LLM-GUIDED CONTRASTIVE LEARNING FOR CROSS-MODAL MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal information has shown great potential in medical image segmentation, particularly when textual semantic cues are effectively exploited. However, existing cross-modal medical image segmentation methods often suffer from coarse cross-modal fusion and alignment due to two factors: (i) a lack of challenging negative examples for contrastive optimization, and (ii) the neglect of spatial correspondences between textual semantics and local image regions. To this end, we propose a cross-modal medical image segmentation framework via large language model (LLM)-guided contrastive learning (LGCSeg), which enhances representation learning and modality alignment. LGCSeg leverages a multimodal LLM to generate textual descriptions of potential infection regions in input images, which serve as semantic references for identifying relevant hard negative samples from a domain-specific medical text corpus. These samples drive contrastive learning, enhancing the discriminative power of visual-language representations. Furthermore, a multidimensional cross-modal interaction block is introduced to enable semantic interaction between visual and textual features along global, horizontal and vertical spatial dimensions. This design achieves both global and spatial-level alignment, facilitating accurate localization of text semantics within corresponding image regions. Experimental results demonstrate that LGCSeg outperforms the base model by 1.95% in Dice and 2.83% in mIoU on QaTa-COV19 (X-ray), and by 13.04% and 19.44% on MosMedData+ (CT), respectively.

1 INTRODUCTION

Medical image segmentation aims to delineate regions of interest for further disease prediction (Azad et al., 2024), diagnosis (De Fauw et al., 2018) and analysis (Ma et al., 2024). Due to significant patient heterogeneity and the camouflaged lesion regions, segmentation methods relying solely on image information often suffer from limited semantic understanding (Cao et al., 2022b; Ruan et al., 2024; Cheng et al., 2025b). Recently, vision-language modeling has emerged as a promising approach (Huang et al., 2024; Wang et al., 2026; Gui et al., 2026). These methods map text and visual features into a common embedding space for semantic alignment. For example, MAML (Zhang et al., 2021) employs encoders to independently extract textual and visual features, followed by a modality-aware strategy to promote fusion. To achieve more effective image-text feature alignment, many methods further introduce contrastive learning, such as CLIP (Radford et al., 2021) and MERC (Yi et al., 2024), which constrain cross-modal representations and enforce semantic consistency. However, medical images exhibit low contrast, homogeneous textures and blurred lesion boundaries, making precise cross-modal alignment challenging.

The challenges of cross-modal fusion and alignment in medical image segmentation primarily manifest in two aspects. First, the absence of an effective mechanism for selecting hard negative examples during contrastive learning–based image–text matching reduces alignment accuracy. Existing methods such as CLIP (Radford et al., 2021), RecLMIS (Huang et al., 2024) and AMVLM (Pan et al., 2025) typically employ samples from the same batch as negatives. However, prior studies show that hard negative examples, which are capable of confusing positive pairs, are more effective in enhancing contrastive learning (Chen et al., 2020; Li et al., 2023a). Second, cross-modal interaction often overlooks the spatial correspondence between textual semantics and local regions of the image, impeding the accurate spatial localization of textual meanings. For example, STPNet (Shan et al., 2025), DMMI (Hu et al., 2023) and RefSegFormer (Wu et al., 2024) interact with textual tokens

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

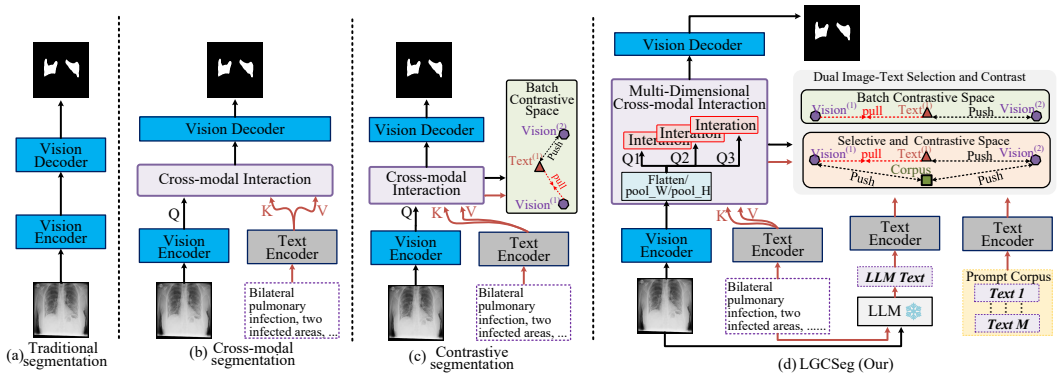


Figure 1: The architecture comparison of different segmentation frameworks. (a) Traditional segmentation relies solely on a single-modal visual stream, leading to limited semantic understanding. (b) Cross-modal segmentation introduces a text encoder and a cross-modal interaction for preliminary semantic fusion, but lacks contrastive learning to enhance modality alignment. (c) Contrastive segmentation with batch-based sampling provides insufficient semantic correspondence between images and text. (d) LGCSeg achieves fine-grained and robust multidimensional interaction. Furthermore, LLM-generated medical descriptions serve as semantic references to construct corpus-level contrastive learning, which complements batch-level contrastive optimization and enhances the model's discriminative capability.

merely by flattening global image features (per pixel or patch). They ignore local positional and directional cues, making it difficult to capture the spatial locations and boundary features of the lesion. Overall, two key questions remain: (1) Can image and text be effectively aligned? (2) Is the alignment sufficiently fine-grained? Motivated by these challenges, this study explores more effective and fine-grained cross-modal alignment and fusion strategies to enhance segmentation performance.

To address the challenge of selecting high-quality hard negative samples in cross-modal contrastive learning, we leverage an LLM to generate challenging textual descriptions based on medical image content and its annotations. These descriptions correspond to potential lesion-evolution regions related to positive samples. Nevertheless, LLM outputs may differ from medical descriptions in format and linguistic style. Inspired by prior work (Shan et al., 2025), we build a standardized textual corpus that encompasses all potential variations present in the infection regions. Based on this corpus, the semantic similarity between the LLM-generated semantics and corpus entries is computed to identify similar yet confusing samples as hard negative examples. These samples are then used to challenge positive pair matching during contrastive learning. To implement this strategy, we propose a dual image–text selective and contrastive (DITSC) block, which not only preserves batch-sampled negative samples but also leverages LLM-generated semantic descriptions to selectively retrieve challenging negative samples from the corpus for enhanced optimization.

Accurate lesion localization requires not only global vision–language alignment but also fine-grained spatial correspondence between textual semantics and image regions. To this end, we propose a multidimensional cross-modal interaction (MDCMI) block. This block establishes vision–language interactions along global, horizontal and vertical dimensions to improve the spatial localization of textual semantics. Specifically, for lateral semantics such as "left/right lung", the model enhances horizontal spatial differentiation through horizontal interactions, whereas vertical interactions strengthen positional understanding for semantics like "upper/middle/lower lung". This design achieves more fine-grained cross-modal correspondence through direction-specific interactions. Furthermore, we incorporate this attention into the image feature decoding process and propose the coordinate attention combined (CAC) block to enhance positional information.

Accordingly, we propose an LLM-guided contrastive learning framework for cross-modal medical segmentation (LGCSeg). It leverages an LLM to guide the selection of hard negative samples and establishes semantic interactions between image structures and textual descriptions across global, horizontal and vertical dimensions, thereby achieving effective contrastive matching and fine-grained cross-modal information fusion. Fig. 1 illustrates that LGCSeg demonstrates substantial improvements compared to existing frameworks. The primary contributions of this study are as follows:

- We propose LGCSeg for cross-modal medical image segmentation, which utilizes an LLM to infer potential lesion evolution and provide an informative reference for contrastive sample matching.

