Simple-ViLMedSAM: Simple Text Prompts Meet Vision-Language Models for Medical Image Segmentation

Supplementary Material

1. Appendix A

1.1. Data Information

In our work, we employ four 2D medical imaging datasets covering a variety of imaging modalities, anatomical structures, and disease types to evaluate the performance of the segmentation model.

- *Kvasir-SEG* is a public medical image dataset designed for colorectal polyp segmentation. Released by the Norwegian Cancer Registry, it contains 1,000 endoscopic images, each annotated with pixel-wise segmentation masks. The image resolutions range from 332×487 to 1920×1072 pixels, covering polyps of various shapes, sizes, and colors. The dataset is split into 700 training cases, 100 evaluation cases, and 200 test cases.
- *ISIC-2018* is a skin lesion segmentation dataset released by the International Skin Imaging Collaboration (ISIC), containing dermoscopic images for skin cancer detection and segmentation tasks. The dataset includes various skin conditions such as melanoma, nevus, and keratosis, with pixel-wise segmentation masks provided. For this dataset, we use 810 cases for training, 90 for validation, and 379 for testing.
- *COVID-QU-Ex* (Chest X-ray) is a radiology dataset for COVID-19, designed for the detection and segmentation of lung infection and abnormal regions. It contains X-ray images covering normal lungs, lung opacities, viral pneumonia, and COVID-19 cases, with pixel-wise segmentation masks provided. We split it into 16,280 / 1,372 / 957 cases for training / evaluation / testing.
- *CT Lung & Heart & Trachea Segmentation* (Chest CT) is a medical CT image segmentation dataset designed for anatomical segmentation of the lungs, heart and trachea. It contains CT scans from 107 patients, with pixel-wise segmentation masks consists of CT scans with segmentation masks for fibrotic lung lesions, collected from 107 patients. It is divided into 7,959 slices for training, 3,010 for validation, and 1,800 for testing.

To evaluate the model's generalization in zero-shot segmentation, we designed experiments targeting two challenges: cross-modality and cross-target segmentation. For the cross-target setting, the model was trained on ISIC, Chest X-ray, and Chest CT and tested on unseen polyp images, and trained on Kvasir-SEG, Chest X-ray and Chest CT and tested on unseen skin lesion images. For the cross-modality setting, the model was trained on Kvasir-SEG, ISIC and Chest CT and tested on unseen X-ray images; similarly, it was trained on Kvasir-SEG, ISIC, and Chest X-ray and tested on unseen CT images. This setup systematically evaluates the model's ability to segment new modalities and targets it has never seen during training.

Task	text prompt	β	σ	ℓ
Kvasir-SEG	polyp	0.1	2	8
ISIC	skin melanoma	0.1	2.0	10
Chest X-ray	lung	0.1	1.0	8
Chest CT	lung	0.1	1.0	8

Table 1. Setting of modality-specific hyperparameters on four datasets.

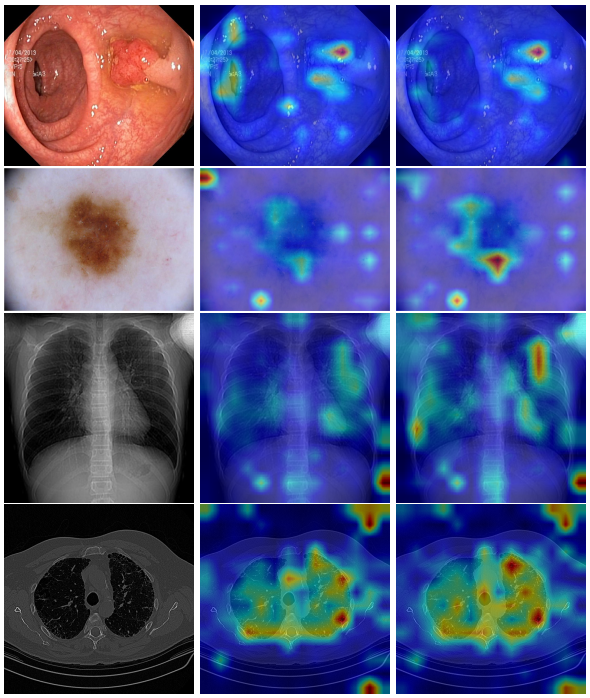

Figure 1. Visualization of initial attribution maps under different text prompts across four datasets.

