BAPM: Boundary-Aware Prompt Mining for Training-Free Few-Shot Medical Image Segmentation

Qi Wen[1], **Jiezhou He**[2,*], **Zhiming Luo**[1,3,*], **Shaozi Li**[4]

[1]Department of Artificial Intelligence, Xiamen University, Xiamen, China
[2]School of Aerospace Engineering, Xiamen University, Xiamen, China
[3]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, China
[4]Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University, Wuyishan, Fujian, China

wenqi7@stu.xmu.edu.cn hdcin@xmu.edu.cn
zhiming.luo@xmu.edu.cn szlig@xmu.edu.cn

Abstract

Few-shot medical image segmentation has shown great potential in reducing annotation costs for clinical applications. Recently, many few-shot methods have explored the Segment Anything Model (SAM) for training-free medical image segmentation via prompt engineering. However, existing approaches mainly focus on locating positive prompts from support images while overlooking informative background cues, making them prone to over-segmentation in anatomically similar regions. Moreover, simply introducing negative prompts cannot effectively suppress boundary leakage, and may even interfere with SAM's mask decoding process, resulting in worse performance than using positive prompts alone. To address this limitation, we propose a training-free support-query framework, termed Boundary-Aware Prompt Mining, which introduces boundary-aware negative prompting for SAM-based few-shot medical image segmentation. Specifically, we introduce a Prototype-guided Positive Prompt strategy, which adopts a multi-center prompting mechanism to construct multiple foreground prototypes, enabling comprehensive spatial coverage. Furthermore, we propose a Boundary-band Ambiguity Filtering strategy that identifies a boundary-adjacent background band and progressively removes ambiguous pixels and unreliable clusters highly similar to the foreground, enabling the selection of reliable negative prompts for suppressing false positives near foreground boundaries. The generated positive and negative prompts are jointly fed into SAM to produce refined segmentation results without additional training. Extensive experiments on Abd-MRI and Abd-CT datasets demonstrate that our method consistently outperforms existing approaches, highlighting its effectiveness under limited annotation settings. Code will be released upon acceptance.

1 Introduction

Medical image analysis requires large amounts of labeled data, which are costly and often scarce. Few-shot medical image segmentation (FSMIS) [10, 20, 26] has shown promising progress, highlighting its potential to significantly reduce per-task annotation costs in practical clinical applications.

*Corresponding authors: Jiezhou He and Zhiming Luo

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

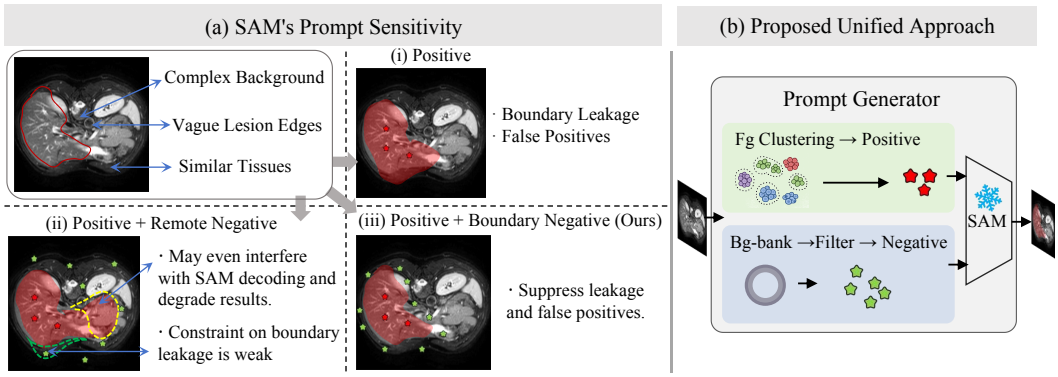


Figure 1: (a) SAM prompt sensitivity. (b) Proposed method.

Existing FSMIS methods can be broadly divided into two paradigms. The first paradigm generates pseudo labels from annotated support images to train segmentation networks, typically via image registration or prototype matching [6, 5, 12, 11, 25]. However, these methods usually require task-specific training and often struggle to generalize across diverse foreground appearances. The second paradigm leverages large pre-trained models such as the Segment Anything Model (SAM) [14], where prompts derived from support images guide segmentation inference [28, 17, 22]. To adapt SAM to medical images, some methods rely on fine-tuning [27, 19, 29, 24], introducing additional computational cost. In contrast, training-free methods directly exploit the pre-trained model [1, 28], but mainly focus on positive prompts while overlooking informative negative prompts, often resulting in false positives and over-segmentation in ambiguous regions.

Medical images exhibit ambiguous foreground boundaries and complex surrounding tissues. In SAM-based FSMIS, when only positive prompts transferred from support images are used, SAM tends to over-segment foreground regions into visually similar areas, as shown in Fig. 1(a)(i). More importantly, we observe that simply adding negative prompts does not necessarily improve performance. As shown in Fig. 1(a)(ii), introducing negative prompts typically leads to two situations: (1) remote negatives provide weak boundary constraints and fail to suppress over-segmentation; (2) some negatives interfere with SAM's mask decoding, resulting in worse performance than positive prompts alone. This suggests that negative prompts should not be arbitrary background cues. As shown in Fig. 1(a)(iii), effective negative prompts should instead be located near ambiguous foreground boundaries, where they can suppress false-positive leakage without interfering with true foreground regions.