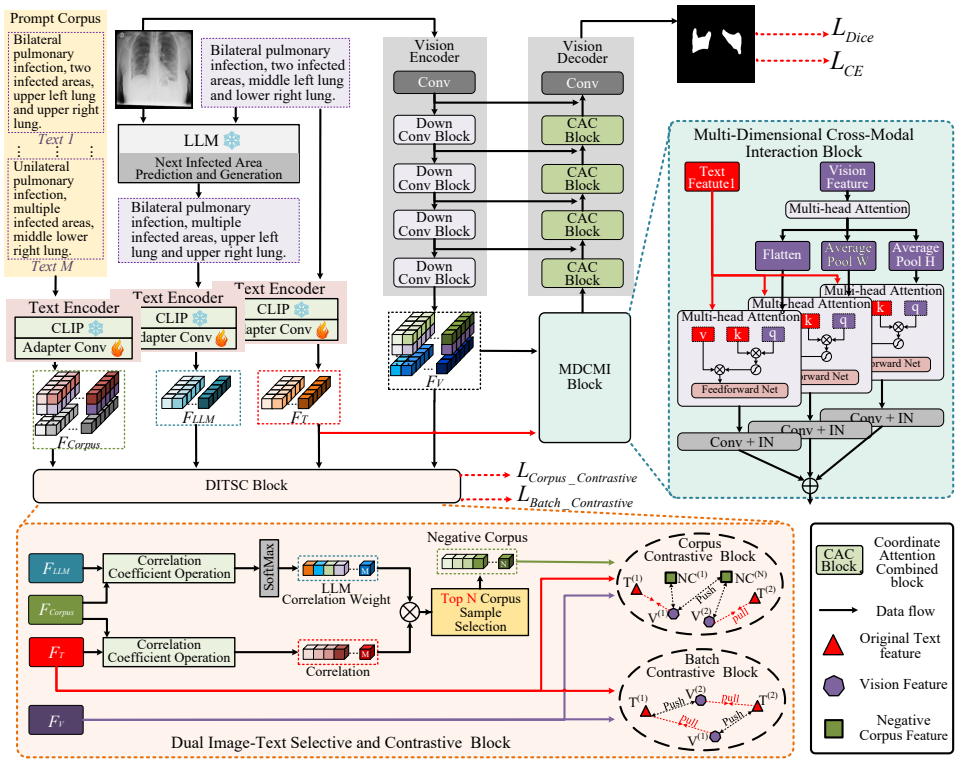


Figure 2: The architecture of the LGCSeg method.

- A DITSC block is proposed to achieve effective semantic alignment between visual and textual features. It incorporates LLM-generated medical descriptions to guide hard negative sample selection, reducing sampling randomness and improving cross-modal alignment robustness.
- We design an MDCMI block to perform semantic interactions between images and text along global, horizontal and vertical dimensions, enhancing the spatial localization of textual semantics within the visual space through direction-specific constraints.
- A CAC block is proposed to enhance positional information. Experiments on QaTa-COV19 and MosMedData+ demonstrate that LGCSeg delivers superior segmentation performance.

2 RELATED WORK

Medical Image Segmentation. Neural networks have made remarkable advances in medical image analysis (Azad et al., 2024; Zhou et al., 2024). For example, the UNet architecture (Ronneberger et al., 2015) features an encoder–decoder with skip connections that enhance segmentation accuracy; UNet++ (Zhou et al., 2018) enhances multi-scale dense connections; AttUNet (Oktay et al., 2018) incorporates attention gate mechanisms; and nnUNet (Isensee et al., 2021) introduces adaptive configuration and training strategies. Owing to its self-attention mechanism for modeling global context, the Vision Transformer (ViT) has been adopted across diverse computer vision applications. Chen et al. (Chen et al., 2021) integrated ViT and proposed TransUNet for segmentation tasks. Wang et al. (Wang et al., 2022) incorporated a channel Transformer in segmentation. To mitigate the computational complexity of ViT, researchers have proposed several frameworks. Swin-Unet (Cao et al., 2022a) adopts a shifted window attention mechanism for modeling. VM-Unet (Ruan et al., 2024) and Mamba-Sea (Cheng et al., 2025a) incorporate Mamba models to perform state-space-based modeling. However, these methods primarily focus on image information alone, while medical imaging faces challenges such as low imaging quality, significant inter-patient variability and high lesion camouflage. These factors hinder further improvements in segmentation performance.

Cross-Modal Medical Image Segmentation. Segmentation methods that rely solely on visual information are often vulnerable to image quality variations and limited by insufficient semantic understanding. With the growing impact of multimodal learning, vision-language modeling has emerged as a promising paradigm (Boecking et al., 2022; Stevens et al., 2024). A key aspect of this paradigm

is the effective integration and alignment of cross-modal information. For cross-modal fusion, Li et al. (Li et al., 2023b) employed a BERT-based model for text embedding and introduced pixel-level attention to achieve multimodal fusion and segmentation. Zhang et al. (Zhang et al., 2024) designed an adapter module to achieve superior fusion and decoding performance. Rahman et al. (Rahman et al., 2025) integrated self- and cross-modal attention to enhance multimodal information fusion. In alignment strategies, contrastive learning has emerged as an effective approach. Radford et al. (Radford et al., 2021) presented CLIP, a large-scale vision-language pretraining framework that achieves image–text matching via contrastive learning. Chen et al. (Chen et al., 2024) incorporated causal intervention into the CLIP model. Li et al. (Li et al., 2024) proposed knowledge-guided contrastive learning. In addition, several studies have integrated cross-modal fusion with contrastive learning. Hu et al. (Hu et al., 2023) introduced contrastive learning for text–vision feature alignment. Huang et al. (Huang et al., 2024) employed conditional reconstruction with contrastive learning. Shan et al. (Shan et al., 2025) built a predefined textual corpus to facilitate text retrieval via contrastive matching. Despite advancements in multimodal fusion and alignment, most cross-modal segmentation methods overlook local directional cues. Additionally, random negative sampling in contrastive learning can weaken representation discrimination.

3 METHODS

To further enhance multimodal information interaction and alignment, we propose LGCSeg, whose overall architecture is illustrated in Fig. 2.

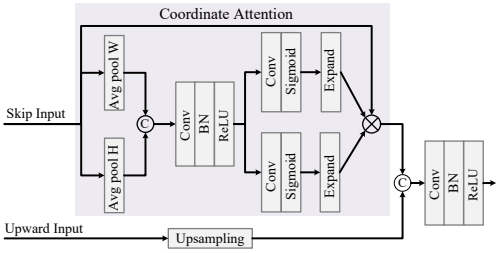


Figure 3: Architecture of the CAC block.

3.1 LLM-BASED TEXT GENERATION

A frozen LLM (GPT-4o mini) (Achiam et al., 2023) was employed to generate challenging textual descriptions, leveraging its multimodal reasoning and generation capabilities. We leveraged the LLM's application programming interface (API) to produce textual outputs from paired text-image inputs (More details can be found in Appendix A). The prompt used was as follows: "The image description is T, which is the confirmed infected area. From a medical perspective, please predict the areas in the image that may be infected in the next stage but have not been described. Please answer in the following manner: no additional information is required. If 'and' is used, both the left and right lungs must be described." This step can be formulated $T_{LLM} = LLM(T, img)$, where T is original textual description and $T \in \mathbb{R}^L$. Img is original image and $img \in \mathbb{R}^{H \times W \times 3}$. The final infection area labels were integrated from all original textual descriptions.