2. Appendix B

2.1. Experimental Details

For BiomedCLIP, input images are resized to 224 × 224 pixels, while for SAM, they are resized to 1024 × 1024 pixels.

Dataset	Simple Text	Clinical Description Examples
Kvasir-SEG	"polyp"	"The image shows a polyp in the colon. It appears as a raised, slightly irregular lesion with a reddish hue. The polyp is located in the colon, and its size and shape are visible.", "The image shows a well-defined, round, and slightly elevated polyp in the colon. It appears to be a relatively large polyp, approximately 1.5 cm in diameter, with a smooth surface. The surrounding colon mucosa appears normal."
ISIC	"skin melanoma"	"The image shows a skin lesion with irregular borders and a mottled, brown-red appearance. The lesion exhibits a heterogeneous texture, with areas of increased and decreased pigmentation. The lesion is well-defined, but the overall appearance suggests a potential concern for melanoma.", "The image shows a skin lesion with irregular borders and a mottled, heterogeneous appearance. The lesion exhibits a mix of colors, including brown, black, and possibly red or blue hues. The lesion has a somewhat raised or irregular surface texture."
Chest X-ray	"lung"	"The lungs appear somewhat hazy and indistinct, with increased opacity in the left lung field. There is a possible area of consolidation or fluid accumulation in the lower left lung. The right lung appears relatively clear, though there may be some subtle increased markings.", "The chest X-ray shows diffuse bilateral infiltrates, likely representing pulmonary edema or pneumonia. The lung markings are prominent and indistinct, obscuring the normal lung architecture."
Chest CT	"lung"	"The lungs appear relatively clear with no obvious large consolidations, effusions, or masses. The lung markings are present, and the overall density is consistent with normal lung tissue.", "The lungs appear relatively clear with normal lung markings visible. There are no obvious consolidations, masses, or significant pleural effusions."

Table 2. Examples of text prompt for different datasets.

For the training protocol, we employ distinct configurations for zero-shot and few-shot learning. In zero-shot mode, training proceeds for 100 epochs with an effective batch size of 8 (physical batch size is 1 and gradient accumulation is 8 steps) using Adam optimization with learning rates of 2e-4 for the LoRA module and 1e-4 for the main model, weight decay of 1e-4, and a cosine learning rate schedule with a 5% warmup phase. For few-shot adaptation (k=10), we fine-tune the model for 50 epochs using Adam with learning rates of 1e-4 (LoRA) and 5e-5 (main model), applying a step decay scheduler that halves the learning rate every 10 epochs to enhance convergence while retaining generalization. We adopt the same LoRA settings as SAMed and H-SAM, in which the rank of LoRA is set to 4. Both configurations maintain the same parameter-efficient fine-tuning approach through LoRA modules.

Following ProxyCLIP, the threshold ϵ is set to 0 to retain only positive similarities, thereby filtering out negative correlations in the attribution map. For the implementation of M2IB, we introduce an information bottleneck into a designated layer ℓ of the CLIP image encoder, and train it following the same procedure as the Multi-Modal Information Bottleneck in the original IBA framework. Among the modality-specific hyperparameters, the scaling factor β modulates the trade-off between informativeness and compression in the IB objective: a larger β enforces stronger compression by limiting the information flow through this layer. The noise parameter σ controls the magnitude of Gaussian noise added to the intermediate representation, and also indirectly affects the compression term. Specifically, smaller σ increases the KL divergence, thereby intensifying the compression effect. Therefore, β and σ are coupled in practice. The choice of insertion layer ℓ further influences the resulting attribution: inserting the bottleneck too early may hinder the model's ability to capture high-level semantics, while too late reduces the bottleneck's impact. Following grid search experiments as reported in the original M2IB work, exploring combinations of $\beta \in \{1, 1.0, 2.0\}$ and $\sigma \in \{1, 1.0, 2.0\}$, and layer index $\ell \in \{8, 9, 10\}$, we select the different optimal combinations for different dataset, as shown in Table 1.