To address this limitation, we propose a training-free support-query framework, termed Boundary-Aware Prompt Mining (BAPM), which explicitly exploits boundary-aware prompt design for SAM-based FSMIS, as illustrated in Fig. 1(b). Specifically, we propose a Prototype-guided Positive Prompt (PP) strategy, which adopts a multi-center prompting mechanism to cluster foreground regions into multiple representative prototypes, enabling comprehensive spatial coverage for positive prompt localization. In parallel, we propose a Boundary-band Ambiguity Filtering (BAF) strategy for negative prompt selection. We first construct a boundary-adjacent band via mask dilation to constrain negative prompts near the foreground boundary, and then remove pixels highly similar to the foreground through ambiguity filtering. Next, adaptive clustering is performed on the remaining background pixels, and clusters with low foreground similarity are selected and transferred to query slices via feature matching to provide reliable negative prompts. The selected positive and negative prompts are jointly fed into SAM to produce segmentation results on unlabeled slices without additional model retraining. Our contributions can be summarized as follows:

- We propose a training-free FSMIS framework, termed BAPM, which introduces boundary-aware prompt mining to address over-segmentation caused by ambiguous boundaries in SAM-based FSMIS.
- We propose a Boundary-band Ambiguity Filtering (BAF) strategy that constructs reliable boundary-adjacent negative prompts, reduces under-segmentation caused by foreground-like negative prompts while effectively suppressing over-segmentation caused by weak boundary constraints.

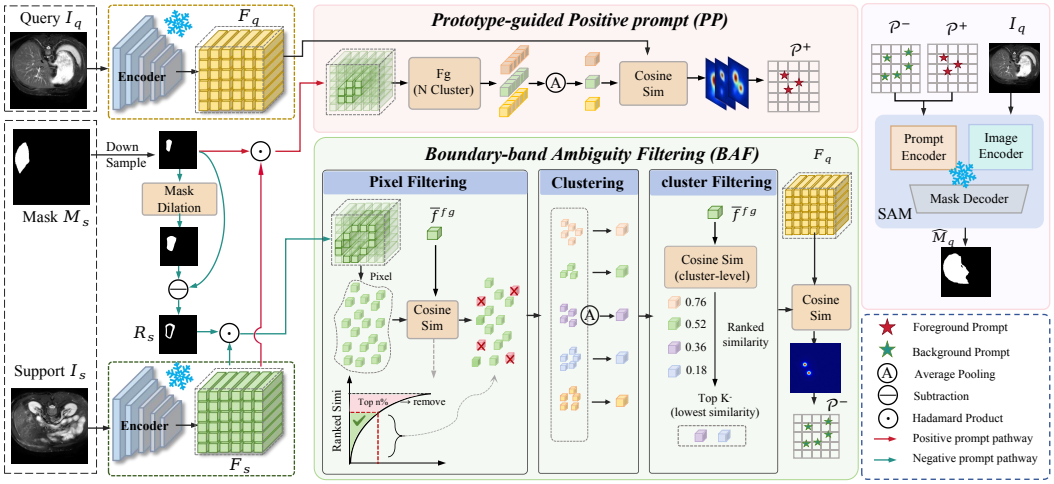


Figure 2: Overview of the proposed BAPM framework. (1) prototype-guided positive prompt (PP), (2) boundary-band ambiguity filtering (BAF), and (3) prompt-driven segmentation inference.

- Extensive experiments on the Abd-MRI and Abd-CT datasets demonstrate that BAPM outperforms existing FSMIS methods and achieves state-of-the-art.

2 Related Works

Few-shot Medical Image Segmentation. Few-shot medical image segmentation aims to learn segmentation models from only a few annotated samples. Existing methods mainly follow several representative paradigms. Prototype-based approaches extract representative prototypes from annotated support images and match them with query features to guide segmentation [33, 37, 38]. Despite their effectiveness, these methods typically require task-specific model training and may struggle with large appearance variations across foreground. Another line of work introduces self-supervised learning to alleviate the reliance on manual annotations by exploiting intrinsic image structures such as superpixels or supervoxels [20, 21, 10]. However, these approaches usually require additional pretraining stages and may not fully exploit support–query correspondence. Some methods further improve segmentation performance through iterative mask refinement or attention-based feature interaction [26, 7, 4]. Although these methods improve segmentation accuracy, they often rely on complex training pipelines and remain sensitive to appearance variations.

SAM for Few-shot Segmentation. Segment Anything Model (SAM) has recently attracted increasing attention in few-shot medical image segmentation. Existing studies mainly adapt SAM to medical imaging tasks through parameter-efficient tuning or additional training modules. Several works fine-tune SAM or introduce lightweight adaptation strategies to bridge the domain gap between natural and medical images, including adapter-based or parameter-efficient transfer learning approaches [27, 29, 24]. Other studies explore fine-tuning SAM using a small number of annotated exemplars or meta-learning strategies to improve few-shot segmentation performance [34, 27, 30, 9, 17]. In addition, prompt-driven frameworks have been proposed to better exploit SAM for medical segmentation, such as automated prompting mechanisms and multimodal vision–language guidance [22, 15, 31]. To further reduce training overhead, several works attempt to leverage SAM in a training-free manner by generating prompts directly from support information or coarse segmentation predictions [1, 28]. However, most existing prompt-based methods mainly focus on generating foreground prompts while ignoring informative background prompts, which may lead to over-segmentation.

3 Methods

3.1 Problem Definition

Following prior work on FSMIS, the task is formulated in an episode-based manner. Each episode consists of a support set $\mathcal{S}$ and a query set $\mathcal{Q}$, where $\mathcal{S} = \{(I_i^s, M_i^s)\}_{i=1}^K$ denotes K annotated

support images with corresponding segmentation masks, and $\mathcal{Q} = \{I^q\}$ denotes the query images to be segmented. I and M represent the image and its corresponding mask, respectively. The goal is to segment target structures in query images and generalize to unseen categories. Most prior works adopt the challenging 1-way 1-shot setting ($K = 1$), which is also considered in this study.

Distinct from conventional methods, we reformulate FSMIS as a prompt generation problem and employ SAM as the segmentation backbone. Given a support-query pair $(\mathcal{S}, \mathcal{Q})$, our objective is to generate a set of prompts Φ from the support image that can effectively guide segmentation on each query image $I^q \in \mathcal{Q}$. The final segmentation is obtained by feeding both the generated prompts Φ and the query image I^q into SAM.

