Text-Guided Medical Image Segmentation via Residual Aggregation and Dense Alignment Representations

Jie Gui[1,2], Hang Tu[2], Yunfeng Dou[3], Wen Sha[1,2,*], Xiuquan Du[1,3,*], Shuo Li[4]

[1]Key Laboratory of Intelligent Computing and Signal Processing,
Ministry of Education, Anhui University, Hefei, China
[2]School of Artificial Intelligence, Anhui University, Hefei, China
[3]School of Computer Science and Technology, Anhui University, Hefei, China
[4]Department of Biomedical Engineering, Case Western Reserve University, Cleveland, OH, USA
{w125111032,w125121009,e125121027}@stu.ahu.edu.cn
{04069,dxqllp}@ahu.edu.cn

Abstract

Specifying segmentation targets with natural language provides flexible semantic conditioning for medical image segmentation. However, existing methods still struggle to reliably ground high-level textual semantics into accurate pixel-level spatial prompts, especially under low-contrast lesions, ambiguous boundaries, and visually similar anatomical contexts. In this paper, we propose **RADAR** (**R**esidual **A**ggregation and **D**ense **A**lignment **R**epresentations), a parameter-efficient framework for text-guided medical image segmentation. RADAR explicitly decouples and synergizes the medical semantic anchoring capability of BiomedCLIP with the prompt-driven mask decoding ability of MedSAM2, while enabling low-cost task adaptation via LoRA. Specifically, Guided Residual Depth Aggregation adaptively aggregates cross-layer features under semantic guidance to alleviate delayed textual injection and deep semantic dilution. Structured Dense Representation Alignment leverages a frozen BiomedCLIP to construct crop-level target, boundary, and context anchors, improving the local discriminability of dense features and the reliability of automatic prompts. Multi-modal Prompt Generation further converts cross-modal semantic responses into composite text, point, and mask prompts for MedSAM2 decoding. Experiments on three public medical image segmentation datasets show that RADAR achieves competitive performance with only 0.74M trainable parameters, demonstrating its effectiveness in parameter-efficient adaptation, semantic grounding, and dense representation enhancement. The source code is available at `https://github.com/JGuiidev/RADAR`.

1 Introduction

Text-guided medical image segmentation dynamically specifies targets via natural language. Unlike closed-set paradigms relying on predefined labels, this task better aligns with real-world clinical scenarios involving diverse pathologies and low-quality imaging [1, 2]. However, natural language typically provides only high-level semantic priors, such as lesion categories, anatomical locations, and regional scopes, whereas segmentation requires precise pixel-level localization and boundary delineation. This discrepancy often leads to localization deviation, boundary ambiguity, and contextual confusion, forming a substantial semantic-to-spatial gap, as illustrated in Figure 1(a). Thus, reliably translating medical textual semantics into executable spatial prompts remains a key challenge.

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

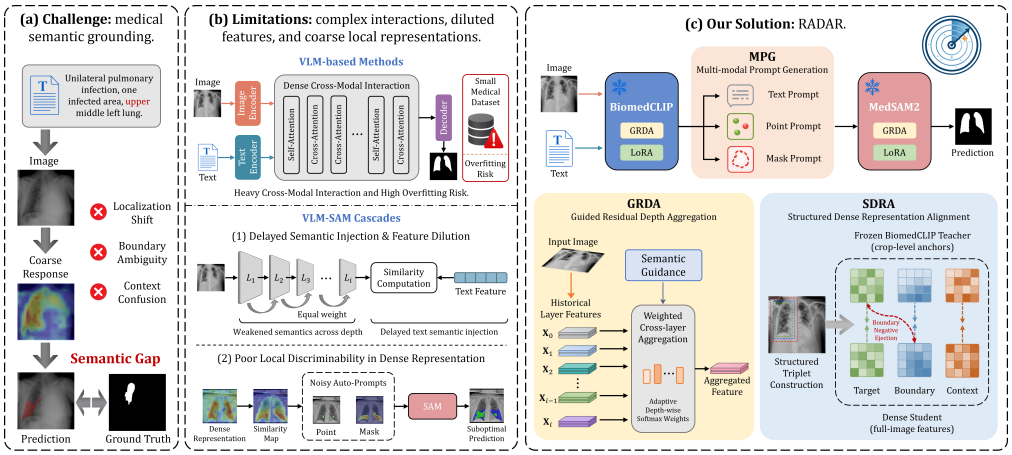


Figure 1: Challenge and limitations of existing methods, and our solution.

Vision-Language Models (VLMs) provide a foundation for this semantic grounding. Early end-to-end multimodal medical segmentation methods rely on elaborate dense cross-modal interactions [3, 4], which often introduce high training costs and may overfit under limited medical annotations. Conversely, prompt-driven models such as SAM and MedSAM exhibit strong class-agnostic mask generation and can recover accurate boundaries from point or box prompts [5, 6]. Yet, they lack native text understanding and depend heavily on manually specified visual prompts.

Recently, foundation models such as CLIP [7] and BiomedCLIP [8] have leveraged contrastive learning to enhance cross-modal alignment, mapping textual and visual representations into a shared semantic space. Several cascaded architectures further combine VLMs with SAM, using VLMs to provide semantic spatial cues and SAM for fine-grained mask decoding [9]. Although these approaches exploit their complementary capabilities, existing VLM-SAM frameworks still face two critical bottlenecks, as summarized in Figure 1(b).

First, textual semantics participate too late, while historical visual evidence can be diluted during deep propagation. Dual-encoder VLMs typically encode visual and textual inputs independently and establish their correspondence only through deep similarity computation or late cross-modal interaction. Consequently, text cannot explicitly determine which intermediate visual representations should be retained for the current target during encoding. Meanwhile, conventional layer-wise feature accumulation often assigns fixed weights to historical representations, gradually weakening discriminative intermediate-layer responses during deep propagation [10].

Second, VLM dense representations lack sufficient local discriminability. VLMs are primarily optimized for image-level or global semantic matching, and their local dense features are not explicitly trained for pixel-level target discrimination [11]. Under low-contrast lesions, blurred boundaries, and visually similar neighboring anatomical structures, text-relevant responses may simultaneously activate target, boundary, and contextual regions. Such ambiguity increases noise in automatically generated prompts and consequently limits the mask decoding quality of downstream SAM models.

To address these limitations, we propose RADAR (Residual Aggregation and Dense Alignment Representations), a parameter-efficient framework for text-guided medical image segmentation. As shown in Figure 1(c), RADAR decouples the medical vision-language semantic anchoring capability of BiomedCLIP from the prompt-conditioned mask decoding capability of MedSAM2, while introducing LoRA into key branches for parameter-efficient adaptation [12]. Guided Residual Depth Aggregation (GRDA) uses textual semantics as a depth-routing condition to adaptively select and aggregate historical visual evidence at each spatial location, rather than repeatedly injecting textual content into the current visual representation. This text-conditioned depth routing mitigates delayed semantic interaction and feature dilution during deep propagation. Structured Dense Representation Alignment (SDRA) further improves the local discriminability of VLM dense representations by constructing target, boundary, and context anchors with a frozen BiomedCLIP teacher. Unlike generic global alignment [11], SDRA emphasizes fine-grained regional differences and thereby improves the reliability of spatial responses. Finally, Multi-modal Prompt Generation (MPG) serves as the

semantic-to-spatial interface between BiomedCLIP and MedSAM2, converting the enhanced cross-modal responses into composite text, point, and mask prompts for prompt-conditioned decoding.

The main contributions of this work are summarized as follows:

- We propose RADAR, a parameter-efficient framework integrating BiomedCLIP for semantic anchoring and MedSAM2 for mask decoding. MPG seamlessly translates cross-modal responses into composite spatial and textual prompts.
- We introduce GRDA, a semantic-guided depth aggregation mechanism that mitigates delayed cross-modal interaction and semantic dilution in deep feature propagation.
- We propose SDRA, which constructs crop-level semantic anchors with a frozen BiomedCLIP to improve dense feature discriminability and prompt reliability.

2 Related work

2.1 Vision-language models in medical segmentation

Traditional medical image segmentation methods, such as FCN [13], U-Net [14], and TransUNet [15], mainly rely on predefined label spaces and thus struggle with open-vocabulary targets and weakly standardized clinical descriptions. Methods such as LAVT [16] and LViT [17] incorporate language guidance through Vision Transformers [18] and cross-modal attention. However, dense cross-modal interaction usually increases training costs and may overfit under limited medical annotations.

Contrastive VLMs provide an alternative foundation for medical semantic grounding. CLIP [7] and ALIGN [19] learn shared image-text spaces, while MedCLIP [20] and BiomedCLIP [8] further improve medical image-text alignment using domain-specific data. Nevertheless, their dense local features are not explicitly trained for pixel-level target discrimination [11, 21]. Directly using VLM similarity responses for segmentation can therefore cause ambiguous activations, boundary leakage, and contextual false responses.

2.2 Promptable foundation models and VLM-SAM cascades

SAM [5] and its medical extensions, including MedSAM [6] and MedSAM2 [22], provide strong prompt-driven mask generation but lack native text understanding. Recent methods therefore combine VLM-based semantic localization with SAM-based mask decoding. In natural images, EVF-SAM [9], GLaMM [23], and DIT [24] explore language grounding or instruction tuning. In the medical domain, MedCLIP-SAMv2 [25], TGSAM-2 [26], SegMoTE [27], and LGA [28] further explore text-guided SAM adaptation. Parameter-efficient techniques such as LoRA [12], QLoRA [29], and MAdapter [30] reduce the cost of adapting these large models.