LoRA Adapters	IPP	BID	ISIC		Chest X-ray		Chest CT	
			Dice% ↑	IoU% ↑	Dice% ↑	IoU% ↑	Dice% ↑	IoU% ↑
✗	✗	✗	70.84	61.36	60.85	50.76	66.45	53.52
✗	✓	✗	73.24	65.20	75.13	62.59	76.75	64.61
✗	✓	✓	75.82	66.59	76.47	65.90	78.79	67.78
✓	✗	✗	74.24	67.20	77.50	64.83	80.07	74.03
✓	✗	✓	75.37	68.31	78.43	66.84	82.33	77.07
✓	✓	✗	78.50	69.56	81.23	69.95	90.45	84.40
✓	✓	✓	**79.65**	**70.67**	**82.60**	**74.82**	**93.62**	**89.25**

Table 3. Results of ablation study on different datasets.

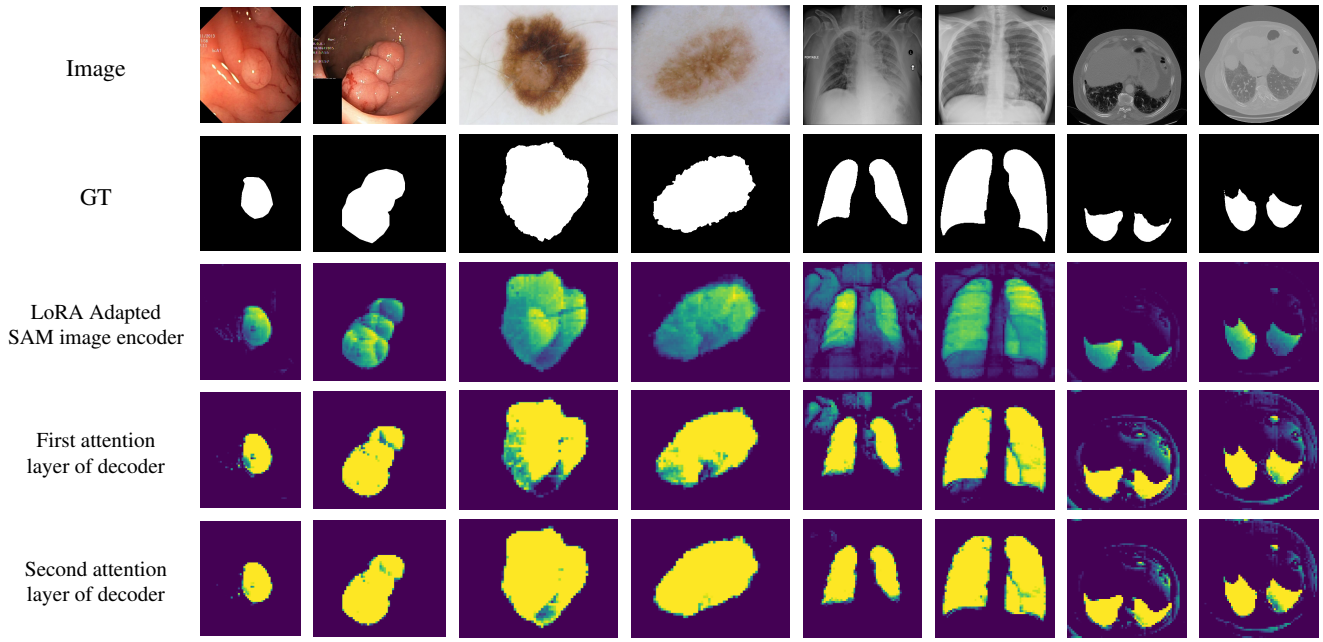

Figure 2. Attention features from different layers.