3.2 CORPUS TEXT GENERATION

Inspired by STPNet (Shan et al., 2025), we structurally decomposed the description of the lung infection into four semantic components: infection type, number of infected areas, left lung infected areas and right lung infected areas. Specifically, the infection types include "Bilateral/Unilateral pulmonary infection." The number of infected areas is categorized as "one/two/three/four/multiple infected areas." The left lung infection areas comprise "upper left lung," "lower left lung," "middle left lung," "upper middle left lung," "middle lower left lung," "upper middle lower left lung," and "all left lung," while those of the right lung follow a similar naming convention. This design is guided by the inherent left–right division of pulmonary anatomy and the empirical distribution of infection counts observed in COVID-19 cases (Li et al., 2023b). By integrating the four components, we constructed a comprehensive corpus of medical descriptions representing various pulmonary infection scenarios.

3.3 FEATURE ENCODER-DECODER

During the text encoding phase, LGCSeg employs a frozen CLIP text encoder (Radford et al., 2021) combined with an adapter convolution layer to form the text encoder. The original, LLM-generated and corpus texts are each input into this encoder $En_t(\cdot)$ to generate corresponding text feature representations F_T, F_{LLM} and F_{corpus}. The visual network model consists of a visual encoder and decoder. Its design adopts a U-shaped architecture, enabling effective multi-scale feature extraction.

This visual-coding process includes 3×3 convolution and down-sampling operations, which can be computed as $F_{Di} = DConv_t(F_{D(i-1)})$, where $DConv$ is defined as a down-convolution block that applies a 3×3 convolution and downsamples the feature map by a factor of 2.

The visual decoder reconstructs the fused image–text features to generate segmentation results. As depicted in Fig. 3, the decoder is constructed using multiple CAC blocks. Each CAC block receives two types of inputs: one from the encoder and the other from the output of the preceding CAC block. The skip-connected features obtained from the encoder are processed through a coordinate attention mechanism to enhance spatial attention representations along both horizontal and vertical directions. Simultaneously, features from the preceding block are upsampled and concatenated with the encoder stream along the channel dimension, followed by a convolution to yield the CAC block output, formulated as follows:

$$Tp_{Ui} = Concat(Upsample(F_{U(i-1)}), CA(F_{Di})), i < 5 \tag{1}$$

$$D_{ui} = Conv(Tp_{ui}) \tag{2}$$

$$Mask = Softmax(Conv(D_{u4})) \tag{3}$$

where $Upsample$ represents Up-sampling, $Concat$ represents channel concatenation and Mask is a segmentation mask.

To improve the precision of aligning textual semantic information with image spatial features, we adopt a multidimensional interaction strategy. Inspired by the directional feature encoding of the coordinate attention (Hou et al., 2021), the MDCMI block leverages a multi-head attention mechanism to interact with textual information across three spatial dimensions of image features: global, horizontal and vertical. This design aims to facilitate finer-grained directional information fusion and alignment. As illustrated in Fig. 2, the block receives feature encodings F_T and F_{img} from T and img via the LGCSeg's feature encoder. To further enhance the semantic relevance of visual features, a multi-head attention is applied to image features F_{img} for global context enrichment. The multi-head attention mechanism is formulated as follows:

$$out = Softmax(\frac{(W_q F_q)^T (W_k F_k)}{\sqrt{\hat{C}}})(W_v F_v) \tag{4}$$

where W_q, W_k and W_v are learnable parameters. When using self-attention, $F_q, F_k, F_v = F_{img}$. After applying the multi-head mechanism to F_{img}, the result $F^{'}_{img}$ is obtained. Subsequently, $F^{'}_{img}$ is processed through global flattening, horizontal pooling followed by flattening, and vertical pooling followed by flattening to extract global information F^{G}_{img}, horizontal information F^{H}_{img} and vertical information F^{W}_{img}, respectively. During feature interaction, we employ the multi-head attention module to integrate image and text features. Notably, the image features at this stage are derived not only from globally flattened representations but also from features processed via horizontal and vertical pooling, thereby enabling more precise spatial localization of textual descriptions within images. The multi-head attention process can be formalized as Eq. (4), where F_T is projected via linear layers to yield F_q and F_k. F^{G}_{img}, F^{H}_{img} and F^{W}_{img} each serve as F_V.

3.4 DITSC BLOCK

LGCSeg employs contrastive learning to match image and text features by measuring their semantic similarity. Specifically, we adopt the Gaussian kernel (Xiao et al., 2014; Carreira-Perpinan, 2007) to compute the similarity between image and text. This choice is motivated by the Gaussian kernel's ability to flexibly capture the nonlinear feature similarity through its adjustable bandwidth parameter, compared to linear or Laplace kernels (Pérez-Cruz & Bousquet, 2004; Zhang et al., 2018). This property aligns better with the assumption of semantic association, where distances between features should smoothly reflect their semantic similarity, thereby drawing positive samples together and pushing negative samples apart. The feature similarity calculation is as follows:

$$K(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \tag{5}$$

where x and y are two different samples, σ is bandwidth and used to control smoothness. Most existing image–text matching approaches generate negative sample pairs only within the mini-batch (Huang et al., 2024; Hu et al., 2023; Radford et al., 2021; Stevens et al., 2024), constraining the diversity and informativeness for learning discriminative semantic boundaries. We address this limitation with the DITSC block, which preserves batch sampling while introducing corpus-level contrast. It utilizes semantically ambiguous or "hard" medical descriptions generated by the LLM

Table 1: Comparison with state-of-the-art methods. The ARSegNet, LViT, ReCLMIS, and Baseline methods were retrained following the experimental settings described in their original papers. The results of STPNet are cited from its original publication. The others are referenced from LViT and ReCLMIS.

Method	Backbone	Description	QaTa-COV19		MosMedData+	
			Dice(%)↑	mIoU(%)↑	Dice(%)↑	mIoU(%)↑
UNet (Ronneberger et al., 2015)	CNN	×	79.02	69.46	64.60	50.73
UNet++ (Zhou et al., 2019)	CNN	×	79.62	70.25	71.75	58.39
nnUNet (Isensee et al., 2021)	CNN	×	80.42	70.81	72.59	60.36
TransUNet (Chen et al., 2021)	Transformer	×	78.63	69.13	71.24	58.44
SwinUnet (Cao et al., 2022b)	Swin Transformer	×	78.07	68.34	63.29	50.19
UCTransNet (Wang et al., 2022)	Hybrid	×	79.15	69.60	65.90	52.69
STPNet (Shan et al., 2025)	Hybrid	×	80.63	71.42	**76.18**	**63.41**
ConVIRT (Zhang et al., 2022)	CNN	✓	79.72	70.58	72.06	59.73
TGANet (Tomar et al., 2022)	CNN	✓	79.87	70.75	71.81	59.28
CLIP (Radford et al., 2021)	Transformer	✓	79.81	70.66	71.97	59.64
LAVT (Yang et al., 2022)	Transformer	✓	79.28	69.89	73.29	60.41
RefSegformer (Wu et al., 2024)	Transformer	✓	84.09	75.48	74.98	61.70
ARSegNet (Wang et al., 2026)	Hybrid	✓	83.23	71.72	70.14	54.60
LViT (Li et al., 2023b)	Hybrid	✓	83.38	74.77	74.17	61.02
ReCLMIS (Huang et al., 2024)	CNN	✓	**84.86**	**76.69**	74.63	61.50
Baseline	CNN	✓	83.73	75.07	68.39	54.23
LGCSeg	Hybrid	✓	**85.36**	**77.20**	**77.31**	**64.77**

to compute semantic correlation weights with an external corpus. In Fig. 2, the weights serve as references for retrieving semantically similar negative samples from external corpora that align with the original input medical descriptions. By incorporating these challenging negative pairs, LGCSeg can enhance cross-modal discriminative capability in complex semantic scenarios. The contrastive learning processes within the batch can be represented as follows:

$$\mathcal{L}_{Batch} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(K(F_{img_i}, F_{T_i})/\tau\right)}{\sum_{j=1}^{B}\exp\left(K(F_{img_i}, F_{T_j})/\tau\right)} \tag{6}$$

where K represents calculation of the Gaussian kernel, τ is the temperature hyperparameter. The corpus contrastive learning process is illustrated in Eq. (7)–(10).

