Focus on Background: Exploring SAM's Potential in Few-shot Medical Image Segmentation with Background-centric Prompting
– *Supplementary Material* –

Yuntian Bo[1] Yazhou Zhu[1] Piotr Koniusz[2,3,✉] Haofeng Zhang[1,✉]
[1]Nanjing University of Science and Technology [2]University of New South Wales [3]Data61♥CSIRO
{yuntian.bo, zyz_nj, zhanghf}@njust.edu.cn, piotr.koniusz@unsw.edu.au

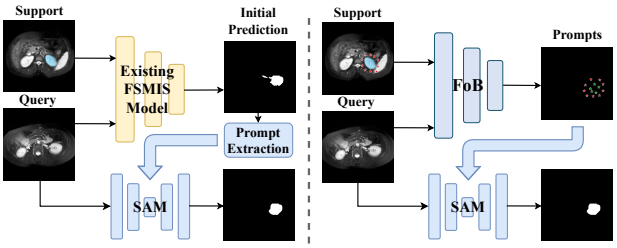

Figure 7. Comparison of the previous method, ProtoSAM [53] (left), and our method (right). ProtoSAM connects an existing FSMIS model with SAM by extracting prompts from the coarse segmentation output of the FSMIS model, which often fails to provide accurate background prompts due to inherent flaws. In contrast, our method predicts both precise background and foreground prompts, improving SAM's use in medical image segmentation.

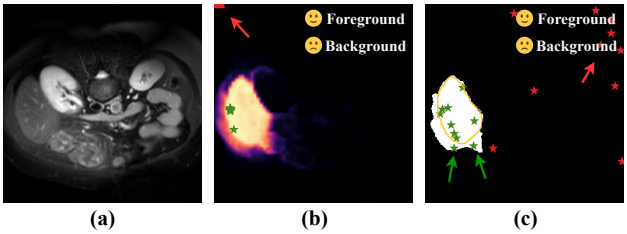

Figure 8. Previous prompt extraction methods based on existing FSMIS model outputs. (a) Original image. (b) Prediction confidence-based method, which selects high-confidence foreground and background points as prompts. (c) Coarse binary mask-based method, which randomly samples prompts within the predicted foreground and background regions of the mask. The yellow line indicates the ground truth. The arrows in (b) and (c) indicate incorrect prompt locations.

A. Discussion: Why FoB Instead of Coarse Mask-based Prompting

Our method introduces a dedicated prompt generator specifically designed for SAM-based automatic segmentation. Figure 7 shows this design fundamentally differs from previous approaches such as ProtoSAM [53], which simply combines an off-the-shelf FSMIS model (SSL-ALPNet [59]) with SAM and extracts prompts directly from the coarse predictions of the FSMIS model. Extracting prompts in this manner follows two strategies (see Figure 8) which we analyze below and explain their suboptimality:

i. **Extracting prompts based on prediction confidence [53].** Figure 8(b) shows that boundaries are hard to be accurately delineated. High-confidence background prompts selected by this method are not useful as they tend to remain far away from object boundaries. This contradicts our objective, as ideal background prompts should be placed adjacent to the outer boundary, guiding SAM to more effectively discriminate foreground from background to suppress over-segmentation.

ii. **Extracting prompts directly from the coarse binary mask.** Figure 8(c) shows due to the inherent limitations of pseudo-label supervision and the generalization capacity of FSMIS, the resulting coarse predictions are inaccurate. This purely geometric sampling strategy lacks semantic/shape awareness of the object. Thus, background prompts may fail to reach the boundaries, and foreground prompts may mistakenly be placed in background regions, or vice versa.

In contrast, FoB is directly supervised to predict boundary-adjacent background prompts to achieve accurate guidance. FoB models relations among prompts by transformer, producing background prompts with more coherent placement and shape that conforms well to the true object boundaries.

B. Additional Ablation Studies

B.1. Design Choices

B.1.1. Transformer *vs*. Mamba for Background-centric Context Modeling.

The recently proposed architecture, Mamba [56], is a competitive alternative to Transformers. In BCM, we treat pixel

✉ Corresponding authors.