3. Appendix C

3.1. Text Prompt Design

We conducted experiments to examine how different text prompt designs affect zero-shot segmentation performance, comparing simple prompts that directly name the object to be segmented, with complex prompts written as longer, clinically descriptive sentences. Representative examples of both prompt types are presented in Table 2. To further analyze their behavioral differences, we visualize the initial attribution maps generated by CLIP under both prompt types. As shown in Fig. 1, prompts with long clinical descriptions tend to produce stronger activations over the target regions and suppress responses in irrelevant areas, likely due to their richer semantic context.

For instance, when localizing polyps, the simple prompt produces spurious activations in the upper-left region, which are far from the true target, whereas the descriptive prompt eliminates such noise. A similar pattern is observed in skin melanoma localization: the descriptive prompt accurately highlights part of the lesion, while the simple prompt fails to activate the target and instead responds strongly to non-lesion regions. These findings highlight the necessity of designing strategies that enhance attribution quality under simple prompts. Although clinically descriptive prompts can offer semantic advantages, they are more costly to obtain and less practical in real-world settings, where simple prompts are more accessible and better aligned with clinical usage.

4. Appendix D

4.1. Ablation of Multi-modal Information Bottleneck

To validate the effectiveness of our proposed Multi-modal Information Bottleneck (M2IB) module, we conducted an ablation study by replacing it with the classic Attention Rollout method for generating initial attribution maps A_{init}. All other components and training configurations remained unchanged to ensure a fair comparison.

As shown in Table 4, substituting M2IB with Attention Rollout leads to consistent performance degradation across all four datasets. Specifically, the Dice scores drop by 2.34%, 2.11%, 1.71%, and 2.75% on Kvasir-SEG, ISIC, Chest X-ray, and Chest CT datasets, respectively. These results demonstrate that M2IB effectively filters redundant background information in medical images and produces more accurate attribution maps, which is crucial for the subsequent segmentation process. The consistent improvements across different imaging modalities (endoscopy, dermoscopy, and radiology) further confirm the robustness and generalizability of the method in our approach.

Table 4. Ablation study of M2IB module. A_{init}: initial attribution map.

A_{init} Generation	Kvasir-SEG	ISIC	Chest X-ray	Chest CT
Attention Rollout	57.49	77.54	80.89	90.87
M2IB	59.83	79.65	82.60	93.62

4.2. Ablation Analysis

To gain deeper insights into the effectiveness and interplay of the proposed modules, we further analyze the ablation results in Table 3 across three distinct medical segmentation tasks: melanoma segmentation on ISIC, lung field segmentation on Chest X-ray and Chest CT. These tasks differ substantially in image modality and contrast distribution, providing a comprehensive evaluation of each component's generalization capability.

The introduction of LoRA-based domain adapters demonstrated stable improvements across all datasets. Specifically, on ISIC, Dice score increased from 70.84% to 74.24% when LoRA was activated, while Chest X-ray improved from 60.85% to 77.50%, and Chest CT from 66.45% to 80.07%. The particularly substantial gains on X-ray and CT modalities suggest that SAM's original ViT-based encoder struggles to represent medical grayscale distributions and subtle organ boundaries, whereas LoRA effectively adapts mid-level attention patterns to better capture medical-specific textures and intensity transitions.

Implicit Pos-Prompter (IPP) consistently enhanced segmentation performance by injecting coarse spatial priors derived from attribution maps. On ISIC, enabling IPP increased Dice from 74.24% to 78.50% when combined with LoRA, while on Chest X-ray, Dice improved from 77.50% to 81.23%. The effect was particularly pronounced on Chest CT, where Dice rose from 80.07% to 90.45%. Since IPP provides SAM with implicit location hints without explicit user prompts, it effectively resolves the ambiguity inherent in zero-shot settings. The stronger improvement on CT images can be attributed to the high structural regularity of lung regions, where IPP successfully highlights the target organ, enabling the decoder to focus on structure-consistent areas while suppressing background noise.