$$K^m(F_{LLM}, F^{(m)}_{Corpus}) = \exp\left(-\frac{\left\|F_{LLM} - F^{(m)}_{Corpus}\right\|^2}{2\sigma^2}\right) \tag{7}$$

where $m = 1, \ldots, M$. The top N most similar samples are selected from all candidates.

$$W_{LLM} = Softmax\left(K^m(F_{LLM}, F^{(m)}_{Corpus})\right) \tag{8}$$

$$F_{Corpus_N} = \text{TopN}\left(W_{LLM} \cdot K^m(F_T, F^{(m)}_{Corpus})\right) \tag{9}$$

$$\mathcal{L}_{Corpus} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp\left(K(F_{img_i}, F_{T_i})/\tau\right)}{\exp\left(K(F_{img_i}, F_{T_i})/\tau\right) + F_N} \tag{10}$$

where $F_N = \sum_{j=1}^{N}\exp\left(K(F_{img_i}, F_{Corpus_j})/\tau\right)$.

3.5 Training Object

The objectives of LGCSeg during training include the Dice loss $\mathcal{L}_{Dice}$, Cross Entropy loss $\mathcal{L}_{CE}$, batch contrastive loss $\mathcal{L}_{Batch}$, and corpus contrastive loss $\mathcal{L}_{Corpus}$. The $\mathcal{L}_{Batch}$ and $\mathcal{L}_{Corpus}$ calculations are as in Eqs. (6) and (10). The calculation methods for $\mathcal{L}_{Dice}$ and $\mathcal{L}_{CE}$ are as follows:

$$\mathcal{L}_{Dice} = 1 - \sum_{i=1}^{N}\sum_{j=1}^{C}\frac{1}{NC}\cdot\frac{2|p_{i,j} \cap m_{i,j}|}{(|p_{i,j}| + |m_{i,j}|)} \tag{11}$$

$$\mathcal{L}_{CE} = -\sum_{i=1}^{N}\sum_{j=1}^{C}\frac{1}{N}\cdot m_{i,j}\cdot log(p_{i,j}) \tag{12}$$

where N is the number of pixels, C is the number of categories, which is set to 1. $p_{i,j}$ represents the probability predicted by LGCSeg that pixel i belongs to category j. $m_{i,j}$ represents the mask of ground truth. In summary, the training objective for LGCSeg is as follows:

$$\mathcal{L}_{Net} = \lambda_1 \cdot (\mathcal{L}_{Dice} + \mathcal{L}_{CE}) + \lambda_2 \cdot \mathcal{L}_{Batch} + \lambda_3 \cdot \mathcal{L}_{Corpus} \tag{13}$$

where λ_i is hyperparameters.

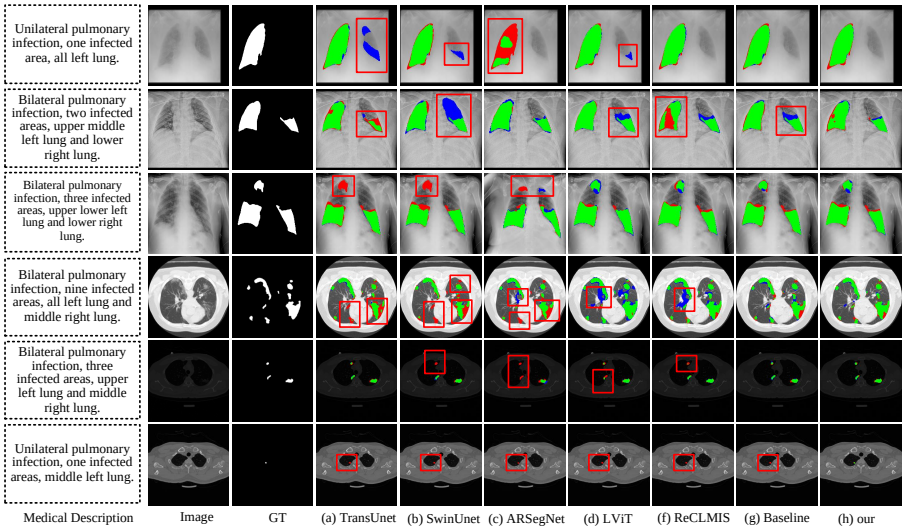


Figure 4: Visualization of segmentation results. Green areas indicate correct segmentation, blue indicates oversegmentation, and red indicates undersegmentation.

4 EXPERIMENTS

Due to space limitations, dataset and implementation details are provided in Appendix.

4.1 METHOD COMPARISON

Tab. 1 summarizes the results. LGCSeg consistently outperforms competing methods on Dice and mIoU metrics. On the QaTa-COV19 dataset, it surpasses STPNet, which lacks medical description inputs, by 5.87% in Dice and 8.09% in mIoU. Compared to ReCLMIS, which uses medical descriptions, LGCSeg achieves additional gains of 0.59% and 0.67% in Dice and mIoU, respectively. Over the baseline, it further improves Dice by 1.95% and mIoU by 2.83%. On the MosMedData+ dataset, LGCSeg outperformed the second-best method, STPNet, achieving a 1.25% improvement in Dice and a 2.14% increase in mIoU. Compared with the baseline, LGCSeg achieved substantial gains of 13.04% in Dice and 19.44% in mIoU. These results underscore the effectiveness of incorporating medical description information, as well as the proposed DITSC, MDCMI and CAC blocks. These components significantly enhance lesion recognition and segmentation accuracy.

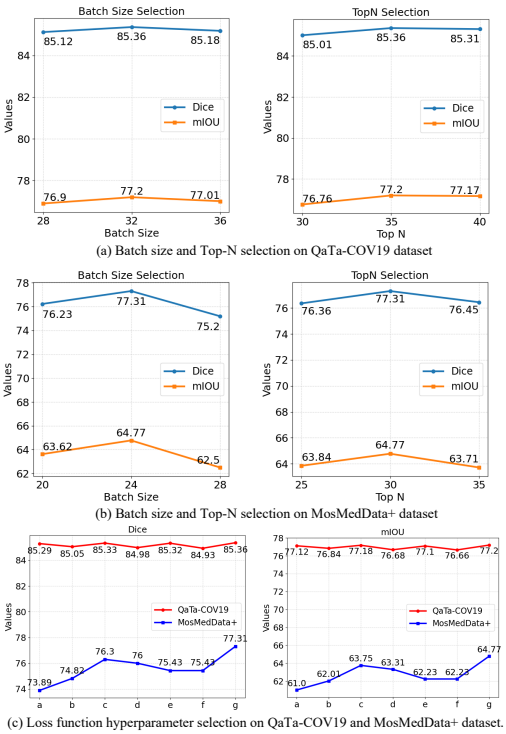


Figure 5: Segmentation results of loss function hyperparameter selection. (a) $\lambda_1 = 4, \lambda_2 = 0.1, \lambda_3 = 0.1$; (b) $\lambda_1 = 6, \lambda_2 = 0.1, \lambda_3 = 0.1$; (c) $\lambda_1 = 5, \lambda_2 = 0.05, \lambda_3 = 0.1$; (d) $\lambda_1 = 5, \lambda_2 = 0.5, \lambda_3 = 0.1$; (e) $\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.05$; (f) $\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.5$; (g) $\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.1$.