Model	mDice	Param. (M)	Latency (ms/img)
FoB-Transformer	86.21	49.31	57.49
FoB-Mamba	85.31	49.90	33.95

Table 5. Comparison of Transformer- and Mamba-based implementations of the BCM module. "mDice" denotes the mean Dice score (%), and "Param." indicates the number of parameters.

r	Liver	RK	LK	Spleen	Mean
19	82.12	83.41	82.28	84.01	82.96
15	**86.51**	86.51	**87.29**	84.54	**86.21**
11	86.37	**87.33**	86.12	**84.93**	86.19
7	79.21	82.66	84.76	80.47	81.78

Table 6. Ablation study (Dice score % reported) on the impact of the dilation kernel size r. Smaller r moves background prompts closer to the foreground.

features as tokens and apply a Transformer for context modeling based on self-attention. Mamba models sequence dependencies through a learnable state-space transition. To investigate Mamba's contextual reasoning, we re-implement FoB by replacing the multi-head self-attention in BCM with Mamba. The resulting performance-efficiency trade-off is summarized in Table 5. We observe that both implementations achieve comparable segmentation performance and exhibit similar parameter scales, indicating their equivalent potential for clinical deployment. However, the Mamba-based BCM demonstrates significantly lower latency due to its linear-time inference complexity. This advantage enables faster segmentation feedback on resource-constrained medical devices.

B.1.2. Discussion: Are Medical SAMs Suitable for This Task?

As shown in Table 1 of the main text, SAM-Med2D [54] underperforms in comparison to the vanilla SAM on abdominal datasets, while significantly outperforming it on Skin-DS. Table 7 presents the results of using FoB to prompt another popular SOTA medical SAM, MedSAM [58], which was not evaluated in the main text due to its support for box prompts only. The results show suboptimal performance which we attribute to the architectural choice: both SAM-Med2D and MedSAM adopt the ViT-B (base) backbone, which is less expressive than the ViT-H (huge) backbone used in the original SAM [57]. Notably, most SOTA medical SAM variants, including SAM-Med2D and MedSAM, rely on ViT-B, due to its reduced data requirements [55], which is a property well aligned with the limited availability of annotated medical data.

The superior performance of SAM-Med2D on some datasets may stem from data overlap between training and testing. Given the scarcity of public medical datasets, it is possible that some test images, especially from Skin-DS, were seen during the training of SA-Med2D-20M [60], which was used to train SAM-Med2D.

In contrast, the few-shot segmentation setting strictly prohibits access to target classes during training. Our intention is to leverage the category-agnostic segmentation ability of the SAM trained on natural images and extend its generalization to the medical domain, whereas using medical SAMs risks violating the fundamental assumptions of few-shot setting. Thus, while we report results from SAM-Med2D and MedSAM for completeness and to assess prompt generalization across model variants, we advocate using the vanilla SAM to comply with the FSMIS protocol.

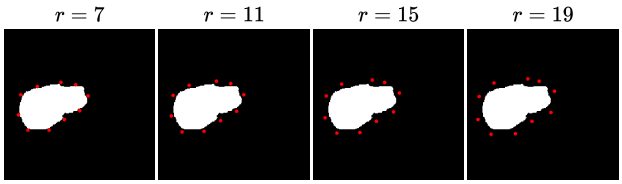

Figure 9. Visualization of support background prompts generated with different r in BPPC. Best viewed under large zoom.

B.2. Hyperparameter Settings

B.2.1. Optimal Proximity of Background Prompts.

We design FoB to predict background prompts that are located close to the foreground region, thereby constraining the over-segmentation errors that extend beyond object boundaries. This section investigates an important question: "Is closer always better?"

In our design, this distance is controlled by a dilation kernel size r. A smaller r generates background prompts closer to the foreground in BPPC. Figure 9 shows support background prompts with different r. As shown in Table 6, reducing the distance between the background prompts and the foreground gradually improves the segmentation accuracy of SAM, indicating that over-segmentation is effectively suppressed. However, when this distance becomes excessively small, the performance drops. We attribute this to the model excessively prioritizing proximity, which increases the risk of background prompts falling inside the true foreground, introducing conflicting signals and deteriorating the segmentation quality.

B.2.2. Additional Hyperparameter Analysis.

We conduct extensive ablation studies on key hyperparameters in our model, including the standard deviation σ for heatmap generation, the number of deformable iterations κ and deformable receptive field size k in SPR, the foreground sampling threshold $\mathcal{T}$, and the temperature parameter τ in the RAC loss. Among these, σ has the most significant im-

Setting	Abd-CT					Abd-MRI					Skin-DS			
	Liv	RK	LK	Spl	Avg.	Liv	RK	LK	Spl	Avg.	Mel	Nev	SK	Avg.
Setting I	69.51	65.79	73.01	57.93	66.56	71.54	78.57	80.46	68.81	74.85	71.59	75.34	68.75	71.89
Setting II	68.61	64.30	67.67	58.68	64.82	67.99	75.57	74.47	62.86	70.22	72.91	71.25	68.49	70.88

Table 7. Segmentation results of MedSAM when prompted by FoB. Bounding box prompts are obtained by computing the minimum enclosing rectangle of the background prompts generated by FoB. These results further validate that medical SAMs are in fact unsuitable for FSMIS tasks, due to inherent architectural limitations and potential violations of the few-shot learning protocol.