3.2 Overview

The overall framework consists of four main components: (1) feature extraction using the frozen SAM image encoder, (2) prototype-guided positive prompt (PP), (3) boundary-band ambiguity filtering (BAF) strategy for negative prompt , and (4) prompt-driven segmentation inference. We first extract feature maps from the SAM image encoder, denoted as $F_s \in \mathbb{R}^{H\times W\times C}$ and $F_q \in \mathbb{R}^{H\times W\times C}$ for the support and query slices, respectively. The extracted features are then fed into the PP and BAF modules to produce corresponding prompts for each query slice. Finally, the generated positive and negative prompts are jointly input into the frozen SAM model to obtain segmentation masks for query slices in a training-free manner.

3.3 Prototype-guided Positive prompt (PP)

To capture the intra-foreground heterogeneity, we collect feature vectors from the foreground region of the support slice:

$$\mathcal{F}_s^{fg} = \{F_s(p) \mid M_s(p) = 1\}, \tag{1}$$

where p indexes spatial locations.

We then perform adaptive density-based clustering (e.g., DBSCAN [8]) on $\mathcal{F}_s^{fg}$ augmented with normalized spatial coordinates to obtain a set of foreground prototypes $\{\mathbf{c}_1^{fg}, \ldots, \mathbf{c}_K^{fg}\}$, where the number of clusters K is automatically determined.

For each foreground prototype $\mathbf{c}_k^{fg}$, we compute its cosine similarity with all spatial locations in the query feature map:

$$S_k^+(x) = \frac{F_q(x) \cdot \mathbf{c}_k^{fg}}{\|F_q(x)\| \, \|\mathbf{c}_k^{fg}\|}, \tag{2}$$

where x denotes a spatial location in the query slice. The location with the highest similarity is selected as the positive prompt, $x_k^+ = \arg\max_x S_k^+(x)$, and the final positive prompt set is $\mathcal{P}^+ = \{x_1^+, \ldots, x_K^+\}$.

3.4 Boundary-band Ambiguity Filtering (BAF)

Remote negative prompts often provide weak boundary constraints, leading to over-segmentation in ambiguous regions. While sampling negatives from boundary-adjacent regions helps suppress over-segmentation by providing stronger spatial constraints near foreground boundaries, it may also introduce foreground-like ambiguous pixels that incorrectly suppress true regions and cause under-segmentation. To address this challenge, we propose a Boundary-band Ambiguity Filtering (BAF) strategy that follows a filtering–clustering–selection pipeline. Specifically, we first perform pixel-level filtering to remove foreground-like ambiguous pixels, then group the remaining background pixels via clustering, and finally select reliable clusters to construct informative boundary-aware negative prompts.

Boundary-proximal band construction. Given the support mask M_s, we first resize it to the feature resolution and construct a boundary-proximal band via morphological dilation with a radius r. The candidate background region is defined as:

$$\mathcal{R}_s = \widetilde{M}_s \ominus M_s, \tag{3}$$

where $\widetilde{M}_s$ denotes the dilated support mask, $\ominus$ denotes the set difference operation.

Pixel-level ambiguity filtering. We first perform pixel-level filtering within the boundary band to remove foreground-like ambiguous pixels, retaining reliable boundary-adjacent background regions for negative prompt generation. We first obtain a mean foreground representation:

$$\bar{\mathbf{f}}^{fg} = \frac{1}{|\{q \mid M_s(q) = 1\}|} \sum_{q \in \{q | M_s(q)=1\}} F_s(q), \tag{4}$$

where $|\cdot|$ denotes the cardinality of a set.

For each pixel $p \in \mathcal{R}_s$, we compute its similarity to the foreground prototype:

$$S^{fg}(p) = \frac{F_s(p) \cdot \bar{\mathbf{f}}^{fg}}{\|F_s(p)\| \, \|\bar{\mathbf{f}}^{fg}\|}. \tag{5}$$

We rank all pixels in $\mathcal{R}_s$ according to $S^{fg}(p)$ and discard the top α_{pixel} fraction with the highest similarity values. The remaining pixels form a filtered background set: $\mathcal{R}_s^{clean} \subset \mathcal{R}_s$.

Spatially-aware clustering. To preserve spatial coherence, we perform adaptive density-based clustering on the filtered features $\{F_s(p) \mid p \in \mathcal{R}_s^{clean}\}$ augmented with normalized spatial coordinates, producing a set of background clusters $\{\mathcal{B}_1, \ldots, \mathcal{B}_N\}$.

Background prototype construction and cluster-level selection. For each background cluster $\mathcal{B}_n$, we compute a prototype by averaging its feature representations:

$$\mathbf{c}_n^{bg} = \frac{1}{|\mathcal{B}_n|} \sum_{p \in \mathcal{B}_n} F_s(p). \tag{6}$$

To measure the discriminability of each background prototype from the foreground, we compute its cosine similarity to the mean foreground feature:

$$\bar{S}_n = \frac{\mathbf{c}_n^{bg} \cdot \bar{\mathbf{f}}^{fg}}{\|\mathbf{c}_n^{bg}\| \, \|\bar{\mathbf{f}}^{fg}\|}. \tag{7}$$

Background prototypes with lower similarity are considered more reliable for suppressing over-segmentation. Therefore, we retain the top β_{cluster} fraction of prototypes with the smallest similarity values:

$$\mathcal{C}^{bg} = \text{Top}_{\beta_{\text{cluster}}}\big(\{\mathbf{c}_n^{bg}\}_{n=1}^{N};\ \text{ascending } \bar{S}_n\big). \tag{8}$$

Query matching for negative prompts. The selected background prototypes $\mathcal{C}^{bg}$ are then transferred to the query slice via feature matching. For each selected prototype $\mathbf{c}_n^{bg} \in \mathcal{C}^{bg}$, we compute its similarity map on the query feature map:

$$S_n^-(x) = \frac{F_q(x) \cdot \mathbf{c}_n^{bg}}{\|F_q(x)\| \, \|\mathbf{c}_n^{bg}\|}, \qquad x_n^- = \arg\max_x S_n^-(x). \tag{9}$$

The final negative prompt set is $\mathcal{P}^- = \{x_n^- \mid \mathbf{c}_n^{bg} \in \mathcal{C}^{bg}\}$.