Existing approaches typically either fuse language repeatedly through dense cross-modal interactions or directly convert the final VLM response into SAM prompts. The former incurs substantial interaction costs, whereas the latter is limited by late semantic interaction and noisy dense responses. In particular, point or mask prompts directly derived from VLM responses [31] may contain background noise, motivating further improvement of semantic propagation and spatial discrimination before prompt generation.

2.3 Feature aggregation and dense representation alignment

Semantic aggregation in deep networks. Residual connections and PreNorm structures are widely used to stabilize deep networks [32, 33], but fixed residual accumulation can progressively dilute the contribution of individual layers. Attention Residuals [10] use depth-wise attention to aggregate historical representations, while BiPVL-Seg [34] progressively injects language information through vision-language fusion. Unlike such cross-modal feature fusion or modality-agnostic depth aggregation, GRDA uses text as a depth-routing condition to select historical visual evidence at each spatial location, enabling text-conditioned and spatially resolved cross-layer retention.

Region-level alignment for dense representations. VLM dense features often lack local discriminability when transferred to segmentation tasks. CLIPSelf [11] and CAT-Seg [35] improve dense

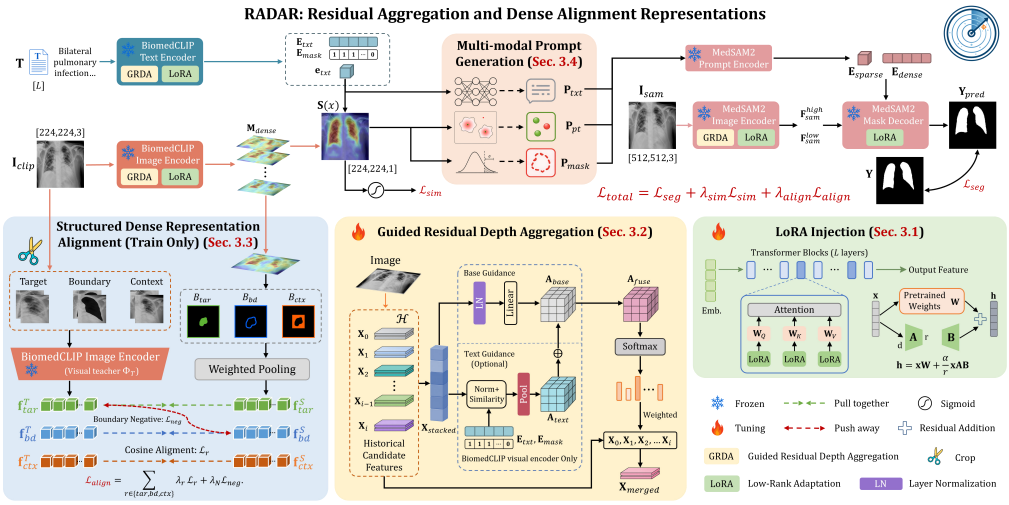


Figure 2: Overview of RADAR. GRDA performs text-conditioned cross-layer routing, SDRA provides training-only dense alignment, and MPG converts cross-modal responses into sparse and dense prompts for MedSAM2. The SDRA teacher is removed at inference.

prediction through self-distillation and cost aggregation, respectively. Unlike generic alignment that mainly encourages teacher-student feature consistency, SDRA explicitly constructs target, boundary, and context regions and models their relative relationships. Its boundary-negative rejection prevents boundary representations from becoming as target-like as the target itself, thereby enforcing relative regional discrimination among target bodies, ambiguous boundaries, and adjacent contexts.

3 Methodology

3.1 Overview

As illustrated in Figure 2, given a medical image $\mathbf{I} \in \mathbb{R}^{H\times W\times C}$ and a text description $\mathbf{T} \in \mathbb{R}^{L}$, we resize the image into $\mathbf{I}_{clip} \in \mathbb{R}^{224\times 224\times 3}$ for BiomedCLIP and $\mathbf{I}_{sam} \in \mathbb{R}^{512\times 512\times 3}$ for MedSAM2. The BiomedCLIP text branch encodes $\mathbf{T}$ into token-level features $\mathbf{E}_{txt} \in \mathbb{R}^{L\times 768}$, a text mask $\mathbf{E}_{mask} \in \{0,1\}^{L}$, and a global text feature $\mathbf{e}_{txt} \in \mathbb{R}^{512}$. Meanwhile, the BiomedCLIP visual branch extracts dense visual features $\mathbf{M}_{dense} \in \mathbb{R}^{768\times 14\times 14}$ from $\mathbf{I}_{clip}$, and the MedSAM2 image encoder extracts multi-scale visual features $\mathbf{F}_{sam}^{low}$ and $\mathbf{F}_{sam}^{high}$ from $\mathbf{I}_{sam}$. Based on $\mathbf{e}_{txt}$, we generate a text prompt $\mathbf{P}_{txt}$. Point and mask prompts, denoted as $\mathbf{P}_{pt}$ and $\mathbf{P}_{mask}$, are automatically mined from the similarity map between $\mathbf{M}_{dense}$ and $\mathbf{e}_{txt}$. The resulting composite prompt set $\{\mathbf{P}_{txt}, \mathbf{P}_{pt}, \mathbf{P}_{mask}\}$ is fed into the MedSAM2 prompt encoder to produce sparse embeddings $\mathbf{E}_{sparse}$ and dense embeddings $\mathbf{E}_{dense}$. Finally, the MedSAM2 decoder combines these prompt embeddings with high- and low-resolution image features to generate the final segmentation mask $\mathbf{Y}_{pred}$.

LoRA injection. To preserve the generalization ability of the pre-trained backbones while enabling parameter-efficient adaptation, we freeze the pretrained backbone weights of BiomedCLIP and MedSAM2 and insert Low-Rank Adaptation (LoRA) modules into selected Transformer layers. For a frozen weight matrix $\mathbf{W} \in \mathbb{R}^{d_{in}\times d_{out}}$, the forward computation is reformulated as

$$\mathbf{h} = \mathbf{x}\mathbf{W} + \frac{\alpha}{r}\mathbf{x}\mathbf{A}\mathbf{B}, \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{d_{in}\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times d_{out}}$ are trainable low-rank matrices, $r \ll \min(d_{in}, d_{out})$ is the rank, and α is the scaling factor. In practice, LoRA is selectively injected into different Transformer layer ranges to control the adaptation capacity and parameter budget.

3.2 Guided Residual Depth Aggregation

To enable text-conditioned selection of historical visual evidence and alleviate feature dilution during deep propagation, we propose Guided Residual Depth Aggregation (GRDA). For the current layer feature $\mathbf{X}_i$, we collect its historical features as $\mathcal{H}_i = \{\mathbf{X}_j\}_{j \le i}$ and stack them along the depth dimension: $\mathbf{X}_{stacked} \in \mathbb{R}^{K \times B \times N \times C}$, where K is the number of candidate layers. Given an aggregation score $\mathbf{A} \in \mathbb{R}^{K \times B \times N}$, the normalized depth-wise weights and merged feature are computed as

$$\boldsymbol{\alpha} = \mathrm{Softmax}\left(\frac{\mathbf{A}}{\tau}, \mathrm{dim} = 0\right), \quad \mathbf{X}_{merged} = \sum_{j \le i} \boldsymbol{\alpha}_j \odot \mathbf{X}_j \tag{2}$$

where τ is a temperature parameter. The aggregated feature is injected through a learnable residual scale:

$$\mathbf{X}_{out} = \mathbf{X}_i + \gamma \left(\mathbf{X}_{merged} - \mathbf{X}_i\right). \tag{3}$$

We initialize γ to 0, so that GRDA starts as an identity mapping and preserves the pretrained representations at the beginning of training.

Base aggregation score. For the BiomedCLIP text encoder and the MedSAM2 image encoder, GRDA computes feature-intrinsic aggregation scores without cross-modal intervention:

$$\mathbf{A}_{base} = \mathrm{Squeeze}\left(\mathrm{Linear}_{C \to 1}\left(\mathrm{LayerNorm}\left(\mathbf{X}_{stacked}\right)\right)\right) \in \mathbb{R}^{K \times B \times N}. \tag{4}$$

Text-conditioned routing score. For the BiomedCLIP visual encoder, we further introduce token-level textual guidance. Specifically, we compute the cosine similarity between the stacked visual features and the text token features $\mathbf{E}_{txt}$, and perform mask-weighted pooling along the text length dimension using the text mask $\mathbf{E}_{mask}$ to exclude padding tokens:

$$\mathbf{S}^{(l)}_{text} = \cos(\mathbf{X}_{stacked}, \mathbf{E}^{(l)}_{txt}), \quad \mathbf{A}_{text} = \frac{\sum_{l=1}^{L} \mathbf{S}^{(l)}_{text} \cdot \mathbf{E}^{(l)}_{mask}}{\sum_{l=1}^{L} \mathbf{E}^{(l)}_{mask}}. \tag{5}$$

Finally, the fused aggregation score is obtained by additive fusion: $\mathbf{A}_{fuse} = \mathbf{A}_{base} + \mathbf{A}_{text}$.