σ	Liver	RK	LK	Spleen	Mean
8	80.02	84.94	85.78	80.21	82.74
4	**86.51**	**86.51**	**87.29**	**84.54**	**86.21**
2	79.76	75.56	75.97	79.91	77.80
1	76.33	61.93	64.46	62.57	66.32

Table 8. Ablation study (Dice score % used) on the impact of the standard deviation σ for generating heatmaps.

k	Liver	RK	LK	Spleen	Mean
16	84.96	76.80	88.33	83.89	83.50
8	86.51	**86.51**	87.29	84.54	**86.21**
4	**87.50**	84.64	85.58	**86.19**	85.98
2	81.05	85.40	**88.63**	83.62	84.68

Table 9. Ablation study (in Dice score %) on the impact of the deformable receptive field size k in SPR.

pact on model performance, as shown in Table 8. It primarily affects both the supervision signal and the prompt prototype generation. A large σ weakens the supervision strength and produces prototypes that fail to accurately represent the prompts. Conversely, a small σ makes the model harder to optimize and results in prototypes that are highly sensitive to noise. Tables 9 and 10 present the hyperparameter analysis of the SPR module. The results show minor performance variance w.r.t. κ and k. The effects of $\mathcal{T}$ and τ are summarized in Tables 11 and 12, respectively. Notably, a smaller τ corresponds to stronger contrastive constraints. As τ decreases, the segmentation performance of SAM steadily improves, suggesting that stronger feature discrimination can effectively prevent background prompts from being misclassified as foreground. Overall, the optimal hyperparameter configuration in our experiments is $\sigma = 4$, $\kappa = 3$, $k = 8$, $\mathcal{T} = 0.90$, and $\tau = 0.10$.

B.3. Robustness Analysis

B.3.1. Does BCM rely on accurate foreground prediction?

In BCM, we first predict a foreground mask to help subsequent modeling differentiate foreground and background regions. While accurate foreground prediction is beneficial, it is not strictly required, *i.e.*, coarse predictions are already sufficient to localize the foreground region because BCM does not rely solely on foreground signals. It explicitly models foreground–background relationships while also leveraging additional contextual cues, *e.g.*, anatomical structures encoded in the query spatial layout. Thus, the performance of BCM is not determined by foreground prediction alone.

Furthermore, the predicted foreground remains reliable even in scenarios where foreground and background are visually ambiguous. This is because the prediction process is supervised by $\mathcal{L}_{fore}$, which encourages semantically discriminative feature learning and promotes precise boundary delineation. In addition, $\mathcal{L}_{rac}$ explicitly focuses on hard boundary regions, further enhancing feature discriminabil-

κ	Liver	RK	LK	Spleen	Mean
7	82.62	83.99	83.94	77.67	82.06
5	84.43	**86.68**	86.63	83.75	85.37
3	**86.51**	86.51	87.29	84.54	**86.21**
1	84.77	84.17	**87.42**	**84.68**	85.26

Table 10. Ablation study (Dice score % used) on the impact of the number of deformable iterations κ in SPR.

$\mathcal{T}$	Liver	RK	LK	Spleen	Mean
0.95	**88.75**	86.03	83.99	**85.27**	86.01
0.90	86.51	**86.51**	**87.29**	84.54	**86.21**
0.85	87.94	83.77	85.62	82.47	84.95
0.80	87.38	84.55	84.01	81.79	84.43

Table 11. Ablation study (Dice score % used) on the impact of the foreground prompt sampling threshold $\mathcal{T}$.

τ	Liver	RK	LK	Spleen	Mean
0.7	81.74	86.31	85.72	80.09	83.47
0.5	84.96	**86.79**	84.94	82.01	84.68
0.3	**87.24**	86.37	86.97	83.17	85.94
0.1	86.51	86.51	**87.29**	**84.54**	**86.21**

Table 12. Ablation study (Dice score % used) on the impact of the temperature τ in the RAC loss.

