m \cdot \bar{A}_m(i,j) + \beta_m \cdot I(D(i,j) > \tau)}{+\gamma_m \cdot \text{Sim}(\bar{A}_m, \bar{A}_{m-1})}\right)}, \quad (7)$$

where, $\bar{A}_l(i,j) = 1/D_l \sum_{d=1}^{D_l} \hat{A}_l(i,j,d)$ is the mean response of the l -th layer self-attention map averaged over the attention-head dimension, reflecting the overall activation intensity of that layer at the position. $D(i,j)$ is the text density factor, τ is the density threshold, and the indicator function $I(D(i,j) > \tau)$ marks whether the pixel lies in a text-dense region. $\text{Sim}(\bar{A}_l,$

$\bar{A}_{l,l-1}$) is the cosine similarity between attention maps of adjacent layers, measuring consistency in their spatial response distributions. $\phi_p, \beta_p, \gamma_l$ are learnable scalar parameters for each layer, controlling the contributions of base response

intensity, text-region enhancement magnitude, and inter-layer consistency constraint, respectively. The entire gating network introduces only $3L = 12$ learnable parameters, achieving adaptive adjustment of fusion weights with almost no increase in model capacity.

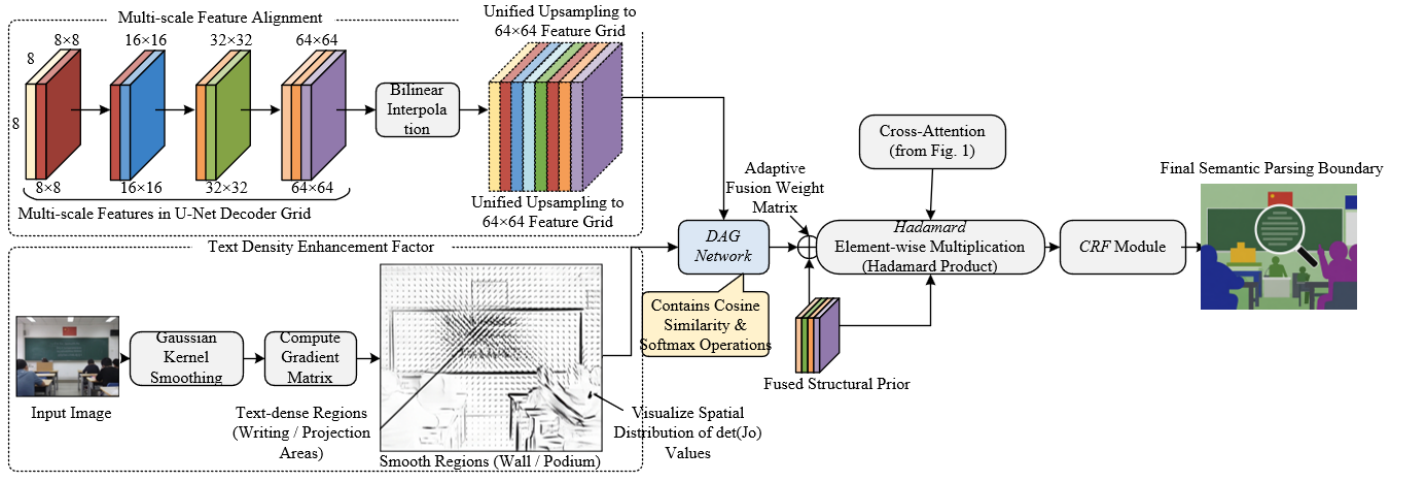


Figure 2. Spatial-density-driven multi-scale attention topology fusion network (Differentiable Attention Gating, DAG)

The dense text arrangement in writing and projection areas of English teaching scenes imposes special requirements on segmentation boundary precision. Edge detection operators commonly used in general scene segmentation methods struggle to effectively distinguish densely arranged character edges from true object boundaries, leading to character adhesion in writing areas and blurred boundaries between projected content and background. A structure-tensor-based text density metric is introduced, encoding texture structure information within each pixel neighborhood into a second-order matrix. For pixel (i,j) , its structure tensor is defined as:

$$J_p(i,j) = G_p * (\nabla I_\sigma(i,j) \nabla I_\sigma(i,j)^T), \quad (8)$$

where, I_σ is the Gaussian-smoothed image with smoothing parameter $\sigma = 0.8$, G_p is a Gaussian kernel with scale parameter $\rho = 1.5$, and $\nabla = (\partial_x, \partial_y)^T$ is the spatial gradient operator. The structure tensor J_p is a positive semi-definite second-order symmetric matrix, whose eigenvalues $\lambda_1(i,j)$ and $\lambda_2(i,j)$ reflect the anisotropic distribution of gradient energy in the pixel neighborhood. The text density factor is defined as the determinant of the structure tensor:

$$D(i,j) = \det(J_p) = \lambda_1(i,j) \cdot \lambda_2(i,j) \quad (9)$$

In text-dense regions, character strokes produce significant gradient responses along multiple directions, keeping both eigenvalues at relatively high levels and raising the determinant response markedly; in smooth regions such as walls or podium surfaces, gradient responses are weak in all directions, and the determinant approaches zero. This metric is rotation-invariant to texture orientation, thus stably responding to English writing oriented horizontally, vertically, or obliquely. In statistics from the self-built dataset, the mean D in writing areas is approximately 8.3 times that of smooth regions, indicating good discriminability of this factor.