Fig. 4 depicts the qualitative segmentation outcomes on the two datasets. The top three rows display segmentation outcomes from the QaTa-COV19 dataset, while the bottom three rows correspond to the MosMedData+ dataset. The visual comparisons clearly demonstrate that LGCSeg outperforms other competing methods. When confronted with large or multiple lesion regions, alternative methods tend to exhibit under-segmentation or over-segmentation issues (as shown in Fig. 4(a)-(d)), whereas LGCSeg yields more accurate and complete segmentation results. This indicates its superior capability in handling randomly distributed lesions with irregular morpholo-

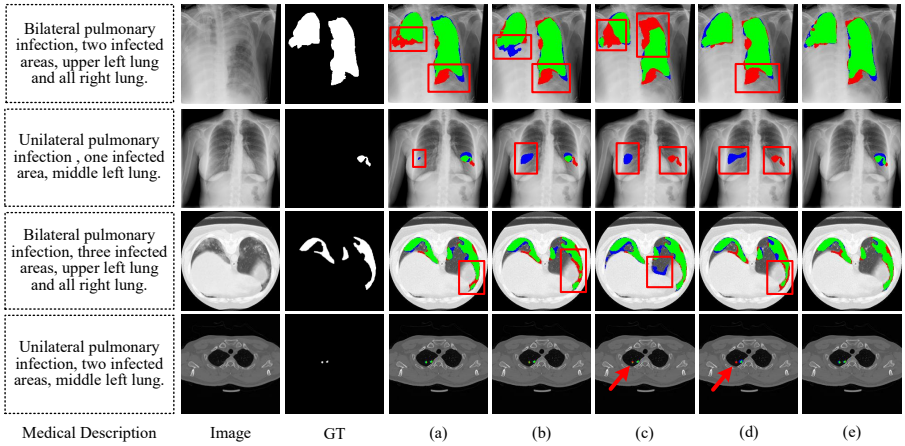


Figure 6: Visualization of segmentation results for ablation of LGCSeg. (a) w/o Batch Contrastive; (b) w/o Corpus Contrastive; (c) w/o MDCMI block; (d) w/o CAC block; (e) LGCSeg.

gies. Furthermore, when segmenting small lesions, LGCSeg achieves more precise delineation than other approaches, effectively alleviating the problem of lesion camouflage. Compared with baseline methods, LGCSeg also produces finer and more accurate lesion boundaries (e.g., Fig. 4(g), second, fourth, and sixth rows), demonstrating that the enhanced semantic contrast alignment and refined cross-modal fusion between textual and visual features contribute to more precise lesion localization and segmentation.

4.2 Parameter Sensitivity Study

Batch Size and Top-N Hyperparameter Selection. We performed ablation studies to investigate the effects of batch size and the Top-N hyperparameter. As illustrated in Fig. 5(a), experiments on the QaTa-COV19 dataset show that both Dice and mIoU reach their peak values at a batch size of 32. Fig. 5(b) shows that LGCSeg achieves optimal performance with a batch size of 24 on MosMedData+ dataset. The Top-N hyperparameter controls the number of top-ranked sample features from the corpus that are most similar to the original and LLM-generated medical description. For QaTa-COV19, segmentation accuracy is highest when $N = 35$. LGCSeg attains its best results when N is set to 30 on MosMedData+ dataset.

Loss Function Weights Selection. The experiment investigates three weighting coefficients in the training objective, namely λ_1, λ_2 and λ_3. In Fig. 5(c), the Dice and mIoU metrics are reported for seven different settings of loss function weights. Following the principle of selecting the configuration that delivers the overall best performance, the weights are set to $\lambda_1 = 5, \lambda_2 = 0.1$ and $\lambda_3 = 0.1$.

4.3 Ablation Experiments

Table 2: Results of ablation experiment.

Method	QaTa-COV19		MosMedData+	
	Dice(%) ↑	mIoU(%) ↑	Dice(%) ↑	mIoU(%) ↑
(a) w/o Batch	85.12	76.97	76.10	63.36
(b) w/o Corpus	85.16	76.98	76.68	63.98
(c) w/o MDCMI	84.78	76.53	75.13	62.30
(d) w/o CAC	84.65	76.19	68.88	54.69
(e) LGCSeg	**85.36**	**77.20**	**77.31**	**64.77**

DITSC Block and Contrastive Learning Analysis. Fig. 6 shows the ablation results evaluating the effectiveness of batch sampling and negative sample selection in corpus-based contrastive learning. The segmentation results without the contrastive block exhibit noticeable under-segmentation (red lesion areas) and over-segmentation (blue lesion areas). In contrast, incorporating the contrastive block significantly improves performance, as demonstrated in Fig. 6(e), effectively reducing both under- and over-segmentation. Tab. 2 presents the quantitative evaluation results. LGCSeg achieves the highest Dice and mIoU scores, indicating that the incorporation of the DITSC block effectively enhances image-text correspondence and thereby improves segmentation performance, particularly on the MosMedData+ dataset.

In the DITSC block, an LLM is used to generate reference medical descriptions, which, together with the original input sample, guide the selection of challenging cases from the corpus. As shown in Fig.7, the Top-N challenging samples are selected as negative samples during corpus-level contrastive training. The selected Top-N samples, influenced by both the original input and the LLM-

generated medical description, are semantically close to the input, helping the model improve semantic boundary discrimination during training.

MDCMI Block and Attention Analysis. Fig. 6 and Tab. 2 present the qualitative and quantitative evaluation results, respectively. Fig. 6(c) shows that removing the MDCMI block causes under-segmentation in large lesion areas (highlighted in red results). Furthermore, for small lesion areas, the model exhibits segmentation errors (the fourth result in Fig. 6(c)). These issues likely arise from suboptimal fusion and misalignment between textual and visual semantic features. By contrast, LGCSeg achieves more precise localization of textual semantics within images through multi-dimensional interactions with image features, including horizontal, vertical, and global dimensions. This mechanism significantly enhances feature integration and alignment, yielding superior segmentation accuracy in both subjective and objective evaluations.

Fig. 8 compares attention maps of LGCSeg using multidimensional interactions with those using traditional global single-dimensional interactions. As shown in Fig. 8, the traditional method can focus on the target vicinity but fails to achieve fine-grained alignment. In contrast, the proposed multidimensional interactions across global, horizontal and vertical dimensions enable more precise integration of image and text features, resulting in more accurate attention and sharper focus on the target regions.

CAC Block Analysis. Fig. 6 and Tab. 2 present the validation results of the CAC block. The CAC block is incorporated into the visual decoder network to better capture fine-grained details in image features. As shown in Fig. 6(d)(e), models without the coordinate attention of the CAC block are prone to both over- and under-segmentation. Quantitative results in Tab. 2 further indicate that removing this coordinate attention from the block leads to a noticeable drop in segmentation performance, including reductions in Dice and mIoU. These results highlight the CAC block's role in enhancing the network's attention to subtle structural details.



Bilateral pulmonary infection, two infected areas, middle lower left lung and middle lower right lung.