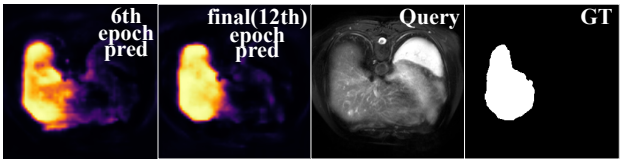

Figure 10. Evolution of the foreground probability map over training epochs. As training proceeds, the foreground prediction becomes highly accurate even under ambiguous boundaries.

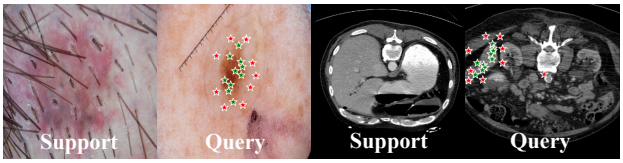

Figure 11. Visualization of matching robustness of the BCM module under significant background variations.

Graph	Liver	RK	LK	Spleen	Mean
No $\mathbf{A}$	81.32	87.81	82.43	78.55	82.53
$\mathbf{A}^{ring}$ only	84.99	**88.26**	84.09	78.95	84.07
$\mathbf{A}^{ada}$ only	83.87	87.00	**84.88**	77.39	83.29
$\mathbf{A}^{ada} + \mathbf{A}^{ring}$	**85.61**	88.18	84.76	**79.31**	**84.46**

Table 13. Ablation study (Dice score % used) on different graph construction strategies.

ity for foreground–background separation. As training progresses, the foreground prediction becomes increasingly accurate, as illustrated in Figure 10.

B.3.2. Is matching robust under significant background variation?

In medical images, the background typically exhibits substantial variation across samples, which can hinder reliable matching for localizing background prompts. However, BCM mitigates this issue by leveraging global contextual information. In particular, it utilizes the foreground region, which is easier to predict than background prompts, as well as additional anatomical context from the query instance to perform instance-adaptive reasoning. This design makes the matching process robust to background variations. When support prompt prototypes are less reliable due to large background discrepancies, BCM compensates by leveraging query contextual cues to refine the matching. We further demonstrate this robustness using the FoB model without SPR in Figure 11. BCM accurately localizes the background prompts through matching, even under significant variations.

B.4. Ablations on Structural Graph in SPR

In SPR, we construct graphs to encode the structural relationships among support background prompt prototypes,

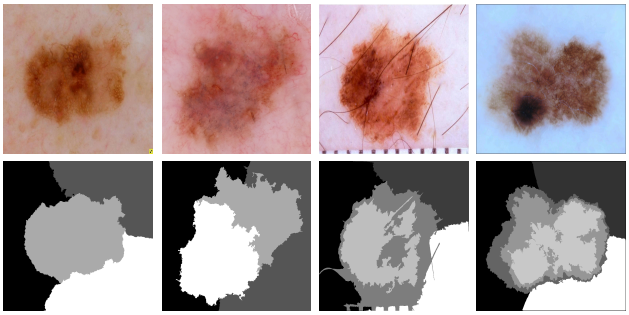

Figure 12. Illustrative examples of superpixel-based pseudo labels on Skin-DS.

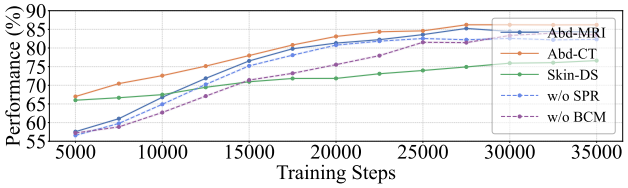

Figure 13. Performance curves of FoB & its variants on 3 datasets.

which are used to regularize the distribution of query prompt prototypes in the feature space. Specifically, we construct both an adaptive graph $\mathbf{A}^{ada}$ and a ring-prior graph $\mathbf{A}^{ring}$ for subsequent modeling.

We ablate different graph construction strategies in Table 13. The quantitative results show that $\mathbf{A}^{ring}$ contributes the most significant performance gain, as it imposes a strong ring-shaped topological prior on the query prototypes. Combining both graphs yields the best performance, as it captures both category-specific structural relationships and a general ring-like prior.

B.5. Performance Curve

We report the performance curves of FoB on three datasets, together with ablation result curves without BCM and SPR, as shown in Figure 13. We observe that on complex multi-object datasets (*e.g.*, abdominal datasets), the initial performance is relatively low, as SAM struggles to accurately separate multiple regions under challenging scenes. As the background prompts become progressively more accurate, the performance improves substantially. The performance of our model increases steadily with the learning of background prompts. Furthermore, BCM accelerates convergence, while SPR consistently improves the upper performance bound, demonstrating the effectiveness of our proposed modules.