Based on the above design, multi-level self-attention maps are fused according to the weights output by the DAG network, and the fusion result is spatially adaptively enhanced by

introducing the text density factor. The fusion process takes the form:

$$M_{fused}(i,j) = \frac{\sum_{l=1}^L \alpha_l(i,j) \cdot \bar{A}_l(i,j) \cdot \min_{(u,v)}(A)}{\max_{(u,v)}(A) - \min_{(u,v)}(A)} \times \left(1 + \gamma \cdot \frac{D(i,j)}{\max_{(u,v)}(D)} \right) \quad (10)$$

The first term is the weighted fused self-attention response min-max normalized to the $[0, 1]$ interval, and the second term is the normalized enhancement factor derived from text density, with $\gamma = 0.3$ as the enhancement coefficient. The mechanism of this formulation is as follows: in text-dense regions such as writing and projection areas, $D(i,j)$ takes larger values, and the fused response is locally amplified, yielding sharper localization of segmentation boundaries; in smooth regions such as walls and podiums, $D(i,j)$ approaches zero, the enhancement factor tends to one, and the fused response is virtually unaffected, preventing spurious responses introduced by over-enhancement in textureless areas. The enhanced self-attention map is then taken as a spatial structural prior and multiplied element-wise with the calibrated cross-attention maps to obtain the final confidence map for each semantic category:

$$M_p^{final} = M_{fused} \odot A^{cross,cal}(p), \quad \forall p \in C_i^{select} \quad (11)$$

where, $\odot$ denotes the Hadamard product. The cross-attention map provides the semantic probability response of a pixel belonging to category p , while the self-attention map provides structural integrity constraints at that location; their product ensures segmentation results are semantically accurate while possessing spatial topological plausibility. The final semantic category of each pixel is determined by taking the maximum response among all activated categories:

$$\hat{y}(i,j) = \arg \max_{p \in C_i^{select}} M_p^{final}(i,j) \quad (12)$$

Pixels not covered by any activated category are labeled as background.

LLaMA-2-7B model:

$$H = LLaMA(S_{graph}, P_{teach}) \quad (16)$$

where, P_{teach} contains instructions requiring the model to infer teaching information such as lesson topic, core knowledge points, and teacher-student interaction focus, and H is the model's output structured teaching semantic description. During this process, all parameters of the LLM remain frozen, and only parameters of the graph attention network and projection layer participate in training. This design effectively controls the number of trainable parameters while retaining the strong reasoning ability of the large model.

The path from segmentation to scene understanding completes an ascending trajectory from pixels to semantics; however, such unidirectional propagation cannot allow errors at the segmentation stage to be perceived and corrected by subsequent modules. To establish a reverse optimization channel, a generative closed-loop feedback mechanism based on image reconstruction is introduced. The inferred teaching semantic description H and segmentation mask M are jointly injected as conditional signals into the decoder of Stable Diffusion to drive the reconstruction of an input image I_{gen} . Specifically, H is converted into a conditional embedding via a fixed text encoder, and the downsampled segmentation mask M is concatenated with feature maps of each U-Net decoder layer along the spatial dimension, serving as a strong constraint on spatial structure. A joint reconstruction loss is constructed by comparing the difference between the reconstructed image and the original input:

$$L_{cyc} = \|I_{gen} - I\|_1 + \lambda_{perc} \cdot \|\Phi(I) - \Phi(I_{gen})\|_2 + \lambda_{ssim} \cdot (1 - SSIM(I, I_{gen})) \quad (17)$$

where, the first term is the pixel-wise L1 reconstruction error, ensuring basic pixel-level fidelity of reconstruction; $\Phi(\cdot)$ is a pre-trained Visual Geometry Group 16-layer (VGG-16) network, with activations of the relu3_3 layer taken as perceptual features, and the second term constrains the distance between reconstructed and original images in perceptual feature space to maintain visual perceptual consistency; the third term adopts the Structural Similarity Index (SSIM) to measure similarity in local contrast and structural information. Weight coefficients λ_{perc} and λ_{ssim} are set to 0.1 and 0.05, respectively.

The core value of this reconstruction loss does not lie in generating realistic teaching images per se, but in its indirect supervision capability over segmentation quality. Its error back-propagation path is:

$$\frac{\partial L_{cyc}}{\partial \Theta_{gate}} \approx \frac{\partial L_{cyc}}{\partial I_{gen}} \cdot \frac{\partial I_{gen}}{\partial M_{fused}} \cdot \frac{\partial M_{fused}}{\partial \Theta_{gate}} \quad (18)$$

where, Θ_{gate} is the parameter set of the DAG network, and M_{fused} is the fused attention map. The computation flow of this link is as follows: the gradient of L_{cyc} with respect to I_{gen} is approximately solved through the cross-attention layers of the generative U-Net, reflecting how sensitively the generated image responds to input conditions; this gradient is then passed to the fused attention map M_{fused} via the dependency of the generator on the segmentation mask, and finally back-tracks through the forward computational graph of the DAG network to the learnable parameters of fusion weights. A key design of

this mechanism is that gradients only update parameters of the DAG network and prompt adapter, while backbone parameters of Stable Diffusion remain frozen. This means that although the reconstruction error originates from the generation process, its optimization objective is solely to adjust the multi-scale attention fusion manner at the segmentation stage, so that segmentation masks better serve subsequent conditional reconstruction. If the reconstruction quality of a local region is markedly lower than surrounding areas, it often implies bias in the spatial constraint provided by the segmentation mask in that region; the error signal guides the DAG network to adjust the relative weights of self-attention across layers at that position, yielding more accurate boundary localization in the next iteration.