Original description

Bilateral pulmonary infection, multiple infected areas, upper left lung and upper right lung.

LLM generation

（1）Bilateral pulmonary infection, two infected areas, lower left lung and lower right lung.（0.0061）
（2）Bilateral pulmonary infection, two infected areas, middle lower left lung and lower right lung.（0.0057）
（3）Bilateral pulmonary infection, two infected areas, lower left lung and middle lower right lung.（0.0054）
（4）Bilateral pulmonary infection, two infected areas, lower left lung and all right lung.（0.0053）
（5）Bilateral pulmonary infection, two infected areas, middle lower left lung and all right lung.（0.0052）
（6）Bilateral pulmonary infection, two infected areas, middle lower left lung and upper middle lower right lung.（0.0051）
（7）Unilateral pulmonary infection, one infected area, lower right lung.（0.0050）
（8）Bilateral pulmonary infection, two infected areas, all left lung and lower right lung.（0.0049）
（9）Bilateral pulmonary infection, two infected areas, middle lower left lung and middle right lung.（0.0049）
（10）Bilateral pulmonary infection, two infected areas, lower left lung and middle right lung.（0.0049）

The Top-N negative samples selected based on LLM intervention and their similarity

Figure 7: LLM-generated descriptions and corresponding Top-N hard-negative samples.

Medical Description: "Bilateral pulmonary infection, two infected areas, lower left lung and lower right lung."

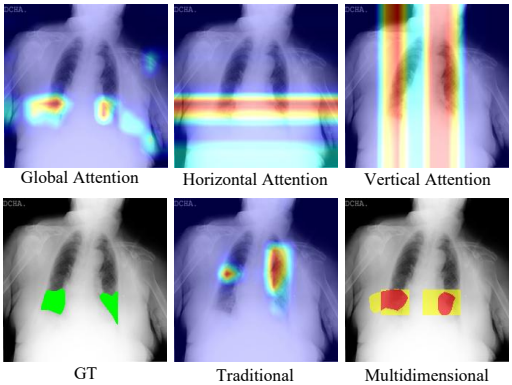


Figure 8: Visualization of Attention in Traditional Interaction and Multidimensional Interaction. The red regions indicate significant attention interactions among global, horizontal and vertical attention.

5 Conclusion

This work presents LGCSeg to enhance cross-modal medical image segmentation through optimized modality fusion and alignment. It integrates a large language model (LLM) with three customized blocks (DITSC, MDCMI, CAC) to fuse visual and textual representations. The LLM generates medical descriptions of infection regions, assisting contrastive learning in DITSC to improve vision-language alignment. MDCMI achieves multi-dimensional spatial semantic interaction for accurate lesion localization, and CAC enhances local feature transmission via coordinate attention. Experiments on the QaTa-COV19 and MosMedData+ datasets demonstrate that LGCSeg precisely captures fine-grained lesion features and exhibits great robustness across diverse medical imaging scenarios.

AI USE STATEMENT

In this work, we used a generative AI tool (GPT-4O-mini) to perform partial grammar checks and spelling corrections. Each time the language editing feature was applied, the author conducted a rigorous review to ensure that the scientific contributions and conclusions remained unaffected. Additionally, we utilized an LLM for the experimental design portion of the study; details are provided in Section 3 of the paper. Throughout this process, we called the LLM's API solely to generate descriptive information for reference purposes. Each time descriptive information was generated, it was manually reviewed to ensure compliance with all relevant requirements. Importantly, the information generated by the LLM served only as a reference and was not used directly; the selection of actual difficult examples was drawn from the corpus we created. The authors bear full responsibility for all scientific contributions and conclusions. We are accountable for the final content of this work, including any text, statements, or artifacts generated with the assistance of generative AI.

ETHICS STATEMENT

This work complies with the ICLR Code of Ethics and does not involve human subjects, human experiments, or personally identifiable data. LLM usage is disclosed in the AI Use Statement; all LLM outputs were manually reviewed by the authors, and all scientific contributions and conclusions are the sole responsibility of the authors.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Reza Azad, Ehsan Khodapanah Aghdam, Amelie Rauland, Yiwei Jia, Atlas Haddadi Avval, Afshin Bozorgpour, Sanaz Karimijafarbigloo, Joseph Paul Cohen, Ehsan Adeli, and Dorit Merhof. Medical image segmentation review: The success of u-net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pp. 1–21. Springer, 2022.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022a.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pp. 205–218. Springer, 2022b.

Miguel A Carreira-Perpinan. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(5):767–776, 2007.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yaxiong Chen, Minghong Wei, Zixuan Zheng, Jingliang Hu, Yilei Shi, Shengwu Xiong, Xiao Xiang Zhu, and Lichao Mou. Causalclipseg: Unlocking clip's potential in referring medical image segmentation with causal intervention. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 77–87. Springer, 2024.

Zihan Cheng, Jintao Guo, Jian Zhang, Lei Qi, Luping Zhou, Yinghuan Shi, and Yang Gao. Mamba-sea: A mamba-based framework with global-to-local sequence augmentation for generalizable medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025a.

Zihan Cheng, Jintao Guo, Jian Zhang, Lei Qi, Luping Zhou, Yinghuan Shi, and Yang Gao. Mamba-sea: A mamba-based framework with global-to-local sequence augmentation for generalizable medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025b.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Aysen Degerli, Serkan Kiranyaz, Muhammad EH Chowdhury, and Moncef Gabbouj. Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In *2022 IEEE International Conference on Image Processing (ICIP)*, pp. 2306–2310. IEEE, 2022.

Jie Gui, Wen Sha, and Xiuquan Du. Towards text-mask consistency in medical image segmentation. In *International Conference on Learning Representations*, volume 2026, pp. 55230–55247, 2026.

Qibin Hou, Daquan Zhou, and Jiashi Feng. Coordinate attention for efficient mobile network design. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13713–13722, 2021.

Yutao Hu, Qixiong Wang, Wenqi Shao, Enze Xie, Zhenguo Li, Jungong Han, and Ping Luo. Beyond one-to-one: Rethinking the referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4067–4077, 2023.

Xiaoshuang Huang, Hongxiang Li, Meng Cao, Long Chen, Chenyu You, and Dong An. Cross-modal conditioned reconstruction for language-guided medical image segmentation. *IEEE Transactions on Medical Imaging*, 2024.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.

Haoxuan Li, Yi Bin, Junrong Liao, Yang Yang, and Heng Tao Shen. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 924–934, 2023a.

Zhe Li, Laurence T Yang, Bocheng Ren, Xin Nie, Zhangyang Gao, Cheng Tan, and Stan Z Li. Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11704–11714, 2024.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 43(1):96–107, 2023b.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Sergey P Morozov, Anna E Andreychenko, Nikolay A Pavlov, AV Vladzymyrskyy, Natalya V Ledikhova, Victor A Gombolevskiy, Ivan A Blokhin, Pavel B Gelezhe, AV Gonchar, and V Yu Chernina. Mosmeddata: Chest ct scans with covid-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020.

Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

Qingtao Pan, Zhengrong Li, Wenhao Qiao, Jingjiao Lou, Qing Yang, Guang Yang, and Bing Ji. Amvlm: Alignment-multiplicity aware vision-language model for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

Fernando Pérez-Cruz and Olivier Bousquet. Kernel methods and their potential use in signal processing. *IEEE signal processing magazine*, 21(3):57–65, 2004.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Md Motiur Rahman, Saeka Rahman, Smriti Bhatt, and Miad Faezipour. Text-assisted vision model for medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Jiacheng Ruan, Jincheng Li, and Suncheng Xiang. Vm-unet: Vision mamba unet for medical image segmentation. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.