C. Details of Superpixel-based Pseudo-labeling for Skin-DS

To the best of our knowledge, we are the first to adopt the Skin-DS dataset for FSMIS. Following the conventional

pseudo-label training paradigm [59], we generate pseudo labels for Skin-DS to enable training FoB without requiring ground-truth annotations of skin disease in Setting I. This lets the model learn robust and generalizable patch-level features, mitigating the risk of overfitting to specific semantics and thus enhancing its ability to generalize to unseen categories during inference. For simplicity and computational efficiency, we adopt the SLIC [52] algorithm for pre-processing Skin-DS. SLIC performs k-means clustering in a joint color–spatial domain, yielding compact and edge-aware superpixel regions. In our implementation, the number of desired superpixels is set to 5, and the compactness parameter is set to 15. We show several processed examples in Figure 12. Moreover, Figure 14 provides qualitative segmentation results on Skin-DS.

D. Additional Visualizations

D.1. Visual Analysis on Structure-guided Prompt Refinement (Detailed)

In Table 3 of the main text, we quantitatively demonstrate the effectiveness of the SPR module. We further provide visualization results to highlight the significant improvements that SPR brings in generating background prompts that better align with the inherent structure. As shown in Figure 15, for several examples, FoB with SPR (w/ SPR) effectively learns to predict smooth, ring-like prompt distributions that closely follow the spatial shape of the target category (as indicated by support prompts), wrapping around the foreground to offer strong constraints to prevent SAM's over-segmentation. In contrast, the predictions of the model trained without SPR are inaccurate, resulting in either outlier prompts located far from the foreground (*e.g.*, top row, second column) or overly compact prompt clusters (*e.g.*, bottom row, second column).

D.2. Generated Background Prompts

A comprehensive visualization of the generated background prompts by FoB across different imaging modalities is presented in Figure 16. We observe that FoB demonstrates remarkable accuracy in localizing background points adjacent to target boundaries. These points are distributed in a morphologically consistent manner around category boundaries, offering strong guidance to constrain the over-segmentation of SAM. Moreover, FoB also yields highly accurate foreground prompts, despite relying solely on basic prototype matching without introducing any additional architectural components.

D.3. Visualization of Segmentation Results

We present the qualitative results of our method in Figures 17 and 14. Compared to conventional approaches and the prior SAM-based method, ProtoSAM, our approach produces more complete foreground segmentation with sharper and more decisive boundaries, benefiting from SAM's strong capability in image segmentation. It also significantly suppresses over-segmentation, a severe issue not only in ProtoSAM but also in conventional methods based on prototypical matching. Our results demonstrate the potential of background-centric few-shot SAM prompting in clinical applications, which achieves strong performance while requiring minimal annotated data.

Algorithm 1 Focus on Background Prompt Generator (1-shot).

Require: Support set $\mathcal{S} = \{(\mathbf{I}^s, \mathbf{M}^s)\}$, query image $\mathbf{I}^q$, numbers of prompts N_p, N_f.
Ensure: Background prompts $\mathcal{P}'_b$, foreground prompts $\mathcal{P}_f$

Feature extraction: Extract support and query features $\mathbf{F}^s$ and $\mathbf{F}^q$ using a shared-weight encoder $f(\cdot)$.
Background Prompt Prototypes Construction (BPPC):
  Sample support background prompt set $\mathcal{P}$ (Eq. (1)).
  Generate Gaussian heatmaps set $\mathbf{G} = [\mathbf{G}^1, \ldots, \mathbf{G}^{N_p}]$ centered at each $\boldsymbol{\mu}^i \in \mathcal{P}$ (Eq. (2)).
  Create background prompt prototype set $\mathbf{P}$ (Eq. (3)).
Background-centric Context Modeling (BCM):
  Get foreground suppressed query image feature $\mathbf{F}_{sup}$ (Eq. (4)).
  Generate coarse background prompt proposal $\mathbf{\Phi}$ with $\mathbf{P}$ and $\mathbf{F}_{sup}$ (Eq. (5)).
  Obtain the contextual modulated feature $\mathbf{F}_m$ using the masked transformer (Eq. (6) & (7)):
  Heatmap prediction: $\hat{\mathbf{H}} \leftarrow \text{Head}(\mathbf{F}_m)$.
  Obtain coarse prompts $\mathcal{P}_b$ by selecting the maximum response in each heatmap: $\mathcal{P}_b \leftarrow \{\arg\max \hat{\mathbf{H}}^i\}_{i=1}^{N_p}$.
Structure-guided Prompt Refinement (SPR):
  Estimate adaptive graph $\mathbf{A}^{ada}$ with support features $\mathbf{P}$ (Eq. (8)).
  Compute ring prior graph $\mathbf{A}^{ring}$ (Eq. (9)).
  Compute $\mathbf{A}$ as a weighted sum of $\mathbf{A}^{ada}$ and $\mathbf{A}^{ring}$ (Eq. (10)).
  Transfer support structure to query to get $\mathbf{Q}'$ (Eq. (11)).
  Iteratively update prompt coordinates:
  for $i \leftarrow 1$ **to** N_p **do**
    Initialize $\mathbf{f} \leftarrow \mathbf{q}_b^{i\prime} \in \mathbf{Q}'$
    for $t \leftarrow 1$ **to** κ **do**
      Predict offset set $\Delta\boldsymbol{\mu}$ (Eq. (12)).
      Compute weights $\mathbf{w}$ using $\mathbf{q}_b^i$ (Eq. (13)).
      Refine location $\boldsymbol{\mu}_b^i \in \mathcal{P}'_b$ and feature $\mathbf{f}$ using $\mathbf{w}$ and $\Delta\boldsymbol{\mu}$ (Eq. (14) & (15)).
    end for
  end for
return $\mathcal{P}'_b, \mathcal{P}_f$