3.5 Prompt-Driven Inference

Given the selected positive prompt x^+ and negative prompt x^- for a query slice I_q, we construct a point-based prompt set

$$\mathcal{P}_q = \{(x, +1) \mid x \in \mathcal{P}^+\} \cup \{(x, -1) \mid x \in \mathcal{P}^-\}, \tag{10}$$

where $+1$ and -1 denote foreground and background labels, respectively.

The prompt set $\mathcal{P}_q$ is fed into the prompt encoder of SAM together with the query image. The image encoder and mask decoder then produce the segmentation mask $\hat{M}_q$ in a single forward pass:

$$\hat{M}_q = \text{SAM}(I_q, \mathcal{P}_q). \tag{11}$$

Table 1: Comparison with state-of-the-art methods on Abd-MRI and Abd-CT datasets. "Train?" indicates whether additional task-specific training or adaptation is required.

Methods	Train?	Abd-MRI					Abd-CT				
		Liv	RK	LK	Spl	Avg.	Liv	RK	LK	Spl	Avg.
ALPNet [20]	✓	76.10	85.18	81.92	72.18	78.84	78.29	71.81	72.36	70.96	73.35
RPT [37]	✓	82.86	89.82	80.72	76.37	82.44	82.57	72.58	77.05	79.13	77.83
SnapSeg [32]	✗	-	-	-	-	-	73.93	69.28	73.03	73.50	72.43
PerSAM [35]	✗	42.44	71.36	64.84	69.14	61.12	65.55	60.31	58.47	65.03	62.34
GMRD [6]	✓	81.42	**90.12**	83.96	76.09	82.90	79.60	74.46	81.70	78.31	78.52
PGRNet [12]	✓	83.27	87.44	81.44	81.72	83.47	82.48	79.88	74.23	72.09	77.17
ProtoSAM [1]	✓	83.14	82.36	82.75	77.98	81.56	84.79	75.67	71.31	70.24	75.50
SAM2 (P2) [23]	✗	71.62	80.82	83.10	78.91	78.61	74.73	81.32	84.54	82.88	80.87
AM-SAM [22]	✓	76.12	84.95	84.17	80.36	81.40	87.28	86.01	84.37	87.11	86.19
SynPo [18]	✗	77.32	83.04	77.32	80.30	81.15	81.32	79.63	75.00	83.76	79.91
FoB [2]	✓	85.61	88.18	84.76	79.31	84.46	86.51	**86.51**	**87.29**	84.54	86.21
Ours	✗	**86.98**$\pm$1.02	88.40$\pm$1.37	**85.31**$\pm$1.56	**84.77**$\pm$0.63	**86.37**$\pm$0.77	**87.85**$\pm$1.18	85.22$\pm$1.24	85.67$\pm$0.69	**87.87**$\pm$0.72	**86.65**$\pm$0.82

4 Experiments

Datasets We conducted comprehensive evaluations on two datasets spanning different imaging modalities and anatomical regions, including Abd-CT [16] and Abd-MRI [13]. Specifically, Abd-CT consists of 30 volumetric CT scans, while Abd-MRI contains 20 3D T2-SPIR MRI volumes. For fair comparison, following FoB [2], we focus on four shared organ categories across both abdominal datasets: liver (Liv), right kidney (RK), left kidney (LK), and spleen (Spl).

Implementation details We use the ViT-H variant of SAM as the backbone image encoder with frozen pre-trained weights, without any fine-tuning. Cosine similarity with L2-normalized features is used for support–query matching. Our method is implemented using PyTorch and evaluated on a workstation equipped with two NVIDIA RTX A4000 GPUs (16GB) with CUDA 12.6. To ensure a fair comparison, we employ the Dice coefficient as the evaluation metric. All results are averaged over five independent runs with different random support selections, and standard deviations are reported where applicable.

4.1 Comparison with state-of-the-art methods

We compare our BAPM model with a diverse set of state-of-the-art FSMIS methods, including training-based approaches such as ALPNet [20], RPT [37], GMRD [6], PGRNet [12], ProtoSAM [1], AM-SAM [22], and FoB variants [2], as well as training-free methods including PerSAM [35], SAM2 (P2) [23], SynPo [18], and SnapSeg [32]. Our experimental results follow the evaluation protocols of SynPo [18], and FoB [2]. All experiments are conducted under the 1-way 1-shot (Setting 1). For fair comparison, we adopt the same data splits and evaluation settings as in FoB. The quantitative results are summarized in Tab. 1.

(1) Overall Performance. Our method achieves the best performance on both Abd-MRI and Abd-CT datasets, reaching Dice scores of 86.37% and 86.65%, respectively. Compared with the second-best method FoB, our approach improves performance by 1.91% on Abd-MRI and 0.44% on Abd-CT.

(2) Comparison under Training-Free Setting. More importantly, our method is entirely training-free, yet still surpasses existing methods that require task-specific training or fine-tuning. Compared with SAM-based few-shot methods that require task-specific training or adaptation, such as ProtoSAM and AM-SAM, our method achieves consistent performance gains improving Dice scores by +4.81% over ProtoSAM and +4.97% over AM-SAM on Abd-MRI. Moreover, compared with training-free methods such as SAM2 (P2), our approach still achieves superior performance, yielding Dice improvements of +7.76% on Abd-MRI and +5.78% on Abd-CT.

(3) Effectiveness of Boundary-Aware Negative Prompts. Compared with recent prompt-based methods that primarily focus on positive prompt design, our approach achieves more precise segmentation by explicitly modeling boundary-aware background band. ProtoSAM and AM-SAM rely mainly on foreground cues, which often lead to over-segmentation around ambiguous foreground boundaries (81.56% and 81.40% Dice on Abd-MRI), whereas our method achieves 86.37% Dice by incorporating boundary-aware negative prompts. This performance gain stems from our ability to explicitly suppress ambiguous background regions near foreground borders.