Dandan Shan, Zihan Li, Yunxiang Li, Qingde Li, Jie Tian, and Qingqi Hong. Stpnet: Scale-aware text prompt network for medical image segmentation. *IEEE Transactions on Image Processing*, 2025.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19412–19424, 2024.

Nikhil Kumar Tomar, Debesh Jha, Ulas Bagci, and Sharib Ali. Tganet: Text-guided attention for improved polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 151–160. Springer, 2022.

Haonan Wang, Peng Cao, Jiaqi Wang, and Osmar R Zaiane. Uctransnet: rethinking the skip connections in u-net from a channel-wise perspective with transformer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 2441–2449, 2022.

Qijie Wang, Xian Lin, and Zengqiang Yan. Towards robust medical image referring segmentation with incomplete textual prompts. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 636–646, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-04981-0.

Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Toward robust referring image segmentation. *IEEE Transactions on Image Processing*, 33:1782–1794, 2024.

Yingchao Xiao, Huangang Wang, and Wenli Xu. Parameter selection of gaussian kernel for one-class svm. *IEEE transactions on cybernetics*, 45(5):941–953, 2014.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18155–18165, 2022.

Zijian Yi, Ziming Zhao, Zhishu Shen, and Tiehua Zhang. Multimodal fusion via hypergraph autoencoder and contrastive learning for emotion recognition in conversation. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 4341–4348, 2024.

Liangwei Zhang, Jing Lin, and Ramin Karim. Adaptive kernel density-based anomaly detection for nonlinear systems. *Knowledge-Based Systems*, 139:50–63, 2018.

Xu Zhang, Bo Ni, Yang Yang, and Lefei Zhang. Madapter: A better interaction between image and language for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 425–434. Springer, 2024.

Yao Zhang, Jiawei Yang, Jiang Tian, Zhongchao Shi, Cheng Zhong, Yang Zhang, and Zhiqiang He. Modality-aware mutual learning for multi-modal medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 589–599. Springer, 2021.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pp. 2–25. PMLR, 2022.

Tianwei Zhou, Songbai Tan, Baoquan Zhao, and Guanghui Yue. Multitask deep neural network with knowledge-guided attention for blind image quality assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):7577–7588, 2024.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *International workshop on deep learning in medical image analysis*, pp. 3–11. Springer, 2018.

Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE transactions on medical imaging*, 39(6):1856–1867, 2019.

A LLM-BASED TEXT GENERATION

A frozen LLM (GPT-4o mini) was employed to generate semantically challenging medical descriptions, leveraging its multimodal reasoning and generation capabilities. As shown in Fig. 9, the LLM-based medical text was generated by calling an application programming interface (API) to produce textual outputs from paired text-image inputs. A manual inspection step filtered the generated content: invalid texts were reprocessed by the LLM, while validated samples constituted the final training dataset. During manual inspection, we verified that the generated description adheres to established medical knowledge of pneumonia, specifically regarding the infection type, lesion count, and laterality (distribution of affected areas in the left versus right lung). The prompt used was as follows: "The image description is T, which is the confirmed infected area. From a medical perspective, please predict the areas in the image that may be infected in the next stage but have not been described. Please answer in the following manner: no additional information is required. If 'and' is used, both the left and right lungs must be described." The construction of these sophisticated medical textual descriptions is finished in advance of model training.

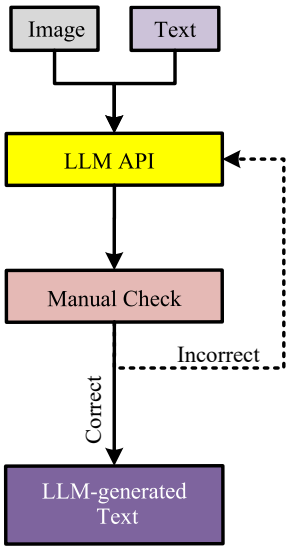


Figure 9: Generating Medical Descriptions via an LLM API.

B MULTIDIMENSIONAL FLATTENING STRATEGY

After applying the multi-head mechanism to F_{img}, the result $F_{img}^{'}$ is obtained. Subsequently, $F_{img}^{'}$ is processed through global flattening, horizontal pooling followed by flattening, and vertical pooling followed by flattening to extract global information F_{img}^{G}, horizontal information F_{img}^{H}, and vertical information F_{img}^{W}, respectively, as follow:

$$F_{img}^{G} = Flatten(F_{img}^{'}) \tag{14}$$

$$F_{B,C,H,1}^{W} = Flatten\left(\frac{1}{W}\sum_{w=1}^{W} X_{B,C,H,W}\right) \tag{15}$$

$$F_{B,C,1,W}^{H} = Flatten\left(\frac{1}{H}\sum_{h=1}^{H} X_{B,C,H,W}\right) \tag{16}$$

where X denotes $F_{img}^{'}$. During multimodal feature interaction, we employ the multi-head attention module to integrate image and text features.

C DATASETS AND IMPLEMENTATION DETAILS

Detailed dataset descriptions. Two major medical datasets were used for experimentation and validation: the QaTa-COV19 (X-ray) (Li et al., 2023b; Degerli et al., 2022) and MosMedData+ (CT) datasets (Li et al., 2023b; Morozov et al., 2020). The QaTa-COV19 dataset comprises 9258 images of chest X-rays, each annotated with manually marked regions of pathological findings. We adopted an extended version of LViT (Li et al., 2023b) for training visual language models. The training set partitioning follows LViT's methodology, yielding 5716 image-text pairs to train, 1429 pairs to validate and 2113 pairs to test. The MosMedData+ dataset comprises 2729 CT images of pulmonary infections. The lesion region descriptions follow the same methodology as the QaTa-COV19 dataset. Data partitioning also employs the Lvit (Li et al., 2023b), with 2183 pairs allocated to train, 273 pairs to validate and 273 pairs to test. The evaluation metrics employed in the experiment are $Dice$ and $mIOU$, which are as follows:

$$Dice = \sum_{i=1}^{N} \sum_{j=1}^{C} \frac{1}{NC} \cdot \frac{2\left|p_{ij} \cap y_{ij}\right|}{\left(\left|p_{ij}\right| + \left|y_{ij}\right|\right)} \tag{17}$$

$$mIoU = \sum_{i=1}^{N} \sum_{j=1}^{C} \frac{1}{NC} \cdot \frac{\left|p_{ij} \cap y_{ij}\right|}{\left|p_{ij} \cup y_{ij}\right|} \tag{18}$$

where N is the number of pixels, C denotes the segmentation categories, p_{ij} represents the prediction result and y_{ij} denotes the ground-truth label.

Detailed implementation. The experimental environment was an Intel Core i9-10980 CPU @ 3.00 GHz and NVIDIA GeForce RTX 3090 GPUs, and our code was based on PyTorch. The Adam optimizer was employed for gradient optimization, with an initial learning rate of 0.0003. The number of training epochs was set to 200. Additionally, hyperparameter tuning during the experiment will be introduced in the ablation experiment section.