The joint optimization objective of the entire GSP-Net is a weighted combination of multiple losses:

$$L_{total} = L_{CE}(M, M_{GT}) + \eta \cdot L_{cyc} + \lambda \cdot \sum_{l=1}^L \|\alpha_l - \alpha_l^{init}\|_2 + \mu \cdot L_{rank} \quad (19)$$

The first term is the cross-entropy segmentation supervision loss, which is active only when labeled data is available; the second term is the closed-loop reconstruction loss, with the weight coefficient η set to 0.5; the third term is a regularization term on attention weights, introduced to constrain the fusion weights α_l from undergoing excessive semantic drift during feedback iterations, with λ set to 0.01 and $\alpha_l^{init} = 1/L$ as the initial uniform weight; the fourth term is the hierarchical ranking loss, defined as:

$$L_{rank} = \sum_{l=1}^{L-1} \max(0, \alpha_{l+1} - \alpha_l + \delta) \quad (20)$$

where, $\delta = 0.05$ is a margin parameter. This loss encourages fusion weights of deeper decoders to be slightly higher than shallower ones; the rationale is that deep self-attention maps possess larger receptive fields and capture overall semantic object structures, thus should dominate fusion, while shallow self-attention maps play a supplementary role only in regions rich in local detail. This ranking constraint maintains a reasonable weight disparity between the two hierarchy types, avoiding deep features from being suppressed by shallow features in extreme cases. In fully annotation-free real teaching scenarios, weights are set as $\eta = 1.0$, $\lambda = 0.1$, and the L_{CE} term is deactivated; at this point, the closed-loop reconstruction loss acts as the primary supervision signal, driving the segmentation module toward self-supervised evolution through continuous interaction in teaching scenes.

3. EXPERIMENTAL DESIGN AND RESULTS ANALYSIS

3.1 Experimental setup

To verify the effectiveness of the proposed method, an English Teaching Scene Dataset (ETSD) is constructed, containing 5000 images collected from real classroom photography, screenshots of online teaching videos, and public teaching resources, uniformly resized to 512×512 resolution. The dataset covers four lesson types: vocabulary, reading, listening & speaking, and grammar, accounting for 27%, 24%, 26%, and 23% respectively. Each image is annotated with pixel-level semantic labels following the 21 categories of the ETST system. The dataset is divided into a training set (3500 images), a validation set (750 images), and

a test set (750 images) at a ratio of 7:1.5:1.5. Annotation work was completed by three graduate students majoring in English education, with an inter-annotator consistency of 92.3% after cross-validation.

For evaluation metrics, the segmentation task adopts mean Intersection over Union (mIoU), Pixel Accuracy (PA), and mean Accuracy (mAcc) as primary indicators. The scene understanding task is evaluated using Scene Graph (SG) Recall@K and teaching semantic reasoning accuracy, where reasoning accuracy is determined via binary judgment of model outputs by three experts in English education. Generated image quality is measured by Fr chet Inception Distance (FID) and Learned Perceptual Image Patch Similarity (LPIPS).

Regarding implementation details, GSP-Net is built upon Stable Diffusion 2.1, with U-Net decoder layers $L = 4$ and inference denoising time steps $T = 20$. Parameters of the DAG network Θ_{gate} are optimized using the AdamW optimizer with a learning rate of 1×10^{-4} , weight decay of 1×10^{-5} , batch size of 8, and trained for 50 epochs. The Adapter adopts the same optimizer with a learning rate of 5×10^{-5} . The enhancement coefficient is set to $\gamma = 0.3$, reconstruction loss weight $\lambda_{perc} = 0.1$, regularization weight $\lambda = 0.01$, and margin parameter $\delta = 0.05$. All experiments are conducted on four NVIDIA A100 GPUs, with a total training time of approximately 36 hours.

Six categories of mainstream models are selected as comparison methods: the zero-shot segmentation baseline

based on diffusion model self-attention maps, DiffSegmenter; the training-free Stable Diffusion mask generation method SeeDiff; the training-agnostic open-vocabulary segmentation method FA-Seg; the multimodal diffusion Transformer attention analysis framework Seg4Diff; the CLIP-based open-vocabulary segmentation method (CLIPSeg); and the in-context learning generalist segmentation model (SegGPT). All comparison methods are re-evaluated on the same ETSD test set using official implementations or author-provided code.

3.2 Semantic segmentation performance comparison

The segmentation performance of GSP-Net and baseline methods on the ETSD test set is compared, and the results are shown in Table 1. GSP-Net achieves an mIoU of 51.8%, representing an improvement of 4.2% over the best baseline Seg4Diff and 6.5% over the best training-agnostic method FA-Seg. In terms of PA and mAcc, GSP-Net also obtains the optimal results, reaching 81.2% and 68.7% respectively. The inference time is 2.73 s per image, slightly higher than DiffSegmenter (2.34 s) and Seg4Diff (1.92 s), mainly due to the extra overhead from text density factor computation and CRF post-processing, which together account for approximately 0.39 s, or 14.3% of the total inference time. Training-agnostic methods generally yield mIoU values below 40%–45% on ETSD, indicating a significant semantic gap exists for general vision-language models in the vertical domain of English teaching scenarios.

Table 1. Comparison of semantic segmentation performance of different methods on the English Teaching Scene Dataset (ETSD) (%)

Method	Publication Year	Training Mode	mIoU	PA	mAcc	Inference Time (s)
CLIPSeg	2025	Training-agnostic	38.2	71.5	52.3	0.42
SegGPT	2025	Training-agnostic	41.7	73.8	55.6	0.68
DiffSegmenter	2025	Training-agnostic	44.1	75.2	58.9	2.34
SeeDiff	2025	Training-agnostic	42.8	74.1	56.7	2.51
FA-Seg	2025	Training-agnostic	45.3	76.4	60.2	0.87
Seg4Diff	2025	LoRA fine-tuning	47.6	78	63.4	1.92
GSP-Net (Ours)	—	Adapter fine-tuning	51.8	81.2	68.7	2.73

Note: mIoU—mean Intersection over Union; PA—Pixel Accuracy; mAcc—mean Accuracy.