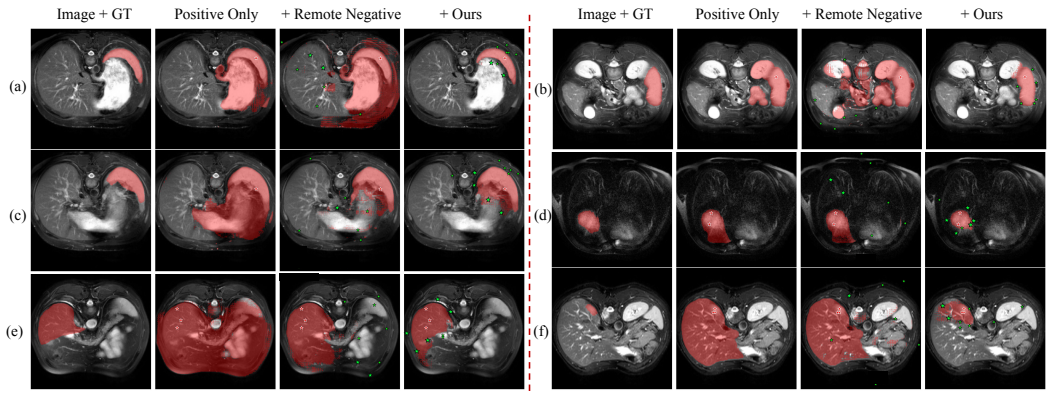


Figure 3: Qualitative results on six representative cases from the Abd-MRI dataset. Red stars denote positive prompts, and green stars denote negative prompts.

Table 2: Ablation study of different prompt design strategies on the Abd-MRI dataset. Pos: positive prompts only. Remote: randomly sampled background negatives. Band: boundary-band negative prompts. Pixel-Filter: pixel-level ambiguity filtering. Cluster-Filter: cluster-level filtering.

Strategy	Liv	RK	LK	Spl	Avg.
Pos	56.10	77.40	70.57	73.25	69.33
Pos + Remote	41.57	56.29	48.67	59.96	51.62
Pos + Band	61.18	78.28	72.48	66.36	69.58
Pos + Band + Pixel-Filter	84.38	85.88	83.36	82.61	84.06
Pos + Band + Pixel-Filter + Cluster-Filter (Ours)	86.98	88.40	85.31	84.77	86.37

4.2 Ablation Study

4.2.1 Ablation study of different prompt design strategies.

We conduct module-level ablation to evaluate different prompt designs, as summarized in Tab. 2. The following observations can be made:

(1) Positive prompts alone lead to over-segmentation. Using only positive prompts results in a relatively low Dice score (69.33%), indicating that foreground cues alone are insufficient to constrain the segmentation. As shown in Fig. 3(a–d), this leads to evident over-segmentation, where predictions leak into surrounding background regions.

(2) Naive negative prompts may be ineffective or even harmful. Remote negative prompts provide weak spatial constraints and may even degrade performance compared with positive-only prompting (51.62% / 69.33%), as they can interfere with SAM's mask decoding (Fig. 3(e–f)).

(3) Boundary-aware negatives effectively constrain over-segmentation. Sampling negative prompts near foreground boundaries significantly improves Dice to 69.58%, outperforming both positive-only and remote-negative settings. This demonstrates that boundary-aware negatives provide more reliable spatial constraints for suppressing over-segmentation.

(4) Ambiguity-aware filtering further improves performance. Applying pixel- and cluster-level filtering further boosts performance to 86.37%. By removing ambiguous background pixels and selecting reliable negative regions, the model achieves better boundary delineation and reduces false positives.

4.2.2 Parameter Ablation

Effect of Feature Layer Selection. We investigate the impact of feature layer selection from the frozen SAM image encoder on segmentation performance. Specifically, features are extracted from different transformer layers (L1, L8, L16, and L32) of the SAM ViT-H encoder, corresponding to shallow, intermediate, and deep representations. As shown in Tab. 3(a), shallow and intermediate layers achieve better performance than deeper layers. In particular, L8 achieves the best overall performance, obtaining the highest Dice scores on RK (88.40%), LK (85.31%), and Spl (84.77%), while L1 achieves the best result on Liver (86.98%). In contrast, deep features from L32 lead to a significant

Table 3: Ablation studies on Abd-MRI (Dice %).

(a) Feature layer selection.

Layer	Liv	RK	LK	Spl	Avg.
L1	**86.98**	85.01	81.55	83.50	84.26
L8	85.71	**88.40**	**85.31**	**84.77**	**86.05**
L16	85.43	86.41	82.96	83.59	84.60
L32	80.34	79.80	74.00	76.18	77.58
L(1+8)	84.75	86.81	83.20	83.91	84.67
L(1+8+16)	85.47	87.12	84.16	83.88	85.16
L(1+8+16+32)	81.47	87.20	83.35	82.96	83.75

(b) Cross-modality comparison.

	Method	Liv	RK	LK	Spl	Avg.
CT→MRI	RobustEMD [36]	60.16	70.26	66.34	53.71	62.61
	FAMNet [3]	73.01	74.68	57.28	58.21	65.79
	FoB [2]	75.05	79.57	70.38	68.21	73.30
	Ours	**82.86**	**81.42**	**80.78**	**80.11**	**81.29**
MRI→CT	RobustEMD [36],	69.82	50.34	63.79	59.88	60.95
	FAMNet [3]	73.57	61.89	57.79	65.78	64.75
	FoB [2]	81.36	58.81	57.18	70.71	67.02
	Ours	**81.94**	**80.61**	**80.92**	**81.85**	**81.33**

(c) Support slice selection on Abd-MRI.