The Bidirectional Interaction Decoder (BID) provided the most substantial single-module improvement by enabling two-way information flow between mask tokens and image features. Without BID, Chest X-ray segmentation with LoRA and IPP yielded 81.23% Dice; with BID enabled, the score increased to 82.60%, while IoU improved from 69.95% to 74.82%, demonstrating notable enhancement in boundary accuracy. Similarly, on CT, Dice increased from 90.45% to 93.62%, indicating that BID is particularly effective on high-resolution volumetric-like slices where richer contextual aggregation is beneficial. These gains confirm that bidirectional cross-attention significantly enhances fine-grained mask refinement, especially in tasks requiring precise edge delineation.

The full model configuration demonstrated synergistic improvements from all three modules. On ISIC, Dice scores progressed monotonically from 70.84% (baseline) to 75.37% (LoRA), 78.50% (LoRA+IPP), and finally 79.65% (full model). Similar progressive trends were observed on Chest X-ray and Chest CT. The consistent additive gains across modalities indicate that LoRA, IPP, and BID address complementary aspects of zero-shot medical segmentation: LoRA adapts visual representations, IPP introduces localization priors, and BID strengthens feature-mask reasoning. This complementary nature is further evidenced by the full model achieving the highest improvements in both Dice and IoU metrics.

Cross-task analysis revealed that each module's effectiveness correlates with specific dataset characteristics. The most substantial LoRA gains occurred in Chest X-ray and CT datasets, where pixel statistics diverge most significantly from natural images. IPP showed maximal improvement on Chest CT, where targets occupy relatively stable anatomical regions. BID demonstrated its strongest impact on high-resolution CT slices, where fine structural detail is particularly critical. These patterns collectively validate the robustness of our proposed framework and underscore the necessity of integrating domain-adaptive features, spatial priors, and enhanced decoder interactions to achieve robust zero-shot generalization across diverse medical imaging modalities.

4.3. Qualitative Results

To illustrate the effectiveness of our implicit positional prompts and how the Bidirectional Interaction Decoder (BID) progressively refines attention to target regions, we visualize intermediate attention features in Fig. 2. The first row presents representative samples from four diverse medical segmentation, while the second row shows the corresponding ground truth (GT) masks. The third row depicts a single feature channel from the final layer of the LoRA adapted SAM image encoder, which broadly highlights semantic regions of interest but also activates irrelevant background areas, especially for small or low-contrast targets.

As shown in the fourth and fifth row of the Fig. 2, it qualitatively demonstrates how the attention becomes noticeably more focused on true target regions, suppressing background noise, after introducing the attribution map in the first cross-attention layer, where SAM features serve as queries and attribution maps as keys and values. In the second attention layer, information flows back from the attribution map to the SAM features. These attention maps are more compact and exhibit stronger correspondence to GT boundaries, demonstrating improved boundary sensitivity for larger organs and enhanced localization for small lesions.

These results highlight that integrating attribution maps as semantic priors effectively guides the BID to refine spatial focus and enhance boundary precision, ultimately improving segmentation accuracy in challenging medical scenarios involving small structures, low contrast, or complex textures.

4.4. Computational Cost Analysis

We further evaluate the computational efficiency of our method by benchmarking it against two existing approaches that also leverage CLIP and SAM architectures. The comparison encompasses three key metrics: inference latency (milliseconds per sample), throughput (samples processed per second), and peak GPU memory usage (megabytes). All experiments were conducted under identical hardware configurations with consistent batch sizes to ensure fair evaluation.