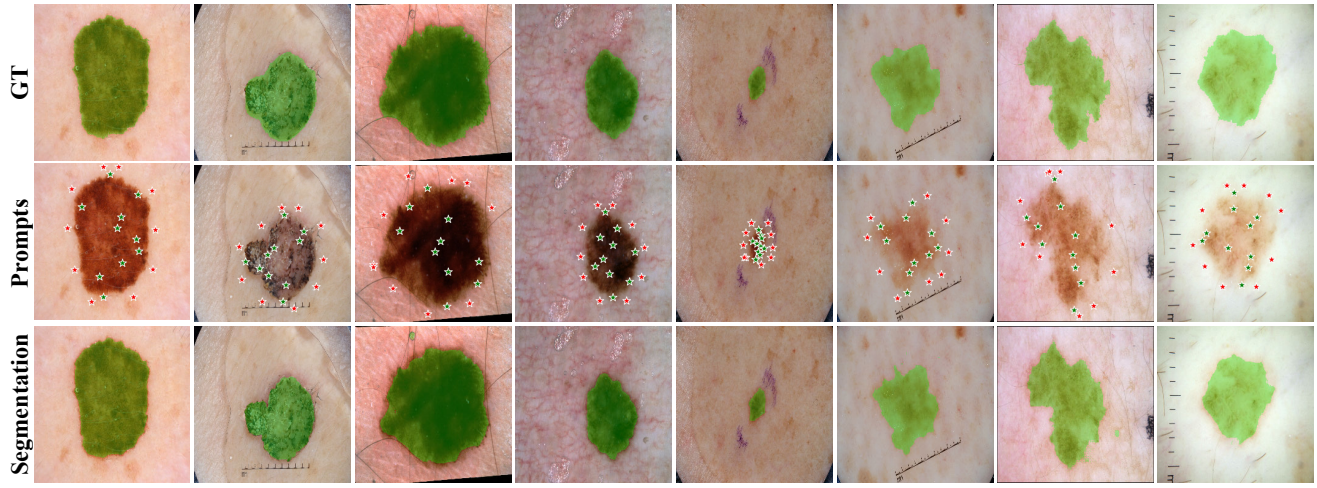

Figure 14. Qualitative segmentation results of our method on Skin-DS.

E. Limitations and Future Work

Although our FoB leverages SAM to achieve accurate segmentation for common medical targets, the current design does not yet support highly irregular and thin structures, such as vessels. This limitation arises because such cases require a larger number of background prompts to avoid erroneous segmentation, whereas our design adopts a fixed number N_p of background prompts. Moreover, such cases are also not well aligned with the ring-shape prior and may benefit from other advanced priors. Future work may explore supporting a dynamic number of background prompts to better adapt to the geometry and scale of target structures, as well as more adaptive strategies to correct the predicted prompts. We hope that our sparse point-matching-based paradigm can foster more SAM-based FSMIS methods and facilitate their practical deployment.

F. Algorithm

Algorithm 1 illustrates the proposed FoB model which comprises three key stages: 1) background prompt prototype generation from the support set via BPPC; 2) contextual modeling for enhanced background prompt localization via BCM; and 3) structure-guided refinement for calibrating erroneous query prompts via SPR.

References

[52] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2274–2282, 2012.