Strategy	Liv	RK	LK	Spl	Avg.
MaxArea	**88.12**	**88.35**	**86.73**	**86.01**	**87.30**
MinArea	70.85	72.14	68.90	69.77	70.42
Rand	85.12	84.90	83.41	82.73	84.04
Rand-High	86.98	88.40	85.31	84.77	86.37
Rand-Low	80.45	81.20	78.96	79.10	79.93

(d) Effect of boundary band radius r.

Radius r	Liv	RK	LK	Spl	Avg.
$r = 2$	84.39	87.36	83.25	83.45	84.61
$r = 3$	85.61	**88.40**	**85.31**	**84.77**	**86.02**
$r = 5$	**86.98**	87.50	84.60	84.26	85.84
$r = 7$	85.57	87.28	83.77	83.69	85.08

performance drop, with the average Dice decreasing to 77.58%. We further explore multi-layer feature fusion by concatenating features from different layers. However, the fused representations do not outperform the single intermediate layer (L8), suggesting that directly combining features from substantially different depths may introduce redundant or inconsistent information. These results indicate that shallow and intermediate layers preserve richer spatial and boundary details, which are more suitable for prototype-based similarity matching and accurate boundary localization in FSMIS.

Effect of support slice selection. We further investigate how different support slice selection strategies affect segmentation performance, including selecting slices with the largest area (MaxArea), smallest area (MinArea), random foreground slices (Rand), high-area slices (Rand-High), and low-area slices (Rand-Low). All results reported in previous sections are obtained under the Rand-High protocol and averaged over five independent runs to ensure statistical stability, while results for random selection strategies are also averaged over multiple independent runs for fair and stable evaluation. This setting reflects practical annotation scenarios, where slices with clear and prominent foreground appearances are more likely to be selected as representative annotations.

As shown in Tab. 3(c), MaxArea achieves the best Dice across all organs (88.12% for liver, 88.35% for RK, 86.73% for LK, and 86.01% for spleen), while MinArea significantly degrades performance, indicating that small or ambiguous regions provide insufficient guidance for reliable segmentation. Compared with purely random selection (Rand), Rand-High yields consistently better results (e.g., +1.86% on liver, +3.50% on RK, +1.90% on LK, and +2.04% on spleen). These observations demonstrate that selecting support slices with sufficiently large and clear target regions is crucial for providing discriminative guidance to the model.

Effect of Boundary-aware Background Region. We further investigate the influence of the boundary-aware background region by varying the radius (r), which controls how far background candidate regions are selected from the target boundary. As shown in Tab. 3(d), the optimal radius varies across organs: the best performance is achieved at $r = 5$ for the liver (86.98%), while the right kidney (RK), left kidney (LK), and spleen (Spl) achieve their optimal performance at $r = 3$ (88.40%, 85.31%, and 84.77%, respectively). Overall, $r = 3$ yields the highest average Dice score (86.02%). This variation is related to anatomical characteristics. Larger organs such as the liver benefit from a slightly larger radius to capture more informative background context, while kidneys and spleen favor boundary-adjacent regions for more discriminative negative prompts.

To further analyze the effect of r, we visualize the distance distribution of selected negative prompts to the organ boundary on LK in Fig. 4. When $r = 3$, most negative prompts are concentrated within 0–10 pixels from the boundary, indicating strong boundary awareness. As r increases, more prompts shift to farther regions, which may weaken their ability to suppress boundary leakage.

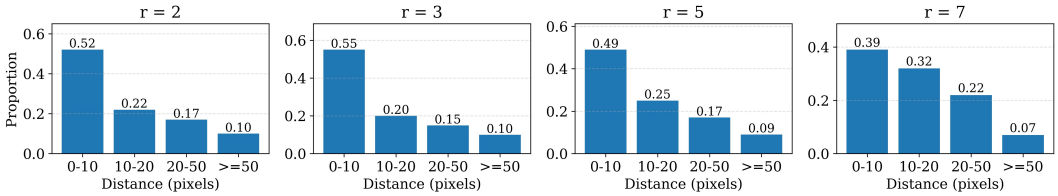


Figure 4: Distribution of negative prompt distances to the organ boundary under different boundary band radius r on LK. The x-axis denotes the distance interval to the boundary, and the y-axis denotes the proportion of negative prompts.

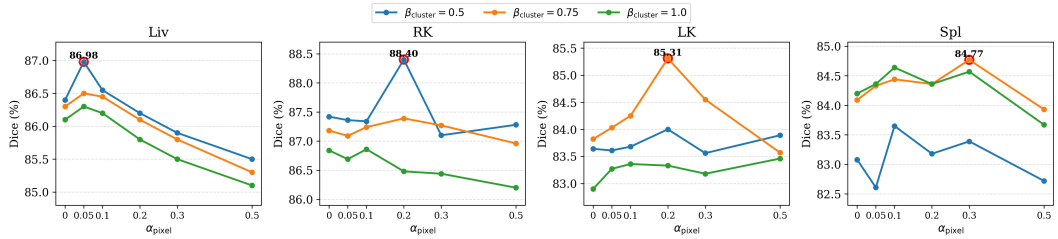


Figure 5: Effect of α_{pixel} and β_{cluster} on Abd-MRI.

Effect of Pixel and Cluster Filtering. We further analyze the impact of the pixel-level filtering ratio (α_{pixel}) and the cluster-level retention ratio (β_{cluster}), which control ambiguous pixel removal and background prototype retention, respectively. As shown in Fig. 5, the optimal α_{pixel} varies across organs. Liver achieves the best performance at $\alpha_{\text{pixel}} = 0.05$ (86.98%), RK and LK at 0.2 (88.04% and 85.31%), while Spl reaches the best result at 0.3 (84.77%). Regarding β_{cluster}, smaller or moderate retention ratios (0.5 or 0.75) generally achieve better performance than $\beta_{\text{cluster}} = 1.0$. These results suggest that removing highly ambiguous background regions is important for constructing discriminative negative prompts. However, excessively retaining background clusters may introduce noisy or redundant information, leading to degraded segmentation performance.