3.3 Structured Dense Representation Alignment

The BiomedCLIP visual encoder is originally optimized for global image-text retrieval, where deep features may lose local boundaries and spatial details required for medical segmentation. Moreover, its dense grid features lack sufficient local discriminability to separate target regions, blurred boundaries, and similar backgrounds. To address this, we modify the last ViT block and introduce Structured Dense Representation Alignment (SDRA) to enhance the local discriminability of dense representations.

Dense feature extraction modification. Following Wu et al. [11] and Zhou et al. [21], we modify the last ViT block of the BiomedCLIP visual encoder by replacing the original attention output with the projected Value-branch representation:

$$\mathbf{y} = \mathbf{x} + \mathrm{Proj}\left(\mathrm{Emb}_v(\mathbf{x})\right), \quad \mathbf{z} = \mathbf{y} + \mathrm{FFN}(\mathbf{y}), \tag{6}$$

where $\mathbf{x}$ denotes the input visual tokens, $\mathrm{Emb}_v(\cdot)$ denotes the Value-branch mapping, and $\mathrm{Proj}(\cdot)$ is the output projection. The output $\mathbf{z}$ is reshaped into a spatial grid to obtain the dense visual feature map $\mathbf{M}_{dense}$.

Structured triplet construction. Given the ground-truth mask $\mathbf{Y}$ during training, we first compute its minimum bounding box B_{orig}. Then, three structured regions are constructed by box expansion: $B_{tar} = \mathrm{Expand}(B_{orig}, v_{tar})$, $B_{bd} = \mathrm{Expand}(B_{tar}, v_{bd})$, and $B_{ctx} = \mathrm{Expand}(B_{tar}, v_{ctx})$. Here, B_{tar} denotes the target-centered region, B_{bd} covers blurred boundaries and transition regions, and B_{ctx} captures surrounding context or similar background structures. These nested structured regions provide region-level supervision that better matches common segmentation errors in medical images.

Crop-level teacher and student representations. For each region $r \in \{tar, bd, ctx\}$, a frozen BiomedCLIP image encoder Φ_T extracts the crop-level teacher anchor: $\mathbf{f}_r^T = \Phi_T(\mathrm{Crop}(\mathbf{I}_{clip}, B_r))$. Meanwhile, the student representation is obtained by mask-weighted pooling over $\mathbf{M}_{dense}$ using the corresponding spatial mask $\mathbf{M}_r$ associated with bounding box B_r: $\mathbf{f}_r^S = \mathrm{Pool}_{mask}(\mathbf{M}_{dense}, \mathbf{M}_r)$. The regional alignment loss is defined as

$$\mathcal{L}_r = 1 - \left\langle \frac{\mathbf{f}_r^S}{\|\mathbf{f}_r^S\|_2 + \epsilon}, \frac{\mathbf{f}_r^T}{\|\mathbf{f}_r^T\|_2 + \epsilon} \right\rangle, \quad r \in \{tar, bd, ctx\}. \tag{7}$$

Boundary negative rejection. Since boundary regions may contain both foreground and background information, they can be incorrectly absorbed into the target representation. We therefore introduce a boundary negative rejection term:

$$\mathcal{L}_{neg} = \mathrm{ReLU}\left(\cos(\mathbf{f}_{bd}^S, \mathbf{f}_{tar}^T) - \cos(\mathbf{f}_{tar}^S, \mathbf{f}_{tar}^T)\right). \tag{8}$$

This term encourages the boundary representation to be less similar to the target teacher anchor than the target student representation.

The final structured alignment loss is

$$\mathcal{L}_{align} = \sum_{r \in \{tar, bd, ctx\}} \lambda_r \mathcal{L}_r + \lambda_N \mathcal{L}_{neg}, \tag{9}$$

where λ_r and λ_N denote the weights of the regional alignment and boundary negative rejection terms, respectively.

3.4 Multi-modal Prompt Generation

To convert BiomedCLIP cross-modal semantics into spatial control signals for MedSAM2, we design Multi-modal Prompt Generation (MPG), which produces a composite prompt set $\mathcal{P} = \{\mathbf{P}_{txt}, \mathbf{P}_{pt}, \mathbf{P}_{mask}\}$.

Text prompt generation. The global text embedding $\mathbf{e}_{txt}$ from the BiomedCLIP text branch encodes the target category and medical semantics. We project it into the MedSAM2 prompt space with a two-layer MLP: $\mathbf{P}_{txt} = \mathbf{W}_2\ \mathrm{GELU}(\mathbf{W}_1 \mathbf{e}_{txt})$, where $\mathbf{W}_1$ and $\mathbf{W}_2$ are learnable projection matrices.

Mask and point prompt generation. To align the dimensionalities, the dense visual features $\mathbf{M}_{dense}$ are first mapped to the shared semantic space using the frozen BiomedCLIP visual projection layer. We then compute the cross-modal cosine similarity map $\mathbf{S}(\mathbf{x})$ between the projected dense features and the global text feature $\mathbf{e}_{txt}$. For mask prompting, the similarity response is normalized and resized to the mask-prompt resolution required by MedSAM2, yielding $\mathbf{P}_{mask}$. For point prompting, $\mathbf{S}(\mathbf{x})$ is resized to the MedSAM2 input resolution and processed by adaptive quantile thresholding to obtain high-confidence foreground regions. Connected components are then extracted from these regions. Positive points are sampled from eroded high-confidence foreground cores, while negative points are sampled from surrounding hard-background regions. The resulting point prompt set is $\mathbf{P}_{pt} = \{(\mathbf{p}_i^+, 1)\} \cup \{(\mathbf{p}_j^-, 0)\}$, where labels 1 and 0 denote foreground and background, respectively. Detailed point prompt generation settings are provided in Appendix A.

Training objective. The final MedSAM2 prediction $\mathbf{Y}_{pred}$ is supervised by a BCE-Dice segmentation loss: $\mathcal{L}_{seg} = 0.5\mathcal{L}_{BCE}(\mathbf{Y}_{pred}, \mathbf{Y}) + 0.5\mathcal{L}_{Dice}(\mathbf{Y}_{pred}, \mathbf{Y})$. We convert the similarity map $\mathbf{S}(\mathbf{x})$ into a probability map using a sigmoid function and supervise it with the same BCE-Dice formulation as an auxiliary loss $\mathcal{L}_{sim}$. Detailed loss definitions are provided in Appendix B. Together with the structured dense representation alignment loss $\mathcal{L}_{align}$, the total objective is

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{sim}\mathcal{L}_{sim} + \lambda_{align}\mathcal{L}_{align}, \tag{10}$$

where λ_{sim} and λ_{align} control the weights of similarity-map supervision and structured alignment, respectively.

4 Experiments

4.1 Experimental settings

Datasets and metrics. We evaluate RADAR on three public medical image segmentation datasets, including two chest infection segmentation datasets, QaTa-COV19 [36] and MosMedData+ [37, 38], and one colonoscopy polyp segmentation dataset, Kvasir [39]. For QaTa-COV19 and MosMedData+, we follow the data splits and text annotations provided by LViT [17]. For Kvasir, we use the text annotations and experimental setting from MAdapter [30], and split the original training set into training and validation subsets with a 9:1 ratio, while keeping the standard test set fixed. All compared methods are retrained and evaluated under the same data splits. We report Dice, mean Intersection over Union (mIoU), 95th percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). Detailed dataset statistics and metric definitions are provided in Appendix C.

Implementation details. All experiments are conducted on Ubuntu 25.04 with an Intel Core i9-14900KF CPU and a single NVIDIA GeForce RTX 5090 D GPU using PyTorch 2.9.0 and CUDA 12.8. The BiomedCLIP and MedSAM2 backbones are frozen during training. For LoRA, we set the rank to $r = 8$ and the scaling factor to $\alpha = 1.0$. The numerical stability term is set to $\epsilon = 10^{-6}$. For structured region alignment, we set $\lambda_{tar} = 0.5$ and $\lambda_{bd} = \lambda_{ctx} = \lambda_N = 0.2$. For the total objective, we set $\lambda_{sim} = 0.2$ and $\lambda_{align} = 0.2$. We train the model with AdamW using an initial learning rate of 3×10^{-4}, a batch size of 8, CosineAnnealingWarmRestarts scheduling, and early stopping.

4.2 Comparison with existing methods

Quantitative comparison. As shown in Table 1, RADAR achieves the best Dice and mIoU on QaTa-COV19, MosMedData+, and Kvasir. Compared with the strongest pure visual baselines, it improves Dice by 6.65, 10.12, and 5.96 percentage points, respectively. Against the best V-L methods, RADAR further improves Dice/mIoU by 0.84/1.01 and 1.53/2.04 percentage points on MosMedData+ and Kvasir. Compared with the best external V-L+S baselines, it improves Dice by 2.65, 3.01, and 2.35 percentage points while maintaining more stable boundary quality. RADAR requires only 0.74M trainable parameters and 46.66G FLOPs, demonstrating strong task-adaptation efficiency with moderate forward computation. Cross-backbone results further support the synergy between medical-domain semantic pretraining and a medical promptable decoder. Robustness to incomplete, noisy, synonymous, and incorrect text is further evaluated in Appendix D.1.