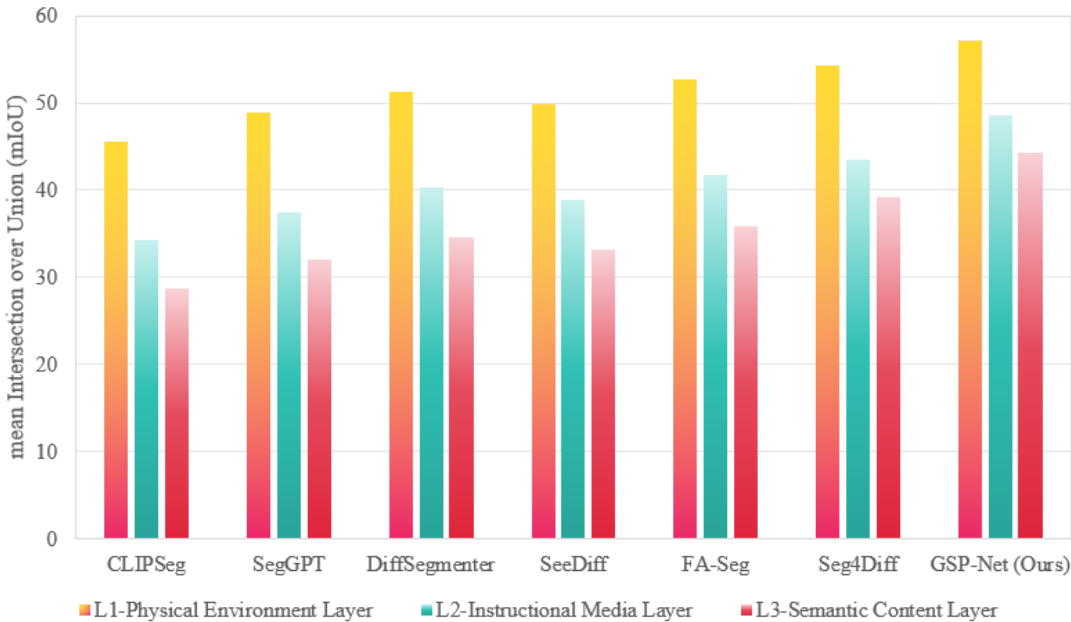


Figure 4. Mean Intersection over Union (mIoU) comparison of different methods across semantic hierarchies (%)

To analyze the performance differences of the proposed method across semantic hierarchies, Figure 4 reports the mIoU of each method on the physical environment layer, instructional media layer, and semantic content layer. All methods achieve the highest mIoU on the physical environment layer and the lowest on the semantic content layer, which aligns with cognitive intuition: physical environment objects possess relatively consistent visual appearances, while semantic content objects exhibit greater visual diversity and blurrier boundaries. The improvement magnitude of GSP-Net across the three layers shows clear divergence: compared with Seg4Diff, the improvement on the semantic content layer is 5.1%, the largest gain; the instructional media layer also sees a 5.1% improvement, while the physical environment layer shows only a 2.8% improvement. This result indicates that the core advantage of the proposed method lies in handling regions of high semantic hierarchy and high text density, verifying the significant contribution of the text density enhancement strategy to segmentation accuracy at the semantic content layer. The mIoU on the instructional media layer is 48.6%, positioned between the physical environment and semantic content layers, reflecting the intermediate nature of the instructional media layer as a bridge between physical carriers and semantic content.

3.3 Ablation study

A series of ablation experiments are designed to verify the effectiveness of each innovation point, and the results are

shown in Table 2. The baseline method adopts fixed text prompts and simple averaging fusion of multi-layer self-attention, achieving an mIoU of 44.1% on the ETSD test set. When scene-aware dynamic prompts are introduced alone, mIoU rises to 46.8%, an increase of 2.7%, indicating that lesson-type-driven semantic category screening effectively narrows the search space of cross-attention, allowing the model to focus on teaching-semantic regions relevant to the current lesson type. When the differentiable attention gating network replaces fixed fusion weights alone, mIoU rises to 47.3%, an increase of 3.2%, verifying the improvement brought by the spatial-location-aware dynamic fusion mechanism to multi-scale self-attention integration quality. Introducing the text density enhancement factor alone raises mIoU to 46.5%, an increase of 2.4%.

The combination of scene-aware prompts and DAG fusion raises mIoU to 49.2%, a 5.1% improvement over the baseline. This gain exceeds the linear expectation of summing their individual contributions, suggesting a synergistic effect between the two: accurate semantic category anchoring provides the gating network with a more discriminative prior of hierarchical attention distribution, thereby amplifying the benefit of dynamic fusion. Further adding text density enhancement pushes mIoU to 50.6%, an additional 1.4% gain on top of scene-aware prompts and DAG fusion. The closed-loop feedback mechanism yields a further 1.2% improvement on the full configuration, bringing mIoU to 51.8%, validating the role of generative reconstruction loss as a self-supervised regularizer promoting segmentation feature learning.

Table 2. Ablation study results on the English Teaching Scene Dataset (ETSD) (mean Intersection over Union, mIoU/%)

Configuration	Scene-Aware Prompt	DAG Fusion	Text Density Enhancement	Closed-Loop Feedback	mIoU	Δ
Baseline (Fixed Prompt + Avg. Fusion)	×	×	×	×	44.1	—
+ Scene-Aware Prompt	✓	×	×	×	46.8	+2.7
+ DAG Fusion	×	✓	×	×	47.3	+3.2
+ Text Density Enhancement	×	×	✓	×	46.5	+2.4
+ Scene-Aware + DAG	✓	✓	×	×	49.2	+5.1
+ Scene-Aware + DAG + Text Enhancement	✓	✓	✓	×	50.6	+6.5
GSP-Net (Full)	✓	✓	✓	✓	51.8	+7.7

Note: DAG—Differentiable Attention Gating; GSP-Net—Generative Scene Parsing Network.