4.2.3 Cross-modality Generalization

We further evaluate cross-modality generalization under CT→MRI and MRI→CT settings by comparing our method with representative few-shot segmentation approaches, including RobustEMD [36], FAMNet [3], and FoB [2]. As shown in Tab. 3(b), existing training-based methods suffer from substantial performance degradation under cross-modality settings due to their reliance on modality-specific representations learned from source-domain training data. In contrast, our method consistently achieves the best performance across all organs, obtaining average Dice scores of 81.29% and 81.33% under the two settings, respectively. We attribute this strong generalization ability to the frozen SAM encoder pretrained on large-scale image distributions, which provides more transferable and modality-agnostic representations. Moreover, our training-free prototype matching and ambiguity-aware prompt mining strategy avoids modality-specific optimization, making the framework more robust to the significant appearance discrepancies between CT and MRI.

5 Conclusion

We propose a training-free FSMIS framework that explicitly incorporates background-aware prompt generation for SAM-based medical image segmentation. By leveraging the annotated support slice, the proposed method constructs discriminative prompts for unlabeled query slices without requiring additional model training. Specifically, foreground prototypes are used to generate reliable positive prompts, while a boundary-aware background modeling strategy is introduced to mine informative negative prompts that effectively suppress false activations. Extensive experiments on multiple medical datasets demonstrate that the proposed approach consistently improves segmentation performance over existing prompt-based methods. However, the proposed method may still face challenges in accurately localizing extremely small lesions due to the limited discriminability of sparse foreground cues and the intrinsic resolution constraints of SAM feature representations. In

future work, we plan to incorporate textual information, such as anatomical descriptions and clinical priors, to further improve prompt generation and segmentation robustness.

Acknowledgments and Disclosure of Funding

This work is supported by the National Natural Science Foundation of China (No. 62276221 and 62376232), the National Key R&D Program of China (No. 2023YFC3503004), the Open Project Program of Fujian Key Laboratory of Big Data Application and Intellectualization for Tea Industry, Wuyi University (FKLBDAITI202505) and the Postdoctoral Fellowship Program of CPSF under Grant Number GZC20260864.

References

[1] L. Ayzenberg, R. Giryes, and H. Greenspan. Protosam: One-shot medical image segmentation with foundational models. *arXiv preprint arXiv:2407.07042*, 2024.

[2] Y. Bo, Y. Zhu, P. Koniusz, and H. Zhang. Focus on background: Exploring sam's potential in few-shot medical image segmentation with background-centric prompting. *arXiv preprint arXiv:2603.21287*, 2026.

[3] Y. Bo, Y. Zhu, L. Li, and H. Zhang. Famnet: Frequency-aware matching network for cross-domain few-shot medical image segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 1889–1897, 2025.

[4] V. I. Butoi, J. J. G. Ortiz, T. Ma, M. R. Sabuncu, J. Guttag, and A. V. Dalca. Universeg: Universal medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21438–21451, 2023.

[5] Z. Cheng, S. Wang, Y. Long, T. Zhou, H. Zhang, and L. Shao. Dual interspersion and flexible deployment for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 2025.

[6] Z. Cheng, S. Wang, T. Xin, T. Zhou, H. Zhang, and L. Shao. Few-shot medical image segmentation via generating multiple representative descriptors. *IEEE Transactions on Medical Imaging*, 43(6):2202–2214, 2024.

[7] H. Ding, C. Sun, H. Tang, D. Cai, and Y. Yan. Few-shot medical image segmentation with cycle-resemblance attention. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2488–2497, 2023.

[8] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 226–231. AAAI Press, 1996.

[9] W. Feng, L. Zhu, and L. Yu. Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars. In *International MICCAI Brainlesion Workshop*, pages 13–22. Springer, 2023.

[10] S. Hansen, S. Gautam, R. Jenssen, and M. Kampffmeyer. Anomaly detection-inspired few-shot medical image segmentation through self-supervision with supervoxels. *Medical Image Analysis*, 78:102385, 2022.

[11] S. Huang, T. Xu, N. Shen, F. Mu, and J. Li. Rethinking few-shot medical segmentation: a vector quantization view. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3072–3081, 2023.

[12] W. Huang, J. Hu, J. Xiao, Y. Wei, X. Bi, and B. Xiao. Prototype-guided graph reasoning network for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(2):761–773, 2024.

[13] A. E. Kavur, N. S. Gezer, M. Barış, S. Aslan, P.-H. Conze, V. Groza, D. D. Pham, S. Chatterjee, P. Ernst, S. Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical image analysis*, 69:101950, 2021.

[14] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[15] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao. Medclip-sam: Bridging text and image towards universal medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 643–653. Springer, 2024.

[16] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.

[17] T. Leng, Y. Zhang, K. Han, and X. Xie. Self-sampling meta sam: enhancing few-shot medical image segmentation with meta-learning. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 7925–7935, 2024.

[18] Y. Liu, H. Xiao, J. Chai, Y. Zhang, R. Wang, Z. Meng, and Z. Luo. Synpo: Boosting training-free few-shot medical segmentation via high-quality negative prompts. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 594–603. Springer, 2025.

[19] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang. Segment anything in medical images. *Nature communications*, 15(1):654, 2024.

[20] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *European conference on computer vision*, pages 762–780. Springer, 2020.

[21] C. Ouyang, C. Biffi, C. Chen, T. Kart, H. Qiu, and D. Rueckert. Self-supervised learning for few-shot medical image segmentation. *IEEE Transactions on Medical Imaging*, 41(7):1837–1848, 2022.

[22] C. M. Pham, P. L. Nguyen, T. T. Nguyen, V. M. H. Phan, and B. P. Nguyen. Unleashing sam for few-shot medical image segmentation with dual-encoder and automated prompting. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 670–680. Springer, 2025.

[23] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. Sam 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations*, 2024.

[24] W. Shi, P. Zhang, Y. Li, and Z. Jiang. Segment anything model for few-shot medical image segmentation with domain tuning. *Complex & Intelligent Systems*, 11(1):37, 2025.