[53] Lev Ayzenberg, Raja Giryes, and Hayit Greenspan. Protosam: One shot medical image segmentation with foundational models. *arXiv preprint arXiv:2407.07042*, 2024.

[54] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiangand Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d. *arXiv preprint arXiv:2308.16184*, 2023.

[55] 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 *ICLR*, 2021.

[56] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[57] 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 *ICCV*, pages 4015–4026, 2023.

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

[59] Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. In *ECCV*, pages 762–780. Springer, 2020.

[60] Jin Ye, Junlong Cheng, Jianpin Chen, Zhongying Deng, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, Hui Sun, Min Zhu, Shaoting Zhang, Junjun He, and Yu Qiao. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks, 2023.

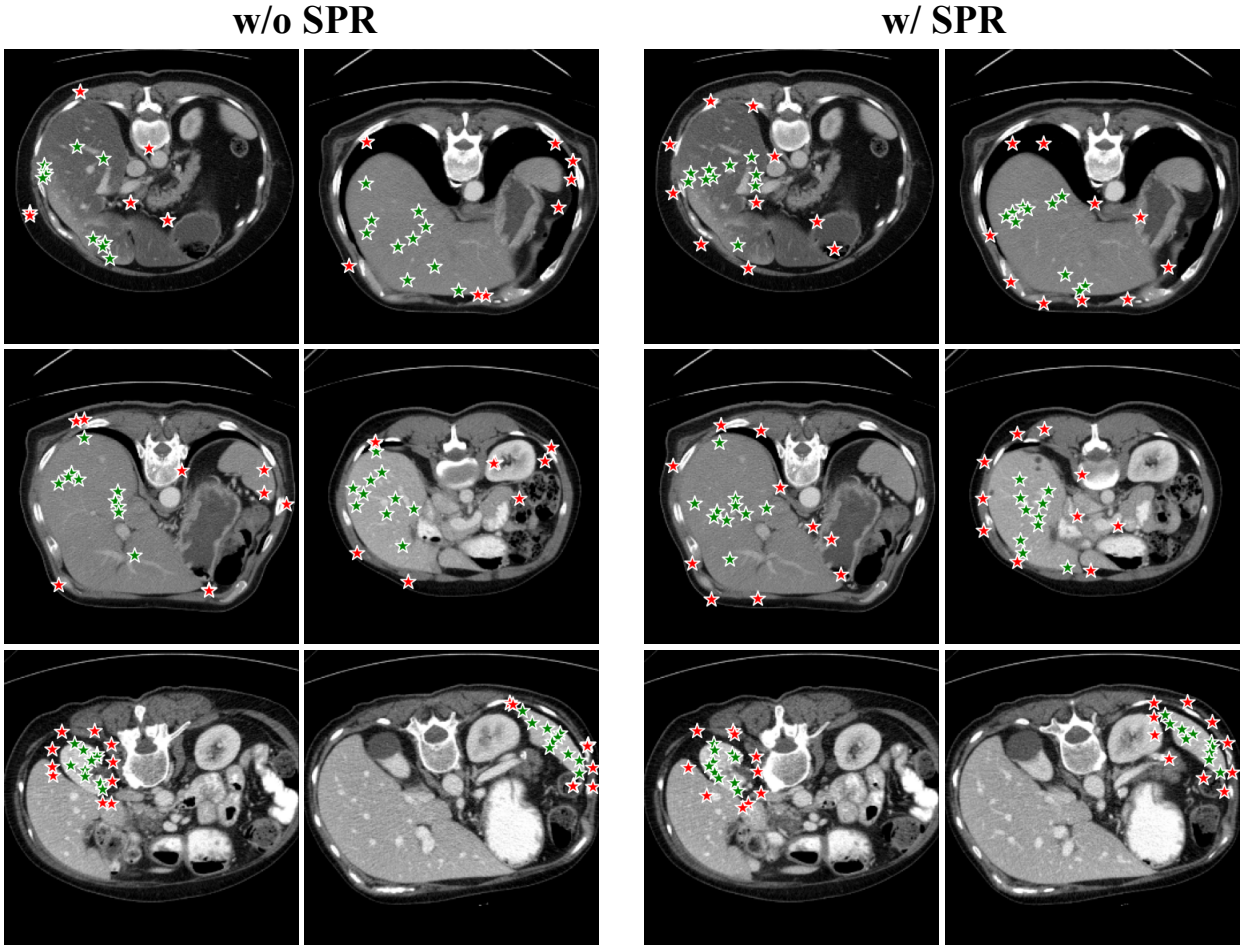

Figure 15. Qualitative effect of SPR on Abd-CT. Our proposed FoB with structure-aware refinement (w/ SPR) significantly outperforms the counterpart that solely uses BCM-predicted prompt sets (w/o SPR).