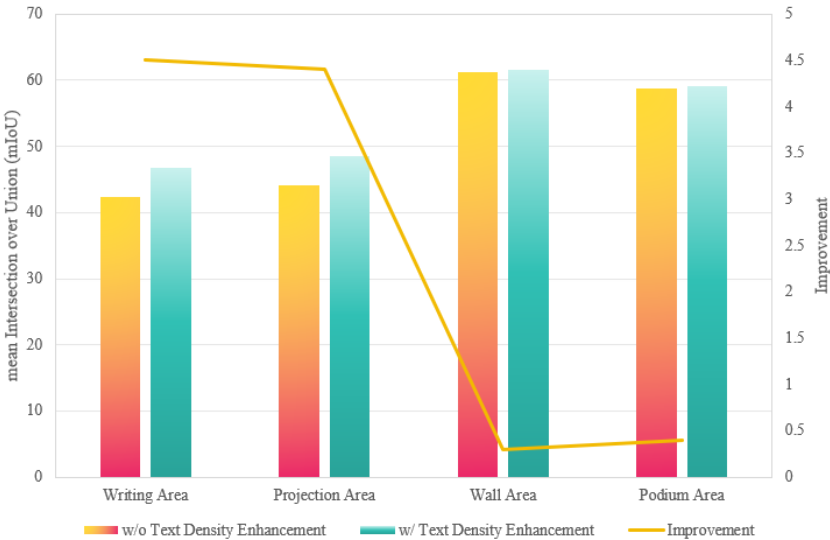


Figure 5. Influence of text density enhancement on segmentation accuracy across different regions (mIoU / %)

To further reveal the specific mechanism of text density enhancement, Figure 5 reports the impact of this module on segmentation accuracy across different regions. Without text density enhancement, the mIoU of writing areas and projection areas is 42.3% and 44.1% respectively, while wall areas and podium areas yield 61.2% and 58.7%. After introducing text density enhancement, the mIoU of writing areas and projection areas rises to 46.8% and 48.5%, gains of 4.5 and 4.4 percentage points, whereas improvements in wall and podium areas are almost negligible, only 0.3 and 0.4 percentage points. This result verifies that the structure-tensor-based text density factor can effectively identify text-dense regions and selectively enhance their attention responses, while preserving original responses in non-text regions. The absolute mIoU of writing areas remains lower than that of projection areas, which may stem from the diversity of handwriting styles and varying writing angles in writing, whereas projection content is usually typeset text with regular fonts.

To verify the fine-grained semantic parsing capability of the proposed method in real English teaching scenarios and its supporting role for subsequent visual scene understanding, Figure 6 presents visualization comparisons among typical classroom images under the original scene, baseline method Seg4Diff, the proposed GSP-Net, and ground-truth annotations. It can be observed that in two representative scenarios—vocabulary lessons and reading lessons—GSP-Net

exhibits higher region completeness and boundary clarity for key semantic objects including projection screens, writing text areas, student subjects, and student subjects. Especially in locally enlarged regions, the proposed method more accurately preserves the structural contours of English vocabulary and reading text, significantly reducing semantic aliasing, boundary adhesion, and local omissions commonly seen in baseline methods. In comparison, although Seg4Diff can produce basic scene segmentation results, it still suffers from category response diffusion, blurred target edges, and insufficient foreground-background separation in text-dense regions; GSP-Net’s performance in these complex regions is already noticeably closer to ground-truth annotations. This result indicates that the semantic modeling mechanism integrating generative AI can effectively improve the collaborative parsing capability for instructional media, semantic content, and scene subjects in English classroom images, enabling the model not only to possess stronger pixel-level segmentation accuracy but also to provide a more reliable structured visual foundation for lesson topic recognition, knowledge point extraction, and teaching behavior understanding, thereby verifying the effectiveness and application potential of the proposed method in English teaching image semantic segmentation and visual scene understanding tasks.



Figure 6. Visualization results of Generative Scene Parsing Network (GSP-Net) semantic segmentation in typical English teaching scenarios

3.4 Scene understanding and generation quality evaluation

The evaluation results of scene graph generation quality are shown in Figure 7. GSP-Net achieves SG Recall@1, @3, and @5 of 41.6%, 61.8%, and 74.3% respectively, representing improvements of 13.2, 15.1, and 16.1 percentage points over the baseline of pure segmentation plus rule-based mapping. The improvement in scene graph generation quality stems mainly from two aspects: the more accurate segmentation results of GSP-Net provide more precise node localization, with significantly higher bounding box accuracy than

comparison methods; the introduction of the graph attention network in SGPNet enables node features to fuse neighborhood information, enhancing the discriminability of node representations. With SG Recall@5 reaching 74.3%, it indicates that when allowing five candidate triplets, GSP-Net can cover most triplets in the ground-truth scene graph.

The expert scoring results for teaching semantic reasoning quality are shown in Table 3. The average score for lesson topic recognition is 4.32, with an accuracy rate of 86.7%, because lesson topics are usually highly correlated with scene type, and the scene classifier already achieves 98.7% accuracy.

The score for core knowledge point extraction is 4.07, with an accuracy rate of 80.0%, indicating that the model can relatively accurately extract key teaching content from the scene graph, yet challenges remain in distinguishing fine-grained knowledge points. The score for teacher-student interaction focus localization is relatively lower at 3.85, with an accuracy rate of 73.3%; this is mainly because interaction focus involves dynamic behavioral information, while the current method performs reasoning solely on single-frame static images and lacks temporal context.