[25] Y. Shu, Y. Wang, L. Qiao, F. Li, B. Xiao, W. Li, and X. Gao. The overlooked matters: Revisiting background, prototype, and activation in few-shot medical image segmentation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3104–3113, 2025.

[26] H. Tang, X. Liu, S. Sun, X. Yan, and X. Xie. Recurrent mask refinement for few-shot medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3918–3928, 2021.

[27] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, and Y. Jin. Medical sam adapter: Adapting segment anything model for medical image segmentation. *Medical image analysis*, 102:103547, 2025.

[28] Q. Wu, Y. Zhang, and M. Elbatel. Self-prompting large vision models for few-shot medical image segmentation. In *MICCAI workshop on domain adaptation and representation transfer*, pages 156–167. Springer, 2023.

[29] Y. Wu, Z. Wang, X. Yang, H. Kang, A. He, and T. Li. Trans-sam: Transfer segment anything model to medical image segmentation with parameter-efficient fine-tuning. *Knowledge-Based Systems*, 310:112909, 2025.

[30] W. Xie, N. Willems, S. Patil, Y. Li, and M. Kumar. Sam fewshot finetuning for anatomical segmentation in medical images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3253–3261, 2024.

[31] Y. Yang, X. Fang, X. Li, Y. Han, and Z. Yu. Cdsg-sam: A cross-domain self-generating prompt few-shot brain tumor segmentation pipeline based on sam. *Biomedical Signal Processing and Control*, 100:106936, 2025.

[32] N. Yu, Z. Cai, Y. Huang, and X. Tang. Snapseg: Training-free few-shot medical image segmentation with segment anything model. In *International Workshop on Trustworthy Artificial Intelligence for Healthcare*, pages 109–122. Springer, 2024.

[33] Q. Yu, K. Dang, N. Tajbakhsh, D. Terzopoulos, and X. Ding. A location-sensitive local prototype network for few-shot medical image segmentation. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 262–266. IEEE, 2021.

[34] K. Zhang and D. Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023.

[35] R. Zhang, Z. Jiang, Z. Guo, S. Yan, J. Pan, X. Ma, H. Dong, P. Gao, and H. Li. Personalize segment anything model with one shot. In *International Conference on Learning Representations (ICLR)*, 2024.

[36] Y. Zhu, M. Li, Q. Ye, S. Wang, T. Xin, and H. Zhang. Robustemd: Domain robust matching for cross-domain few-shot medical image segmentation. *Artificial Intelligence in Medicine*, 167:103197, 2025.

[37] Y. Zhu, S. Wang, T. Xin, and H. Zhang. Few-shot medical image segmentation via a region-enhanced prototypical transformer. In *International conference on medical image computing and computer-assisted intervention*, pages 271–280. Springer, 2023.

[38] Y. Zhu, S. Wang, T. Xin, Z. Zhang, and H. Zhang. Partition-a-medical-image: Extracting multiple representative subregions for few-shot medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 73:1–12, 2024.

A Technical appendices and supplementary material

Technical appendices with additional results, figures, graphs, and proofs may be submitted with the paper submission before the full submission deadline (see above). You can upload a ZIP file for videos or code, but do not upload a separate PDF file for the appendix. There is no page limit for the technical appendices.

Note: Think of the appendix as "optional reading" for reviewers. The paper must be able to stand alone without the appendix; for example, adding critical experiments that support the main claims to an appendix is inappropriate.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

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 claims made in the abstract and introduction are consistent with the proposed methodology and experimental results. All major claims are supported by quantitative comparisons, ablation studies, and qualitative visualizations presented in Sections 4.2.

   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 limitations of the proposed method are discussed in the conclusion section, including the potential difficulty in accurately localizing extremely small lesions due to limited foreground cues.

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 contain formal theoretical results, theorems, or mathematical proofs. The proposed framework is primarily validated through empirical experiments and quantitative evaluations on multiple medical image segmentation datasets.

   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 detailed descriptions of the proposed framework, experimental settings, datasets, evaluation metrics, and implementation details. Hyperparameters and training configurations are clearly specified, and the code will be publicly released to facilitate reproducibility.

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 publicly available datasets. The anonymized implementation code, training scripts, and reproduction instructions are publicly available at `https://anonymous.4open.science/r/neurips-fsmis/`.

   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 provides detailed descriptions of the datasets, evaluation metrics, implementation details, hyperparameter settings, and experimental protocols.

   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: [Yes]

   Justification: Experimental results are reported as mean $\pm$ standard deviation across multiple independent runs, as shown in Tab. 1. The reported standard deviations reflect the performance variability caused by different random selections of support slices under the same evaluation protocol.

   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 specifies the hardware and software environment used for experiments, including GPU type and implementation details. The computational setup is described in the implementation details section4.

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 complies with the NeurIPS Code of Ethics.

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 proposed method may positively impact medical image analysis by improving segmentation performance under limited annotation settings, potentially reducing clinical annotation burden and increasing the accessibility of medical AI systems. Potential negative impacts include segmentation errors or failure cases in challenging scenarios, which may affect downstream clinical decision-making if the system is used without appropriate expert supervision.

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 proposed work focuses on medical image segmentation and does not involve high-risk generative models, scraped datasets, or systems with significant misuse potential. All experiments are conducted on publicly available medical 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: All datasets, pre-trained models, and external codebases used in this work are publicly available and properly cited in the paper. The corresponding licenses and terms of use are respected to the best of our knowledge.

    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: [N/A]

Justification: The paper does not introduce new datasets, benchmarks, or large-scale pre-trained models as primary research assets.

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 experiments, user studies, or research with human subjects. All experiments are conducted on publicly available medical image 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 crowdsourcing experiments, prospective data collection, or research with human subjects. All experiments are conducted on publicly available and previously anonymized medical image datasets.

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: Any use of AI-assisted tools was limited to language editing and writing refinement.

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.