Table 1: Quantitative comparison across three datasets. Dice and mIoU are reported in percentages (%), while HD95/ASSD are reported in pixels. Params denotes the number of trainable parameters. "Visual", "V-L", and "V-L+S" denote pure vision models, vision-language models, and VLM-SAM cascades, respectively. BioCLIP denotes BiomedCLIP. CLIP-SAM2, BioCLIP-SAM2, and CLIP-MedSAM2 correspond to RADAR backbone variants equipped with the full GRDA+SDRA configuration. Best and second-best results are **bolded** and underlined, respectively.

Method	Type	Trainable Params ↓	FLOPs ↓	QaTa-COV19				MosMedData+				Kvasir			
				Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
U-Net [14]	Visual	31.04M	41.91G	76.95	66.97	30.75	8.68	40.57	30.41	42.55	14.81	54.82	41.38	54.33	20.06
UNet++ [40]	Visual	9.16M	26.72G	76.60	66.29	35.00	9.54	65.35	50.78	23.28	5.81	61.74	49.13	55.25	19.20
Swin-UNet [41]	Visual	27.17M	**5.92G**	77.47	68.05	26.28	7.55	63.98	50.61	22.45	5.89	83.90	75.98	19.51	6.37
TransUNet [15]	Visual	105.28M	49.36G	77.38	67.27	30.21	8.01	63.52	49.26	23.83	5.81	86.19	78.50	23.28	6.64
LGMS [42]	V-L	43.95M	11.18G	83.33	74.33	15.78	4.20	71.83	58.22	13.74	3.29	90.62	84.45	16.74	4.55
LViT [17]	V-L	39.93M	27.08G	83.21	74.43	17.43	4.73	73.97	60.89	18.55	4.40	88.14	81.07	21.56	5.59
MMI-UNet [43]	V-L	56.18M	11.04G	83.86	75.20	16.10	4.35	70.80	57.23	14.89	4.14	90.42	84.61	**14.16**	**3.82**
LAVT [16]	V-L	189.73M	49.69G	83.81	75.27	**14.07**	**3.88**	67.00	53.28	16.81	4.24	89.54	83.46	14.68	4.66
RecLMIS [4]	V-L	69.44M	26.69G	83.87	75.16	15.36	4.19	74.63	61.42	18.88	4.31	84.12	75.78	23.76	6.42
MTGT [44]	V-L	56.73M	128.56G	77.22	67.71	27.24	8.12	68.01	54.22	16.78	4.09	71.76	62.40	36.98	11.86
HiMix [45]	V-L	44.73M	11.27G	82.56	73.25	18.89	5.04	72.57	58.96	14.64	3.91	89.94	84.09	17.85	4.60
DIT [24]	V-L+S	103.38M	92.82G	79.99	70.40	28.79	7.81	72.22	59.00	15.19	3.58	89.80	83.97	16.70	5.01
LGA [28]	V-L+S	97.91M	381.18G	81.47	72.22	21.90	6.27	72.46	58.92	20.09	4.76	86.37	78.60	23.15	6.77
CLIP-SAM2	V-L+S	1.13M	252.66G	81.83	72.33	16.97	5.22	74.58	61.28	14.71	3.74	88.25	82.26	16.75	4.91
BioCLIP-SAM2	V-L+S	**0.74M**	254.55G	81.62	71.94	18.81	5.47	73.15	59.35	16.58	4.48	90.02	84.60	17.82	4.95
CLIP-MedSAM2	V-L+S	1.13M	93.85G	83.35	74.36	15.97	4.77	73.90	60.43	14.83	3.83	90.27	84.80	16.37	4.51
RADAR (Ours)	V-L+S	**0.74M**	46.66G	**84.12**	**75.32**	14.27	4.09	**75.47**	**62.43**	**13.16**	**3.26**	**92.15**	**86.65**	14.79	3.91

Qualitative analysis. The qualitative results further support these observations. As shown in Figure 3, pure visual methods generally suffer from unstable localization and boundary deviations, making it difficult to distinguish low-contrast lesions from visually similar anatomical structures, as illustrated in the fourth column. The small-target CT examples in the fourth and fifth rows further show that several methods are prone to missed detections, localization shifts, or background false activations when segmenting tiny lesions. Although vision-language methods introduce textual semantics, they fail to convert such semantics into explicit and reliable spatial prompts, leading to incorrect activations in non-target lung regions, as shown in the first and second rows. Existing V-L+S cascaded methods improve contour quality with SAM-style decoders, but remain highly dependent on the prompts generated by the front-end VLM. Once noisy similarity responses are produced, the errors are propagated to the downstream mask decoding process, as observed in the ninth and tenth columns. In contrast, RADAR exhibits more stable semantic grounding and boundary recovery across different modalities and target scales, validating its advantages in small-target perception, background suppression, and boundary consistency.

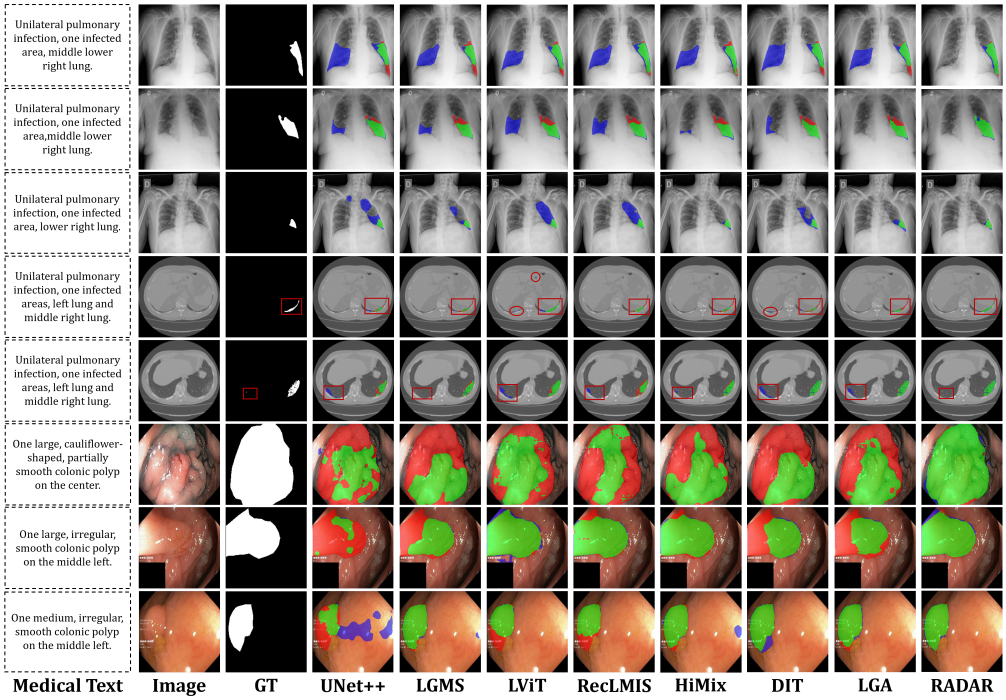


Figure 3: Qualitative comparison on representative QaTa-COV19, MosMedData+, and Kvasir samples covering low-contrast lesions, small targets, and ambiguous boundaries. Green, red, and blue indicate true positive, false negative, and false positive pixels, respectively. RADAR shows improved target localization, background suppression, and boundary recovery across these challenging cases.

4.3 Ablation studies

Effect of main component. Table 2 validates the effectiveness of the main components in RADAR. Starting from the MedSAM2-only baseline, progressively adding text, point, and mask prompts consistently improves performance, showing that MPG effectively converts textual semantics into spatial constraints. Adding GRDA on top of Full MPG further raises Dice to 72.78% and 91.35% on MosMedData+ and Kvasir, respectively, indicating the benefit of text-conditioned cross-layer routing. Removing SDRA, Dense Mod., or GRDA consistently degrades performance, confirming their complementary roles in local discrimination, spatial detail preservation, and semantic propagation. The full RADAR achieves the best results on both datasets. Additional comparisons with direct and learnable prompt-generation alternatives are provided in Appendix D.2, with further ablations in Appendix E.

Table 2: Main component ablation study on MosMedData+ and Kvasir. Dense Mod. denotes the modification of the last BiomedCLIP ViT block for dense feature extraction.

Variant	MPG			GRDA	Dense Mod.	SDRA	MosMedData+		Kvasir	
	Text	Point	Mask				Dice (%) ↑	mIoU (%) ↑	Dice (%) ↑	mIoU (%) ↑
MedSAM2-only	–	–	–	–	–	–	69.19	55.33	88.11	81.07
+ Text	✓	–	–	–	–	–	69.86	55.68	88.73	82.17
+ Text + Point	✓	✓	–	–	–	–	71.14	57.47	89.93	83.62
Full MPG	✓	✓	✓	–	–	–	71.47	57.67	90.36	84.29
Full MPG + GRDA	✓	✓	✓	✓	–	–	72.78	59.40	91.35	85.61
w/o SDRA	✓	✓	✓	✓	✓	–	73.71	60.22	91.63	86.25
w/o Dense Mod.	✓	✓	✓	✓	–	✓	75.29	62.21	91.77	86.24
w/o GRDA	✓	✓	✓	–	✓	✓	74.63	61.30	90.84	85.28
RADAR (Ours)	✓	✓	✓	✓	✓	✓	**75.47**	**62.43**	**92.15**	**86.65**

Effect of GRDA and SDRA. Table 3 validates the transferability and complementarity of GRDA and SDRA across different backbone combinations. On CLIP-SAM2, BioCLIP-SAM2, and CLIP-MedSAM2, introducing either GRDA or SDRA consistently improves performance over the baseline without both modules, indicating that semantic-guided cross-layer aggregation and structured dense alignment are not tied to a specific backbone. Furthermore, jointly incorporating GRDA and SDRA achieves the best Dice and mIoU across all backbone combinations, demonstrating that the two modules complement each other in stabilizing semantic propagation and enhancing local dense discriminability, thereby improving medical semantic grounding.