The evaluation results of generated image quality are shown in Figure 8. Direct reconstruction using Stable Diffusion yields an FID of 28.6, indicating a certain distribution gap between generated and original images. After adding the segmentation

mask as a spatial condition, FID drops to 24.1 and LPIPS decreases from 0.312 to 0.278, showing that spatial structural constraints from the segmentation mask effectively improve reconstruction fidelity. With further injection of teaching semantic reasoning results, FID drops to 21.3 and LPIPS to 0.251, indicating that high-order semantic information injection further enhances content consistency of generated images. The value of the generative closed-loop feedback is doubly confirmed here: from an application perspective, semantically enhanced reconstructed images can serve as teaching materials for auxiliary explanation; from a learning perspective, reconstruction errors continuously optimize the segmentation module via the back-propagation path of Eq. (19).

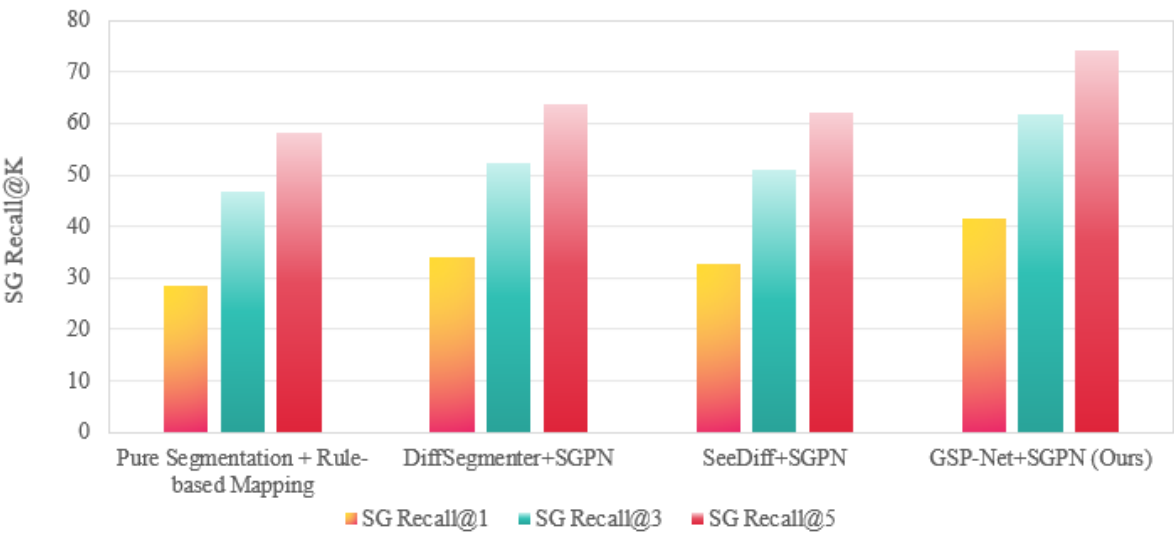


Figure 7. Comparison of scene graph generation performance across methods (SG Recall@K / %)

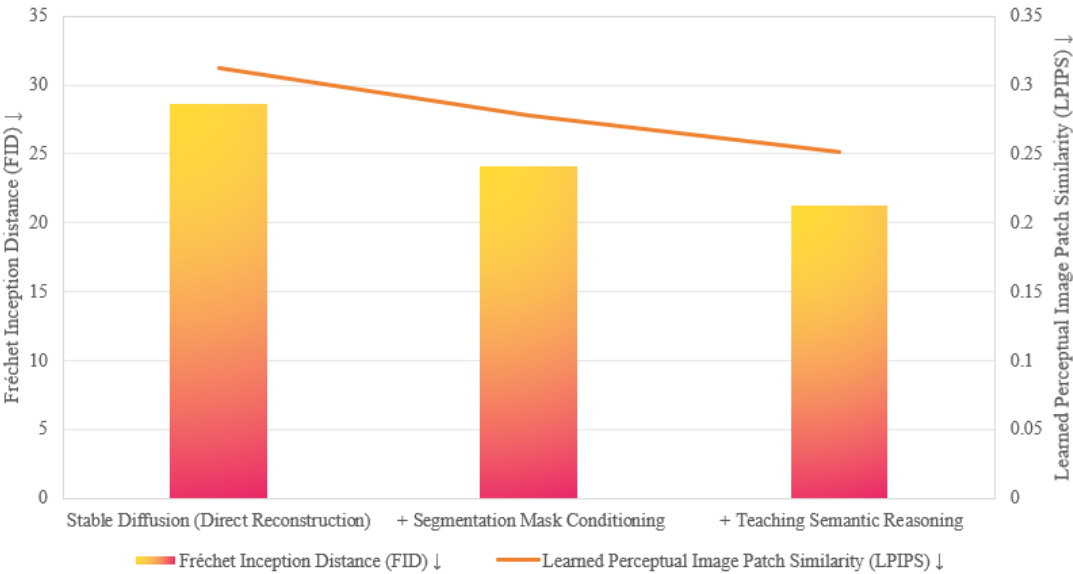


Figure 8. Comparison of generated image quality across methods

Table 3. Expert scoring of teaching semantic reasoning quality

Reasoning Dimension	Average Score	Standard Deviation	Accuracy (Score ≥ 4)
Lesson Topic Recognition	4.32	0.58	86.70%
Core Knowledge Point Extraction	4.07	0.72	80.00%
Teacher-Student Interaction Focus Localization	3.85	0.81	73.30%
Overall Average	4.08	—	80.00%

Table 4. Zero-shot generalization performance across scenarios (mean Intersection over Union, mIoU/%)