Descriptions of compared segmentation methods. To fully demonstrate the superiority of LGC-Seg, this study conducts a comprehensive comparison between methods that operate without medical descriptions and language-guided methods (which add medical description inputs). The compared models encompass representative and widely adopted image segmentation frameworks, including UNet (MICCAI, 2015) (Ronneberger et al., 2015), UNet++ (MICCAI, 2018) (Zhou et al., 2019), nnUNet (Nature Methods, 2021) (Isensee et al., 2021), TransUNet (arXiv, 2021) (Chen et al., 2021), SwinUNet (arXiv, 2021) (Cao et al., 2022b), CLIP (ICML, 2021) (Radford et al., 2021), UCTransNet (AAAI, 2022) (Wang et al., 2022), ConVIRT (PMLR, 2022) (Zhang et al., 2022), TGANet (MICCAI, 2022) (Tomar et al., 2022), LAVT (CVPR, 2022) (Yang et al., 2022), LViT (TMI, 2023) (Li et al., 2023b), RefSegFormer (TIP, 2024) (Wu et al., 2024), ReCLMIS (TMI, 2024) (Huang et al., 2024) STPNet (TIP, 2025) (Shan et al., 2025) and ARSegNet (MICCAI, 2026) (Wang et al., 2026).

D VISUALIZATION RESULTS OF LOSS FUNCTION WEIGHTS SELECTION EXPERIMENTS

Fig. 10 shows visual segmentation results under different loss function weight configurations. Both L_{Dice} and L_{CE} contribute to improving overall segmentation accuracy. As shown in Fig. 10(a)(b)(g), an excessively large λ_1 results in missed or incorrectly segmented lesion regions. In the LGCSeg, setting λ_1 to 4 or 5 yields visually satisfactory outcomes. Furthermore, Fig. 10(c)(d)(g) demonstrate that $\lambda_2 = 0.1$ produces the most stable and reliable visual performance. Finally, Fig. 10(e)-(g) indicates that when λ_3 is set to either a high value (0.5) or a low value (0.05), certain lesion regions remain partially segmented (highlighted in red). The optimal segmentation quality is achieved when $\lambda_3 = 0.1$.

E VISUALIZATION RESULTS OF ABLATION EXPERIMENTS

To further demonstrate the superiority of our proposed method, we provide more representative and comprehensive visualization results of the ablation experiments, as shown in Figure 2. The selected

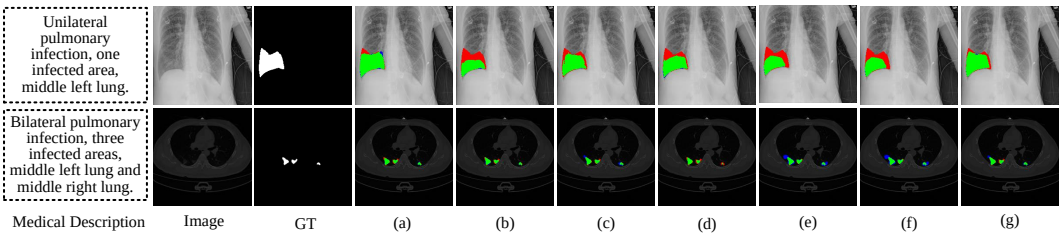


Figure 10: Segmentation results of loss function hyperparameter selection. (a)$\lambda_1 = 4, \lambda_2 = 0.1, \lambda_3 = 0.1$; (b)$\lambda_1 = 6, \lambda_2 = 0.1, \lambda_3 = 0.1$; (c) $\lambda_1 = 5, \lambda_2 = 0.05, \lambda_3 = 0.1$; (d)$\lambda_1 = 5, \lambda_2 = 0.5, \lambda_3 = 0.1$; (e)$\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.05$; (f)$\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.5$; (g)$\lambda_1 = 5, \lambda_2 = 0.1, \lambda_3 = 0.1$.

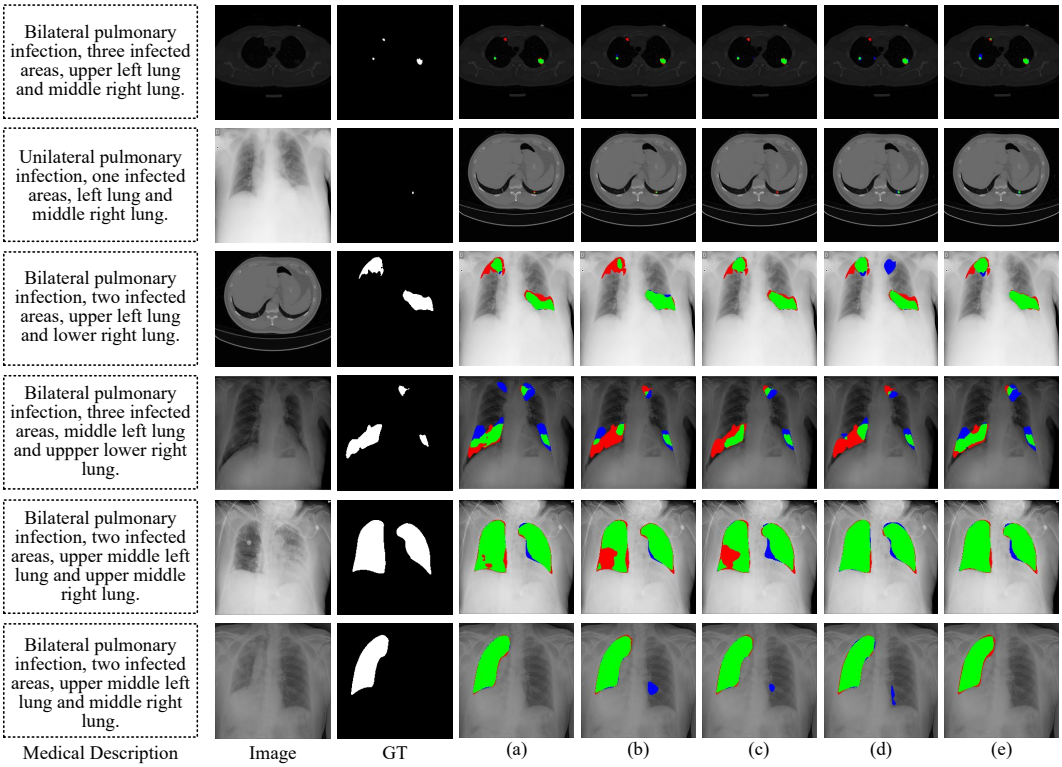


Figure 11: Visualization of segmentation results for ablation of LGCSeg's primary blocks. (a) w/o Batch Contrastive; (b) w/o Corpus Contrastive; (c) w/o MDCMI block; (d) w/o CAC block; (e) LGCSeg. In the segmentation results, red areas indicate under-segmentation of the lesion, while blue areas indicate over-segmentation of the lesion.

cases include images with one, two and three lesions, as well as cases involving single-lung infection and bilateral or multiple-lung infections.

As shown in Fig. 11, the first, third, and fourth groups of images contain three infected regions. When the batch-level comparison, corpus-level comparison, the MDCMI block, or the CAC block is removed, the models fail to segment the infected region in the upper left lung effectively. This demonstrates that LGCSeg performs well in fine-grained lesion segmentation.

The fifth group shows a case with extensive bilateral lung infection. The comparison indicates that models without corpus-level contrastive learning or the MDCMI block produce noticeable segmentation errors, as highlighted by the red regions in the figure. These results demonstrate the effectiveness of our LLM-guided negative-sample selection strategy and multiscale modal interaction mechanism in improving segmentation accuracy.

The sixth group shows a case with a single infected region in one lung. By comparing the results in Fig. 11 (b), (c),(d) and (e), we observe that models without corpus-level comparison, the MD-CMI block, or the CAC block are prone to over-segmentation and may incorrectly identify normal regions as lesions, as indicated by the blue regions in the figure. In contrast, the complete LGC-Seg framework exhibits stronger robustness, effectively suppresses noise interference, and achieves more accurate segmentation.