Table 3: Cross-backbone ablation study of GRDA and SDRA on MosMedData+ and Kvasir. BioCLIP denotes BiomedCLIP.

Backbone	Variant	GRDA	SDRA	MosMedData+				Kvasir			
				Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
CLIP-SAM2	w/o GRDA & SDRA	×	×	71.86	57.90	19.06	5.06	86.11	79.08	21.73	6.25
	GRDA only	✓	×	73.84	60.35	17.41	4.36	87.57	81.03	19.62	5.58
	SDRA only	×	✓	73.01	59.33	17.91	4.72	87.79	81.21	19.63	5.47
	GRDA + SDRA	✓	✓	**74.58**	**61.28**	**14.71**	**3.74**	**88.25**	**82.26**	**16.75**	**4.91**
BioCLIP-SAM2	w/o GRDA & SDRA	×	×	71.70	57.85	18.71	5.08	88.91	82.56	21.46	6.01
	GRDA only	✓	×	72.10	58.25	17.04	4.68	89.56	83.69	19.48	5.46
	SDRA only	×	✓	72.47	58.91	17.28	4.93	89.45	83.60	19.07	5.31
	GRDA + SDRA	✓	✓	**73.15**	**59.35**	**16.58**	**4.48**	**90.02**	**84.60**	**17.82**	**4.95**
CLIP-MedSAM2	w/o GRDA & SDRA	×	×	72.56	58.99	16.03	4.42	88.82	82.15	21.13	6.02
	GRDA only	✓	×	73.31	59.75	16.95	4.23	89.84	83.95	**15.35**	**4.41**
	SDRA only	×	✓	73.23	59.80	15.79	4.05	89.75	83.58	19.83	5.47
	GRDA + SDRA	✓	✓	**73.90**	**60.43**	**14.83**	**3.83**	**90.27**	**84.80**	16.37	4.51

Ablation study on GRDA variants. Table 4 examines two internal design choices of GRDA: aggregation score design and branch placement. For score design, the Base and Text scores are individually effective, while their additive fusion achieves the best Dice/mIoU on both datasets. Text-only yields slightly better boundary metrics on MosMedData+, indicating complementary structural and semantic cues and making Base+Text the most balanced choice overall. For branch placement, restricting GRDA to S-I or removing either B-V or B-T degrades performance, whereas jointly applying it to S-I, B-V, and B-T performs best overall. These results support feature-intrinsic cross-layer aggregation across branches together with text-conditioned routing in the BiomedCLIP visual branch.

Visualization analysis of SDRA. Figure 4 illustrates the effect of SDRA on the dense representations of BiomedCLIP on QaTa-COV19 and MosMedData+. The second and third columns show the last-layer dense outputs of the BiomedCLIP visual encoder without and with SDRA, respectively. Without SDRA, the dense responses are scattered over non-target lung regions, bone structures, and visually similar anatomical backgrounds, indicating that the original BiomedCLIP dense features mainly capture coarse text-related semantics but lack fine-grained local discriminability. After introducing SDRA, the responses become more concentrated around the true lesion regions, with irrelevant background activations clearly suppressed and better spatial consistency with the ground truth. These

Table 4: Ablation of GRDA score design and branch placement on MosMedData+ and Kvasir. S-I, B-V, and B-T denote the MedSAM2 image, BiomedCLIP visual, and text encoders, respectively. Base and Text denote feature-intrinsic and text-conditioned scores in B-V.

Variant	Score	GRDA branches	MosMedData+				Kvasir			
			Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
Aggregation score design										
Base-only	Base	S-I, B-V, B-T	75.13	62.01	13.37	3.28	91.75	86.15	17.60	4.53
Multiplicative	Base × Text	S-I, B-V, B-T	75.35	62.32	13.22	3.37	91.90	86.21	16.69	4.24
Text-only	Text	S-I, B-V, B-T	75.22	62.11	**12.83**	**3.20**	92.00	86.44	15.48	3.93
GRDA branch placement										
S-I only	–	S-I	74.82	61.72	13.27	3.41	91.00	84.98	19.49	4.72
w/o B-V	–	S-I, B-T	75.30	62.12	13.32	3.36	91.33	85.45	18.76	4.81
w/o B-T	Base + Text	S-I, B-V	75.38	62.28	13.92	3.40	91.84	86.14	17.23	4.37
RADAR (Ours)	Base + Text	S-I, B-V, B-T	**75.47**	**62.43**	13.16	3.26	**92.15**	**86.65**	**14.79**	**3.91**

results demonstrate that SDRA effectively enhances the local discriminability of dense representations, providing a more reliable spatial basis for automatic prompt generation and fine-grained segmentation. More extended visual evidence and discussion of SDRA are provided in Appendix F.

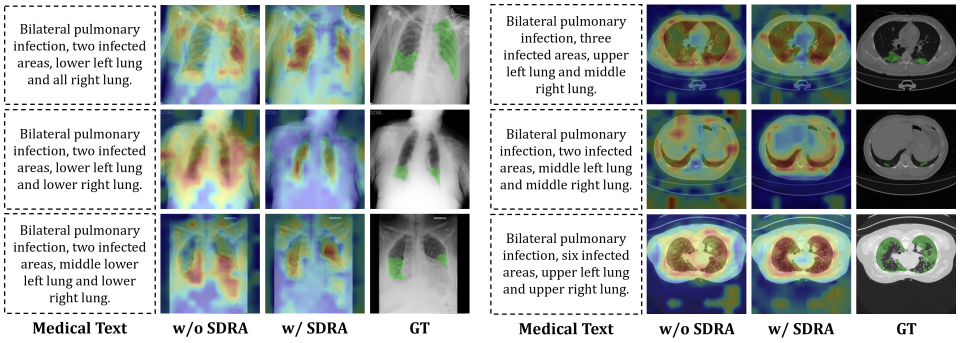


Figure 4: Qualitative visualization of SDRA. The columns show the medical text, BiomedCLIP dense features without SDRA, BiomedCLIP dense features with SDRA, and ground-truth masks.

5 Conclusion

We propose RADAR, a parameter-efficient framework for text-guided medical image segmentation. RADAR decouples the semantic anchoring capability of BiomedCLIP from the mask decoding ability of MedSAM2 and enables low-cost adaptation via LoRA. GRDA preserves target-relevant visual evidence through text-conditioned cross-layer routing, SDRA improves local dense-feature discriminability through structured regional alignment, and MPG converts enhanced cross-modal responses into executable text, point, and mask prompts. Together, these components form a continuous semantic-to-spatial pathway from textual guidance to pixel-level segmentation. Experiments across multiple medical segmentation benchmarks demonstrate that RADAR achieves strong and consistent performance with only 0.74M trainable parameters, highlighting its effectiveness for parameter-efficient collaboration between medical VLMs and promptable segmentation models.

6 Limitations

Despite its strong performance, RADAR still has two main limitations. First, it relies on the pretrained priors of BiomedCLIP and MedSAM2, and current validation mainly focuses on 2D tasks without native 3D, multi-organ, or simultaneous multiclass evaluation. Second, MPG still relies on heuristic thresholding, connected-component analysis, and morphological operations, which may be unstable under low contrast, complex targets, or inaccurate and incomplete text. Future work will explore learnable prompt generation and more robust cross-modal alignment.

Acknowledgments

This work was supported in part by the Jiangsu Provincial Association of Traditional Chinese Medicine Project (Grant No. CYTF2024089) and the Science and Technology Innovation Project of Anhui University Chuzhou Research Institute (Grant No. KJCX-202408). The authors gratefully acknowledge the High-performance Computing Platform of Anhui University for providing computing resources.

References

[1] Taha Koleilat, Hojat Asgariandehkordi, Omid Nejatimanzari, Berardino Barile, Yiming Xiao, and Hassan Rivaz. Medclipseg: Probabilistic vision-language adaptation for data-efficient and generalizable medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1406–1417, June 2026.

[2] Aryan Das, Tanishq Rachamalla, Koushik Biswas, Swalpa Kumar Roy, and Vinay Kumar Verma. Uncertainty-aware vision-language segmentation for medical imaging. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8490–8499, 2026.

[3] Qijie Wang, Xian Lin, and Zengqiang Yan. Towards robust medical image referring segmentation with incomplete textual prompts. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15966 of *Lecture Notes in Computer Science*, pages 636 – 646. Springer Nature Switzerland, September 2025.

[4] 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*, 44(4):1821–1835, 2025. doi: 10.1109/TMI.2024.3523333.

[5] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

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

[7] 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*, pages 8748–8763. PmLR, 2021.

[8] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2025. doi: 10.1056/AIoa2400640.

[9] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *Image and Vision Computing*, 173:106028, 2026. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2026.106028. URL `https://www.sciencedirect.com/science/article/pii/S0262885626001356`.