Method	Other Subjects	Online Edu. Screenshots	Scanned Textbooks	Average
DiffSegmenter	32.4	38.7	35.2	35.4
SeeDiff	31.2	37.5	33.8	34.2
FA-Seg	34.1	40.2	36.9	37.1
Seg4Diff	36.8	42.5	39.4	39.6
Generative Scene Parsing Network (GSP-Net, Ours)	41.2	48.6	44.7	44.8

Table 5. Comparison of inference efficiency across methods

Method	Inference Time (s/img)	GPU Mem (GB)	Params (M)
CLIPSeg	0.42	4.2	632
SegGPT	0.68	6.8	1,200+
DiffSegmenter	2.34	7.5	860
SeeDiff	2.51	7.8	860
FA-Seg	0.87	5.1	860
Seg4Diff	1.92	7.2	880
Generative Scene Parsing Network (GSP-Net, Ours)	2.73	8.1	865

3.5 Cross-scenario generalization capability

To verify the generalization ability of GSP-Net, zero-shot tests are conducted on three unseen scenarios: teaching scenes of other subjects, screenshots of online education platforms, and scanned textbook images. The results are shown in Table 4. GSP-Net achieves a mIoU of 44.8% across the three cross-scenario tasks, representing a 5.2% improvement over the best baseline Seg4Diff. The best performance is observed in the online education screenshot scenario, with an mIoU of 48.6%, because online education screenshots share the closest visual distribution with teaching video screenshots in the ETSD training data. Performance in other-subject scenarios is relatively weaker at 41.2%; this is because semantic categories in mathematics and physics classrooms are not fully covered by the ETST system, yet categories at the physical environment layer remain universal, thus maintaining acceptable segmentation accuracy. The mIoU on scanned textbook images is 44.7%, positioned between the former two, indicating that categories at the instructional media and semantic content layers of ETST possess certain cross-scenario transferability.

3.6 Efficiency analysis and discussion

Table 5 reports the inference time and GPU memory consumption of each method. The inference time of GSP-Net is 2.73 s per image, higher than FA-Seg and CLIPSeg but broadly comparable to DiffSegmenter. The extra time overhead mainly comes from text density factor computation (~0.21 s), CRF post-processing (~0.18 s), and DAG network forward pass (~0.09 s), totaling approximately 0.48 s. The parameter count of GSP-Net is 865 M, essentially consistent with that of Stable Diffusion 2.1, with the additionally introduced Adapter and DAG parameters accounting for less than 0.5%. GPU memory usage is 8.1 GB, which allows easy deployment on a single A100 GPU, demonstrating feasibility in practical teaching application scenarios.

Although GSP-Net achieves strong performance on the ETSD dataset, several limitations remain. First, segmentation accuracy for extremely dense text still has room for improvement; the current mIoU of writing areas is 46.8%, significantly lower than the average level of 57.1% at the physical environment layer. Second, scene graph construction heavily relies on segmentation quality, so severe segmentation errors propagate to subsequent scene understanding and

generation modules. Third, the current method is based on single-frame static images and lacks modeling capability for temporal dynamic information in teaching scenes. Finally, the scale of the ETSD dataset is still relatively small compared with general segmentation datasets, which may limit the upper bound of the model’s generalization ability. Future work will focus on temporal modeling for video teaching scenarios, construction of larger-scale teaching scene datasets, and more efficient lightweight deployment.

4. CONCLUSION

This paper addresses the tasks of image semantic segmentation and visual scene understanding in English teaching scenarios, and proposes a GSP-Net. Taking a pre-trained diffusion model as the backbone, the method is designed around three core challenges in teaching scenes: clearly hierarchical semantic categories, densely distributed text regions, and the disconnection between perception and reasoning tasks. At the semantic guidance level, a three-tier category system covering physical environment, instructional media, and semantic content is constructed, and a lesson-type-driven dynamic prompt encoding mechanism is introduced to make cross-attention precisely anchor semantic regions relevant to the current teaching context. At the fine-grained segmentation level, a differentiable attention gating network is designed to adaptively fuse multi-level self-attention based on spatial location and local text density, while a structure-tensor-derived text density factor is utilized to enhance boundary responses in writing and projection areas. At the semantic reasoning and self-supervised evolution level, a differentiable scene graph propagation network is constructed to transform segmentation results into structured scene representations and drive a LLM to complete reasoning over lesson topics and knowledge points; furthermore, a closed-loop feedback is established via generative image reconstruction loss, enabling the segmentation module to continuously self-optimize in annotation-free scenarios.

Experiments on the self-built English teaching scene dataset show that GSP-Net achieves an mIoU of 51.8%, a 4.2% improvement over the best baseline, and yields optimal results across the physical environment, instructional media, and semantic content layers, with the most significant gain observed at the semantic content layer. Ablation studies verify

the independent contribution and synergy of each module, with text density enhancement delivering accuracy gains of 4.5 and 4.4 percentage points for writing and projection areas respectively. Scene graph generation recall and expert ratings on teaching semantic reasoning both significantly outperform comparison methods, while generated image quality evaluation further confirms the promoting effect of closed-loop feedback on segmentation feature learning. Cross-scenario generalization experiments demonstrate the adaptability of the proposed method across different teaching environments. This study provides a feasible technical pathway for the in-depth application of generative AI in educational visual understanding. Future work will focus on temporal modeling for teaching videos, construction of larger-scale datasets, and lightweight model deployment.