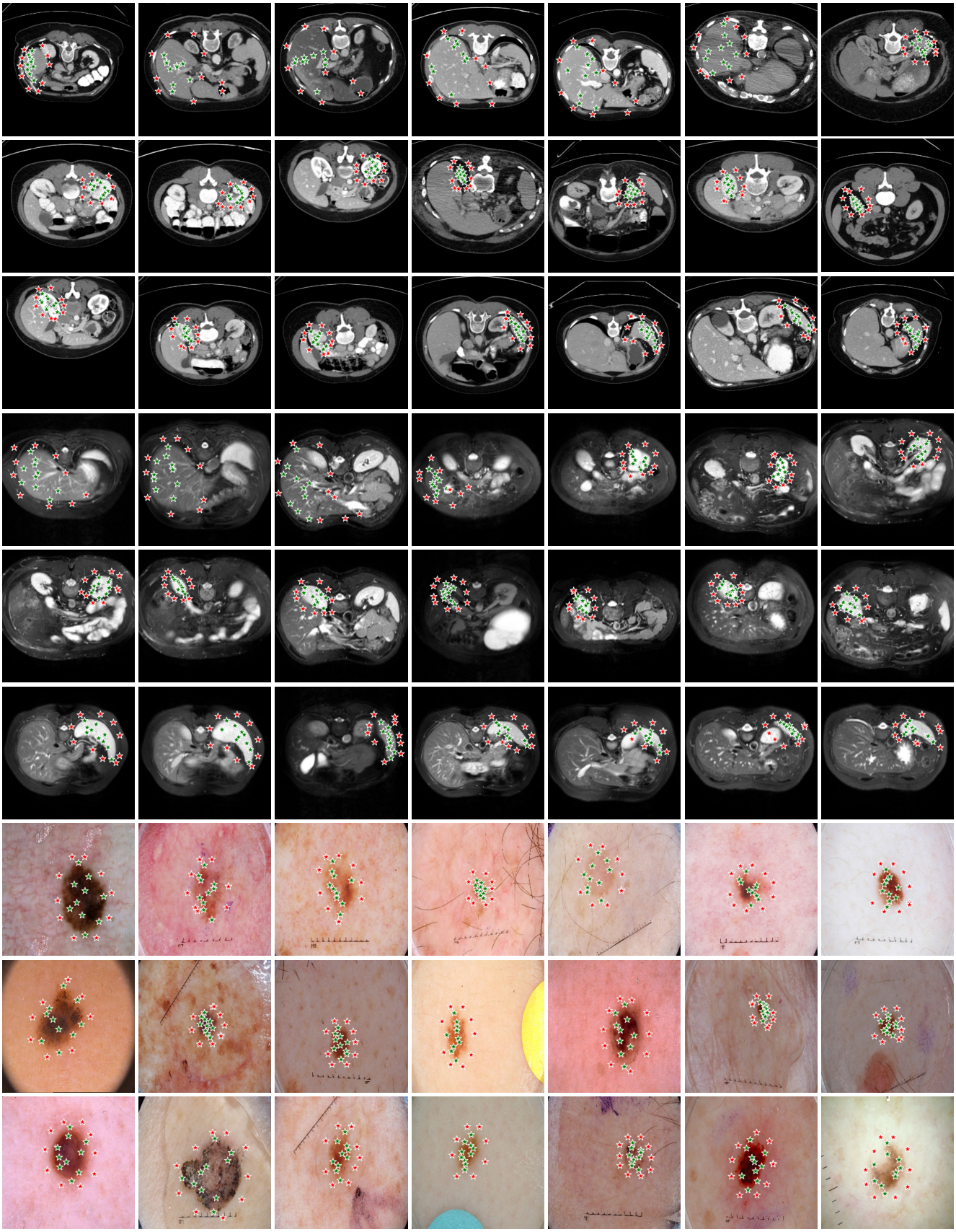

Figure 16. Visualization of prompts generated by the proposed FoB. Rows 1–3 correspond to Abd-CT, rows 4–6 to Abd-MRI, and rows 7–9 to Skin-DS. FoB produces highly reliable background prompts that play a crucial role in constraining SAM's over-segmentation.

Figure 17. Qualitative comparison of segmentation results on Abd-MRI (upper) and Abd-CT (lower).