[10] Kimi Team, Guangyu Chen, Yu Zhang, Jianlin Su, Weixin Xu, Siyuan Pan, Yaoyu Wang, Yucheng Wang, Guanduo Chen, Bohong Yin, Yutian Chen, Junjie Yan, Ming Wei, Y. Zhang, Fanqing Meng, Chao Hong, Xiaotong Xie, Shaowei Liu, Enzhe Lu, Yunpeng Tai, Yanru Chen, Xin Men, Haiqing Guo, Y. Charles, Haoyu Lu, Lin Sui, Jinguo Zhu, Zaida Zhou, Weiran He, Weixiao Huang, Xinran Xu, Yuzhi Wang, Guokun Lai, Yulun Du, Yuxin Wu, Zhilin Yang, and Xinyu Zhou. Attention residuals, 2026.

[11] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DjzvJCRsVf.

[12] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[13] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015. doi: 10.1109/CVPR.2015.7298965.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[15] Jieneng Chen, Jieru Mei, Xianhang Li, Yongyi Lu, Qihang Yu, Qingyue Wei, Xiangde Luo, Yutong Xie, Ehsan Adeli, Yan Wang, Matthew P. Lungren, Shaoting Zhang, Lei Xing, Le Lu, Alan Yuille, and Yuyin Zhou. Transunet: Rethinking the u-net architecture design for medical image segmentation through the lens of transformers. *Medical Image Analysis*, 97:103280, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103280. URL https://www.sciencedirect.com/science/article/pii/S1361841524002056.

[16] Zhao Yang, Jiaqi Wang, Xubing Ye, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. Language-aware vision transformer for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5238–5255, 2025. doi: 10.1109/TPAMI.2024.3468640.

[17] 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, 2024. doi: 10.1109/TMI.2023.3291719.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[19] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[20] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

[21] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 696–712, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19815-1.

[22] Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos, 2025. URL https://arxiv.org/abs/2504.03600.

[23] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024.

[24] Xiaorui Huang, Gen Luo, Chaoyang Zhu, Bo Tong, Yiyi Zhou, Xiaoshuai Sun, and Rongrong Ji. Deep instruction tuning for segment anything model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 905–914, 2024.

[25] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. Medclip-samv2: Towards universal text-driven medical image segmentation. *Medical Image Analysis*, 106: 103749, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103749. URL https://www.sciencedirect.com/science/article/pii/S1361841525002968.

[26] Runtian Yuan, Ling Zhou, Jilan Xu, Qingqiu Li, Mohan Chen, Yuejie Zhang, Rui Feng, Tao Zhang, and Shang Gao. Tgsam-2: Text-guided medical image segmentation using segment anything model 2. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15969 of *Lecture Notes in Computer Science*, pages 565 – 574. Springer Nature Switzerland, September 2025.

[27] Yujie Lu, Jingwen Li, Sibo Ju, Yanzhou Su, He Yao, Yisong Liu, Min Zhu, and Junlong Cheng. Segmote: Token-level mixture of experts for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 36332–36342, June 2026.

[28] Jihong Hu, Yinhao Li, Hao Sun, Yu Song, Chujie Zhang, Lanfen Lin, and Yen-Wei Chen. Lga: A language guide adapter for advancing the sam model's capabilities in medical image segmentation. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15012 of *Lecture Notes in Computer Science*, pages 610 – 620. Springer Nature Switzerland, October 2024.

[29] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=OUIFPHEgJU.

[30] Xu Zhang, Bo Ni, Yang Yang, and Lefei Zhang. Madapter: A better interaction between image and language for medical image segmentation. In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15009 of *Lecture Notes in Computer Science*, pages 425 – 434. Springer Nature Switzerland, October 2024.

[31] Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26081–26090, 2025. doi: 10.1109/CVPR52734.2025.02429.

[32] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[34] Rafi Ibn Sultan, Hui Zhu, Chengyin Li, and Dongxiao Zhu. Bipvl-seg: Bidirectional progressive vision-language fusion with global-local alignment for medical image segmentation, 2025. URL https://arxiv.org/abs/2503.23534.

[35] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.

[36] Aysen Degerli, Serkan Kiranyaz, Muhammad E. H. 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)*, pages 2306–2310, 2022. doi: 10.1109/ICIP46576.2022.9897412.

[37] 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.

[38] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs. Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem. *European radiology experimental*, 4(1):50, 2020.

[39] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In Yong Man Ro, Wen-Huang Cheng, Junmo Kim, Wei-Ta Chu, Peng Cui, Jung-Woo Choi, Min-Chun Hu, and Wesley De Neve, editors, *MultiMedia Modeling*, pages 451–462, Cham, 2020. Springer International Publishing. ISBN 978-3-030-37734-2.

[40] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In Danail Stoyanov, Zeike Taylor, Gustavo Carneiro, Tanveer Syeda-Mahmood, Anne Martel, Lena Maier-Hein, João Manuel R.S. Tavares, Andrew Bradley, João Paulo Papa, Vasileios Belagiannis, Jacinto C. Nascimento, Zhi Lu, Sailesh Conjeti, Mehdi Moradi, Hayit Greenspan, and Anant Madabhushi, editors, *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00889-5.

[41] 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 Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 205–218, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25066-8.

[42] Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 724–733, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43901-8.

[43] Phuoc-Nguyen Bui, Duc-Tai Le, and Hyunseung Choo. Visual-textual matching attention for lesion segmentation in chest images. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 702–711, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72114-4.

[44] Longxuan Zhao, Tao Wang, Xinlin Zhang, Yuanbin Chen, Ence Yang, and Tong Tong. Mtgt: Multiscale text feature-guided transformer in medical image segmentation. *Image and Vision Computing*, 165:105846, 2026. ISSN 0262-8856. doi: https://doi.org/10.1016/j.imavis.2025.105846. URL `https://www.sciencedirect.com/science/article/pii/S0262885625004342`.

[45] Soojin Hwang, Jaeyoon Sim, and Won Hwa Kim. Himix : Hierarchical visual-textual mixing network for lesion segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5332–5341, March 2026.

A Details of point prompt generation

The main paper describes the overall point-prompt generation process in MPG. Here, we provide the implementation details. Given the cross-modal similarity map $\mathbf{S}(\mathbf{x})$ produced by BiomedCLIP, where $\mathbf{x}$ denotes a spatial location, we first construct high-confidence candidate regions and then sample positive and negative prompt points from high-confidence foreground cores and neighboring hard-background regions, respectively.

MPG first computes a sample-adaptive threshold according to the similarity distribution. We use a quantile threshold with $q = 0.90$ and constrain it to be non-negative:

$$\mathbf{R}_{fg}(\mathbf{x}) = \mathbb{I}\left[\mathbf{S}(\mathbf{x}) \geq \max\left(Q_{0.90}(\mathbf{S}), 0\right)\right], \tag{11}$$

where $Q_{0.90}(\mathbf{S})$ denotes the 0.90 quantile of the similarity map and $\mathbb{I}[\cdot]$ is the indicator function.

We then perform connected-component analysis and remove isolated regions smaller than 16 pixels. For each valid component, its area and peak similarity response are recorded, and at most two main candidate components are retained. This suppresses isolated false activations and provides stable spatial support for subsequent prompt sampling.

Before morphological processing, the candidate regions are represented at the 512×512 input resolution required by MedSAM2. Therefore, the morphological kernel sizes and point-distance constraints described below are defined in this spatial resolution.

Positive prompt points are sampled from high-confidence foreground cores. For each retained component $\mathbf{R}_{cc}$, morphological erosion is applied:

$$\mathbf{R}_{core} = \mathrm{Erode}\left(\mathbf{R}_{cc}, k_e\right), \tag{12}$$

where the erosion kernel size is $k_e = 5$ with two iterations. If the eroded region contains fewer than 4 pixels, the erosion strength is gradually reduced. If no valid foreground core remains, the original connected component is used as a fallback. Positive points are preferentially selected from high-response locations in $\mathbf{R}_{core}$. We first sample at most one point from each retained component and then supplement from the union of foreground cores until $N_+ = 2$ positive points are obtained.

Negative points are sampled from nearby hard-background regions. For each retained component, we construct a local background ring using dilation kernels of 17 and 27 pixels:

$$\mathbf{R}_{ring} = \mathrm{Dilate}\left(\mathbf{R}_{cc}, 27\right) \setminus \left(\mathrm{Dilate}\left(\mathbf{R}_{cc}, 17\right) \cup \mathbf{R}_{sel}\right), \tag{13}$$

where $\mathbf{R}_{sel}$ denotes the union of all retained foreground candidate components. If $\mathbf{R}_{ring}$ is empty, the inner dilated region excluding the retained foreground components is used as a fallback.

During sampling, at most 128 candidate points are retained from each candidate region. Positive points are selected in descending order of similarity response, while negative points are preferentially sampled from high-response locations in the background ring to form hard negatives. A minimum distance of 12 pixels is enforced to avoid excessive clustering. We use $N_+ = 2$ positive and $N_- = 2$ negative points:

$$\mathbf{P}_{pt} = \left\{\left(\mathbf{p}_i^+, 1\right)\right\}_{i=1}^{2} \cup \left\{\left(\mathbf{p}_j^-, 0\right)\right\}_{j=1}^{2}, \tag{14}$$

where $\mathbf{p}_i^+$ and $\mathbf{p}_j^-$ denote positive and negative points, and labels 1 and 0 indicate foreground and background, respectively.