REFERENCES

- [1] Croitoru, F.A., Hondru, V., Ionescu, R.T., et al. (2023). Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850-10869. <https://doi.org/10.1109/TPAMI.2023.3261988>
- [2] Cao, H., Tan, C., Gao, Z., et al. (2024). A survey on generative diffusion models. *IEEE Transactions on Knowledge and Data Engineering*, 36(7): 2814-2830. <https://doi.org/10.1109/TKDE.2024.3361474>
- [3] Bonechi, S., Andreini, P., Corradini, B.T., et al. (2025). An analysis of pre-trained stable diffusion models through a semantic lens. *Neurocomputing*, 614: 128846.
- [4] Xiao, C., Yang, Q., Zhou, F., Zhang, C.S. (2024). From text to mask: Localizing entities using the attention of text-to-image diffusion models. *Neurocomputing*, 610: 128437. <https://doi.org/10.1016/j.neucom.2024.128437>
- [5] Wang, J., Li, X., Zhang, J., et al. (2025). Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*, 34: 1895-1907. <https://doi.org/10.1109/TIP.2025.3551648>
- [6] Xie, J., Li, W., Li, X., Ong, Y.S., Loy, C.C. (2025). MosaicFusion: Diffusion models as data augmenters for large vocabulary instance segmentation. *International Journal of Computer Vision*, 133(4): 1456-1475. <https://doi.org/10.1007/s11263-024-02223-3>
- [7] Zhang, J., Huang, J., Jin, S., Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5625-5644. <https://doi.org/10.48550/arXiv.2304.00685>
- [8] Shi, H., Dao, S.D., Cai, J. (2025). LLMFormer: Large language model for open-vocabulary semantic segmentation. *International Journal of Computer Vision*, 133(2): 742-759. <https://doi.org/10.1007/s11263-024-02171-y>
- [9] Li, J., Huang, Y., Wu, M., Zhang, B., Ji, X., Zhang, C. (2024). CLIP-SP: Vision-language model with adaptive prompting for scene parsing. *Computational Visual Media*, 10(4): 741-752.
- [10] Cheng, S., Huang, J., Wang, X., Huang, L., Wei, Z. (2025). Image-text aggregation for open-vocabulary semantic segmentation. *Neurocomputing*, 630: 129702. <https://doi.org/10.1016/j.neucom.2025.129702>
- [11] Mo, Y., Wu, Y., Yang, X., Liu, F., Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493: 626-646. <https://doi.org/10.1016/j.neucom.2022.01.005>
- [12] Velastegui, R., Tataarchenko, M., Karaoglu, S., Gevers, T. (2024). Image semantic segmentation of indoor scenes: A survey. *Computer Vision and Image Understanding*, 248: 104102. <https://doi.org/10.1016/j.cviu.2024.104102>
- [13] Wang, S., Wang, F., Zhu, Z., Wang, J.X., Tran, T., Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications*, 252: 124167. <https://doi.org/10.1016/j.eswa.2024.124167>
- [14] Fütterer, T., Goldberg, P., Bühler, B., et al. (2025). Artificial intelligence in classroom management: A systematic review on educational purposes, technical implementations, and ethical considerations. *Computers and Education: Artificial Intelligence*, 9: 100483. <https://doi.org/10.1016/j.caeai.2025.100483>
- [15] Li, Y., Qi, X., Saudagar, A.K.J., Badshah, A.M., Muhammad, K., Liu, S.A. (2023). Student behavior recognition for interaction detection in the classroom environment. *Image and Vision Computing*, 136: 104726. <https://doi.org/10.1016/j.imavis.2023.104726>
- [16] Pang, S., Lai, S., Zhang, A., Yang, Y., Sun, D. (2023). Graph convolutional network for automatic detection of teachers' nonverbal behavior. *Computers and Education: Artificial Intelligence*, 5: 100174. <https://doi.org/10.1016/j.caeai.2023.100174>
- [17] Long, S., He, X., Yao, C. (2021). Scene text detection and recognition: The deep learning era. *International Journal of Computer Vision*, 129(1): 161-184. <https://doi.org/10.1007/s11263-020-01369-0>
- [18] Naiemi, F., Ghods, V., Khalesi, H. (2022). Scene text detection and recognition: A survey. *Multimedia Tools and Applications*, 81(14): 20255-20290. <https://doi.org/10.1007/s11042-022-12693-7>
- [19] Rainarli, E., Suprpto, Wahyono. (2021). A decade: Review of scene text detection methods. *Computer Science Review*, 42: 100434. <https://doi.org/10.1016/j.cosrev.2021.100434>
- [20] Ghosh, M., Mukherjee, H., Obaidullah, S.M., et al. (2023). Scene text understanding: Recapitulating the past decade. *Artificial Intelligence Review*, 56(12): 15301-15373. <https://doi.org/10.1007/s10462-023-10530-3>
- [21] Liu, Q., Jiang, X., Jiang, R. (2025). Classroom behavior recognition using computer vision: A systematic review. *Sensors*, 25(2): 373. <https://doi.org/10.3390/s25020373>
- [22] Mu, S., Cui, M., Huang, X. (2020). Multimodal data fusion in learning analytics: A systematic review. *Sensors*, 20(23): 6856. <https://doi.org/10.3390/s20236856>
- [23] Alfredo, R., Echeverria, V., Jin, Y., et al. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6: 100215. <https://doi.org/10.1016/j.caeai.2024.100215>
- [24] Yürüm, O.R. (2025). Technology-enhanced multimodal learning analytics in higher education: A systematic literature review. *IEEE Access*, 13: 92057-92073. <https://doi.org/10.1109/ACCESS.2025.3572467>
- [25] Guerrero-Sosa, J.D.T., Romero, F.P., Menéndez-Domínguez, V.H., et al. (2025). A comprehensive review of multimodal analysis in education. *Applied Sciences*, 15(11): 5896.

<https://doi.org/10.3390/app15115896>
[26] Chango, W., Lara, J.A., Cerezo, R., Romero, C. (2022).
A review on data fusion in multimodal learning analytics

and educational data mining. Wiley Interdisciplinary
Reviews: Data Mining and Knowledge Discovery, 12(4):
e1458. <https://doi.org/10.1002/widm.1458>