As presented in Table 5, our method demonstrates superior computational efficiency across multiple dimensions. In comparison to SaLIP, our approach achieves a substantial reduction in inference latency, from 4113.09 ms to 503.10 ms per sample, while simultaneously delivering an eightfold improvement in throughput. Although MedCLIP-SAM v2 exhibits marginally lower peak memory consumption, our method significantly outperforms it in both inference speed and processing capacity, with latency reduced by more than half and throughput nearly tripled. Our method operates with a peak memory footprint of 7910.61 MB, which is slightly above the 7243.30 MB required by MedCLIP-SAM v2. This trade-off, however, yields considerably faster inference and higher throughput, underscoring the superior overall efficiency of our approach.

Table 5. Comparison of inference efficiency and GPU memory usage

Method	Latency (ms / sample)	Throughput (samples/s)	Peak memory (MB)
SaLIP	4113.09	0.24	11614.38
MedCLIP-SAM v2	1349.85	0.74	7243.30
Ours	503.10	1.99	7910.61

5. Appendix E

5.1. Failure Case Analysis

To comprehensively evaluate the robustness of our approach under adverse conditions, we conduct an in-depth analysis of scenarios on ISIC dataset where the initial CLIP-based localization proves insufficient. These cases are particularly instructive as they reveal the complementary strengths of our proposed modules.

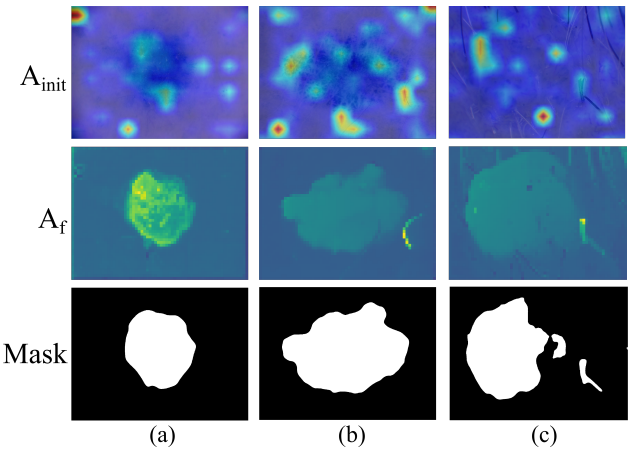

Figure 3. Illustrative case studies on the ISIC dataset. A_{init}: initial attribution map. A_f: refined attribution map.

As visualized in Figure 3(a), certain lesions with ambiguous boundaries or atypical presentations pose difficulties for the CLIP-based initialization. In this instance, the initial attribution maps may only partially cover the target region or include irrelevant background areas. Nevertheless, the proposed Implicit Pos-Prompter (IPP) demonstrates the ability to progressively refine the localization cues and highlight remaining challenges for future work.

Figure 3(b) examines scenario where IPP-generated attribution maps, while improved over the initial CLIP estimates, still exhibit imperfections such as fragmented coverage or boundary imprecision. In these scenarios, the Bidirectional Interaction Decoder (BID) plays a crucial role.

Through cross-attention–based feature interaction, BID effectively integrates global semantic priors with fine-grained structural details, allowing the model to refine boundary structures and suppress residual noise. The design of IPP followed by BID therefore forms a complementary pipeline: IPP focuses on generating reliable coarse localization cues, while BID further enhances spatial coherence and boundary accuracy.

Figure 3(c) presents the most challenging category of case, where both IPP and BID ultimately fail to produce acceptable segmentation results. These difficult instances typically involve lesions with extreme variations in appearance or ambiguous boundaries that challenge even expert interpretation. These observations suggest potential directions for future work, such as incorporating richer contextual priors, leveraging domain-specific medical knowledge, or introducing more advanced multimodal interaction mechanisms to improve robustness under highly complex conditions.

Overall, the qualitative analysis demonstrates the complementary roles of the proposed modules. The IPP module primarily enhances semantic localization by generating refined implicit positional prompts from multimodal information, whereas the BID module focuses on structural refinement through bidirectional feature interaction. Although the framework shows strong robustness across diverse clinical scenarios, the remaining failure cases highlight opportunities for further improving the integration of semantic and structural cues in medical image segmentation.