B Loss function details

This section provides the detailed definitions of the Binary Cross-Entropy (BCE) and Dice losses used in RADAR. Let $\mathbf{Y} \in \{0,1\}^{H \times W}$ denote the ground-truth mask and $\mathbf{P} \in [0,1]^{H \times W}$ a predicted probability map. The spatial domain is denoted by Ω.

Binary Cross-Entropy loss. The BCE loss is defined as

$$\mathcal{L}_{BCE}(\mathbf{P}, \mathbf{Y}) = -\frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} \left[\mathbf{Y}(\mathbf{x}) \log\left(\mathbf{P}(\mathbf{x}) + \epsilon\right) + \left(1 - \mathbf{Y}(\mathbf{x})\right) \log\left(1 - \mathbf{P}(\mathbf{x}) + \epsilon\right)\right], \tag{15}$$

where ϵ is a small constant for numerical stability.

Dice loss. The Dice loss is defined as

$$\mathcal{L}_{Dice}(\mathbf{P}, \mathbf{Y}) = 1 - \frac{2\sum_{\mathbf{x}\in\Omega} \mathbf{P}(\mathbf{x})\mathbf{Y}(\mathbf{x}) + \epsilon}{\sum_{\mathbf{x}\in\Omega} \mathbf{P}(\mathbf{x}) + \sum_{\mathbf{x}\in\Omega} \mathbf{Y}(\mathbf{x}) + \epsilon}. \tag{16}$$

The final MedSAM2 segmentation loss is

$$\mathcal{L}_{seg} = 0.5\mathcal{L}_{BCE}\left(\mathbf{Y}_{pred}, \mathbf{Y}\right) + 0.5\mathcal{L}_{Dice}\left(\mathbf{Y}_{pred}, \mathbf{Y}\right). \tag{17}$$

For similarity-map supervision, the cross-modal cosine similarity is converted into a probability map using a sigmoid function:

$$\mathbf{P}_{sim}(\mathbf{x}) = \sigma\left(\mathbf{S}(\mathbf{x})\right), \tag{18}$$

where $\sigma(\cdot)$ denotes the sigmoid function. The ground-truth mask is resized to the corresponding spatial resolution, denoted as $\mathbf{Y}_{sim}$, and the auxiliary similarity loss follows the same BCE-Dice formulation:

$$\mathcal{L}_{sim} = 0.5\mathcal{L}_{BCE}\left(\mathbf{P}_{sim}, \mathbf{Y}_{sim}\right) + 0.5\mathcal{L}_{Dice}\left(\mathbf{P}_{sim}, \mathbf{Y}_{sim}\right). \tag{19}$$

Together with the structured dense representation alignment loss $\mathcal{L}_{align}$, the total objective is

$$\mathcal{L}_{total} = \mathcal{L}_{seg} + \lambda_{sim}\mathcal{L}_{sim} + \lambda_{align}\mathcal{L}_{align}. \tag{20}$$

We set $\lambda_{sim} = 0.2$ and $\lambda_{align} = 0.2$ in all experiments.

C Dataset and evaluation metrics

C.1 Dataset details

Our main evaluation uses three public medical image segmentation datasets: QaTa-COV19, MosMedData+, and Kvasir.

QaTa-COV19 [36] is a chest X-ray infection segmentation dataset for COVID-19-related lesions. We follow the data split and text annotations provided by [17]. The training, validation, and testing sets contain 5716, 1429, and 2113 images, respectively.

MosMedData+ [37, 38] is a chest CT infection segmentation dataset with pixel-level infection annotations. We follow the split and text annotation protocol of [17]. The training, validation, and testing sets contain 2183, 273, and 273 CT slices, respectively.

Kvasir [39] is a colonoscopy polyp segmentation dataset containing 1000 images. We follow the text annotations and experimental setting of [30]. The standard testing set is kept fixed, while the remaining images are divided into training and validation subsets with a 9:1 ratio using random seed 42, resulting in 810, 90, and 100 images for training, validation, and testing, respectively.

All compared methods are retrained and evaluated using the same data splits, preprocessing pipeline, and evaluation protocol. Text-guided methods additionally use the same image-text annotations. The dataset split statistics are summarized in Table 5.

Table 5: Statistics of the dataset splits used for training, validation, and testing.

Dataset	Modality	Task	Training	Validation	Testing	Total
QaTa-COV19 [36]	X-ray	Chest infection	5716	1429	2113	9258
MosMedData+ [37, 38]	CT	Chest infection	2183	273	273	2729
Kvasir [39]	Colonoscopy	Polyp segmentation	810	90	100	1000

C.2 Evaluation metrics

We use Dice, mean Intersection over Union (mIoU), 95th percentile Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASSD). Predicted probability maps are binarized at a threshold of 0.5 before evaluation. All metrics are computed per test sample and then averaged over the test set. Dice and mIoU are reported as percentages, while HD95 and ASSD are measured in pixels.

Let P_i and G_i denote the predicted and ground-truth foreground pixel sets of the i-th test sample, respectively. Dice is defined as

$$\mathrm{Dice} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|P_i \cap G_i| + \epsilon}{|P_i| + |G_i| + \epsilon}, \tag{21}$$

where N denotes the number of test samples.

In our binary segmentation setting, mIoU denotes the sample-averaged foreground IoU:

$$\mathrm{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i| + \epsilon}{|P_i \cup G_i| + \epsilon}. \tag{22}$$

For boundary metrics, let ∂P_i and ∂G_i denote the predicted and ground-truth boundary point sets. The shortest Euclidean distance from a point a to a boundary set B is

$$d(a, B) = \min_{b \in B} \|a - b\|_2. \tag{23}$$

For samples with non-empty predicted and ground-truth boundaries, the bidirectional distance set is

$$D_i = \{d(a, \partial G_i) \mid a \in \partial P_i\} \cup \{d(b, \partial P_i) \mid b \in \partial G_i\}. \tag{24}$$

Since distance metrics are undefined when either boundary is empty, such samples are excluded from HD95 and ASSD:

$$\mathcal{V} = \{i \mid \partial P_i \neq \emptyset,\ \partial G_i \neq \emptyset\}. \tag{25}$$

The same exclusion rule is applied consistently to all compared methods.

HD95 is defined as

$$\mathrm{HD95} = \frac{1}{|\mathcal{V}|}\sum_{i \in \mathcal{V}} \mathrm{Percentile}_{95}(D_i), \tag{26}$$

and ASSD as

$$\mathrm{ASSD} = \frac{1}{|\mathcal{V}|}\sum_{i \in \mathcal{V}} \frac{\sum_{a \in \partial P_i} d(a, \partial G_i) + \sum_{b \in \partial G_i} d(b, \partial P_i)}{|\partial P_i| + |\partial G_i|}. \tag{27}$$

Higher Dice and mIoU indicate better region overlap, while lower HD95 and ASSD indicate smaller boundary errors.

D Additional analysis

D.1 Robustness to text perturbations

We evaluate the dependence of RADAR on textual guidance by fixing the trained checkpoint and all test images while changing only the test-time text. Incomplete prompts remove disambiguating location, number, or extent attributes and therefore represent under-specified descriptions.

As shown in Table 6, mild noise and synonym substitutions cause relatively small performance drops, whereas incomplete descriptions produce larger degradation. Empty text reduces Dice by 7.22 and 7.67 percentage points on QaTa-COV19 and MosMedData+, respectively, with substantially increased boundary errors. Incorrect text further degrades regional and boundary quality. These results show that RADAR is reasonably robust to semantic-preserving text variations but still depends on correct image-text correspondence.

D.2 Comparison with learnable prompt generation

To examine whether the handcrafted spatial prompt conversion in MPG can be replaced by alternative strategies, we keep BiomedCLIP, MedSAM2, GRDA, SDRA, losses, and all other training settings fixed. RADAR-SimMask directly uses the normalized similarity response as the mask prompt, while RADAR-Extrema samples extrema from the response map as spatial prompts. RADAR-LMPG

Table 6: Robustness to different test-time text conditions on QaTa-COV19 and MosMedData+. The model checkpoint and input images are fixed, and only the textual descriptions are changed. Best and second-best results are **bolded** and underlined, respectively.

Text type	QaTa-COV19				MosMedData+			
	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
Original	**84.12**	**75.32**	**14.27**	**4.09**	**75.47**	**62.43**	**13.16**	**3.26**
Empty	76.90	66.92	33.90	10.13	67.80	54.40	20.49	5.23
Incomplete	82.01	71.95	20.32	5.67	73.72	60.53	13.77	3.60
Mild Noise	83.57	74.72	17.54	4.85	75.02	61.97	13.40	3.28
Synonym	83.26	74.32	19.95	5.39	74.56	61.41	13.79	3.33
Wrong	70.33	58.38	46.25	13.98	70.63	57.00	26.34	6.86

instead concatenates the final BiomedCLIP spatial features with the similarity map and uses a three-layer convolutional head to predict a mask prompt and positive/negative heatmaps. Spatial softmax and differentiable soft-argmax are then used to obtain point coordinates.

As shown in Table 7, direct similarity-map and extrema prompts are more sensitive to false activations and uncertain boundaries. RADAR-LMPG is a viable learnable alternative and slightly improves Kvasir Dice and mIoU, but performs worse on all MosMedData+ metrics and on Kvasir boundary metrics while increasing trainable parameters from 0.74M to 0.91M (+22.97%). The original MPG therefore provides a more balanced performance–parameter trade-off in the current setting.

Table 7: Comparison of different prompt-generation strategies on MosMedData+ and Kvasir. Best and second-best results are **bolded** and underlined, respectively.

Method	Trainable Params	MosMedData+				Kvasir			
		Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
RADAR-SimMask	0.74M	74.81	61.68	14.05	3.33	91.00	84.98	19.49	4.72
RADAR-Extrema	0.74M	75.19	61.98	13.71	3.57	91.55	85.75	16.84	4.31
RADAR-LMPG	0.91M	75.10	61.97	13.38	3.29	**92.24**	**86.74**	16.73	4.03
RADAR	0.74M	**75.47**	**62.43**	**13.16**	**3.26**	92.15	86.65	**14.79**	**3.91**

E Additional ablation studies

E.1 Ablation study on LoRA injection layers

Table 8 compares different LoRA injection ranges. Full-layer adaptation slightly improves Kvasir overlap but increases trainable parameters to 1.23M, while overly narrow injection ranges reduce cross-dataset stability. The adopted configuration applies LoRA to all MedSAM2 image-encoder layers and the last four BiomedCLIP visual/text layers, providing a favorable accuracy–parameter trade-off with 0.74M trainable parameters.

Table 8: Ablation study on LoRA injection layers. S-I, B-V, and B-T denote the MedSAM2 image encoder, BiomedCLIP visual encoder, and BiomedCLIP text encoder, respectively. The MedSAM2 decoder LoRA layers are fixed to S-D 0–1 in all settings. Best and second-best results are **bolded** and underlined, respectively.

Setting	LoRA injection layers	Trainable Params ↓	MosMedData+				Kvasir			
			Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓	Dice ↑	mIoU ↑	HD95 ↓	ASSD ↓
All layers	S-I 0–11, B-V/B-T 0–11	1.23M	74.67	61.44	14.29	3.49	**92.29**	**86.92**	14.92	**3.76**
BioCLIP late-6	S-I 0–11, B-V/B-T 6–11	0.86M	73.52	60.19	14.35	3.88	92.03	86.63	15.17	3.90
BioCLIP last-2	S-I 0–11, B-V/B-T 10–11	**0.62M**	74.78	61.51	13.77	3.53	91.52	85.62	17.45	4.56
MedSAM2 late-6	S-I 6–11, B-V/B-T 0–11	0.69M	74.22	60.94	**12.97**	3.28	91.98	86.48	16.74	4.32
MedSAM2 late-4	S-I 8–11, B-V/B-T 0–11	0.67M	72.17	58.43	13.20	3.63	91.89	86.86	16.88	4.45
RADAR (Ours)	S-I 0–11, B-V/B-T 8–11	0.74M	**75.47**	**62.43**	13.16	**3.26**	92.15	86.65	**14.79**	3.91

F Extended visualization analysis of SDRA

Figure 5 further illustrates the effect of SDRA on BiomedCLIP dense representations and subsequent prompt generation using representative samples from QaTa-COV19 and MosMedData+. The visualization includes dense features before and after SDRA, text-image similarity maps, generated point prompts, final predictions, and ground-truth masks.

Without SDRA, high-response regions are often distributed across non-target lung regions, bone structures, and visually similar backgrounds. After introducing SDRA, responses become more concentrated around lesion regions with fewer irrelevant activations. The resulting similarity maps and point prompts also show better spatial consistency with the target regions. These observations suggest that SDRA improves the spatial concentration of cross-modal responses and reduces noisy prompt propagation, providing more informative spatial cues for MedSAM2 decoding.

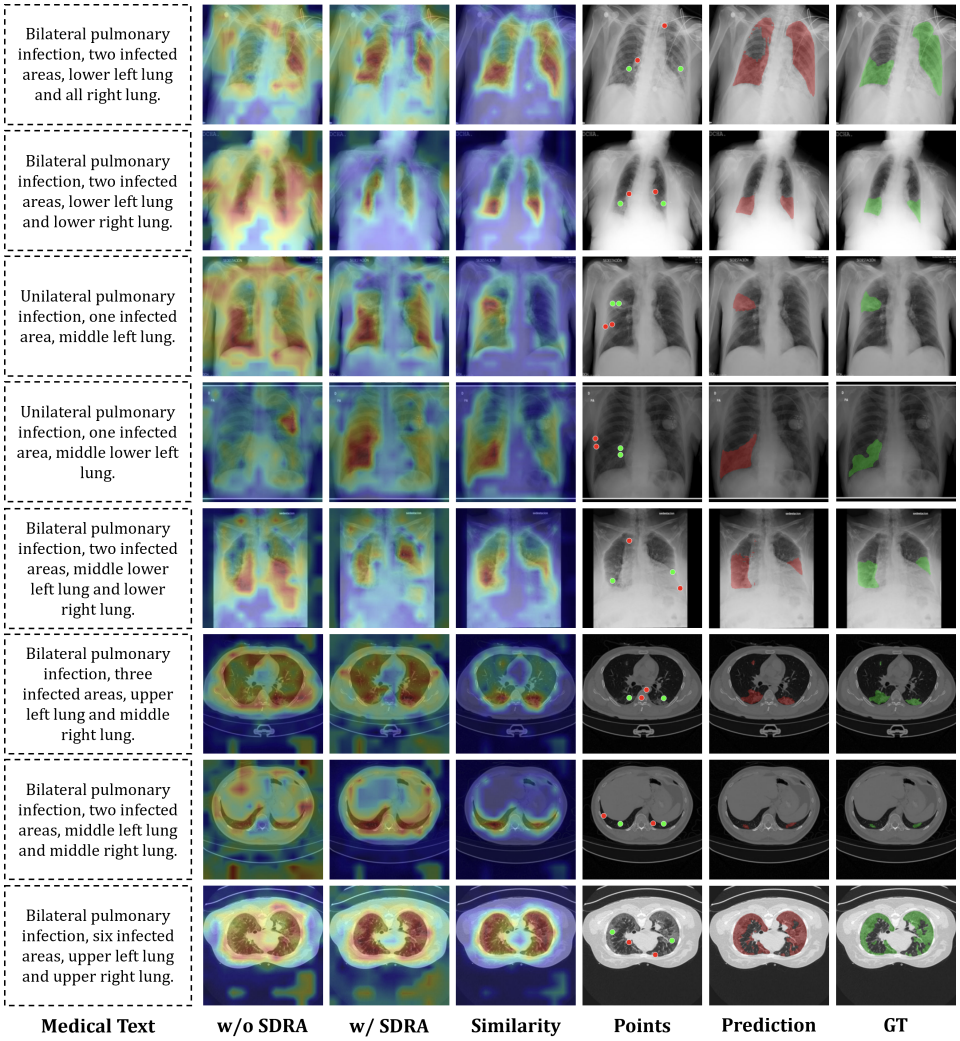


Figure 5: Extended visualization of SDRA on QaTa-COV19 and MosMedData+. The columns show the medical text, BiomedCLIP dense features without and with SDRA, text-image similarity maps, generated point prompts, predictions, and ground-truth masks.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction are aligned with the proposed RADAR framework, including parameter-efficient adaptation, GRDA, SDRA, and MPG. The empirical claims are supported by experiments, ablation studies, and qualitative results on three public datasets.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated Limitations section discussing the dependence on pretrained priors and the heuristic nature of the prompt mining strategy.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present theoretical results or formal theorems. The mathematical expressions are used to define model components and training objectives.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides dataset splits, model architecture details, training hyperparameters, evaluation metrics, loss definitions, and ablation settings. Additional implementation details are provided in the appendix and supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments are conducted on public datasets, and the anonymized implementation code with reproduction instructions is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, data splits, evaluation metrics, optimizer, learning rate, batch size, scheduler, LoRA settings, loss weights, and implementation environment in the Experiments section and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests because retraining all baselines across multiple random seeds is computationally expensive. To reduce variability and ensure fair comparison, all methods are retrained and evaluated under the same data splits, text annotations, preprocessing pipeline, and evaluation protocol. We acknowledge this as a limitation and leave multi-seed evaluation for future work.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports the main hardware and software environment, including CPU, GPU type, GPU memory, CUDA version, and PyTorch version. All experiments are conducted on a single NVIDIA GeForce RTX 5090 D GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses public datasets and does not involve new data collection from human subjects. The proposed method is intended for medical image segmentation research and is not presented as a standalone clinical diagnostic system.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The work may support medical image analysis by improving text-guided segmentation. However, incorrect predictions could be harmful if used without expert oversight, so the method should not be used as a standalone clinical decision system.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release high-risk generative models, scraped datasets, or assets with obvious dual-use misuse risks. The released code is intended for reproducing medical image segmentation experiments on public datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper cites the original sources of the datasets, pretrained models, and baseline methods used in the experiments. We use these assets for research purposes and follow their publicly available terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper provides anonymized code in the supplementary material, including implementation files and instructions for reproducing the main experiments. No new dataset is introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or new research with human subjects. All experiments are conducted on public medical image segmentation datasets.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve new human subject data collection or intervention. The experiments use public datasets released by their original providers.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method does not use LLMs as an important, original, or non-standard component. Any language models used only for writing, editing, or formatting do not affect the methodology, experiments, or scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.