APEX-SAM: Anatomy-aware Prompting with Expert Retrieval for Training-free Medical Image Segmentation

Zhihao Mao*, Bangpu Chen*, Qi Lei, Jiaqi Tan, and Kun Sun†

School of Computer Science, China University of Geosciences (Wuhan), China
htxmzh@gmail.com, chenbangpu@cug.edu.cn, kevinlei@cug.edu.cn
jiaqitan639@gmail.com, sunkun@cug.edu.cn

Abstract. Training-free cross-domain few-shot medical image segmentation aims to segment unseen anatomies without parameter updates, addressing the high cost of dense annotation and domain-specific fine-tuning in clinical practice. Existing support-driven prompting methods face three limitations: support exemplars are randomly selected without quality assurance, geometric alignment is poorly modeled, and multi-modal prompt capabilities remain underexploited. We present APEX-SAM, a retrieval-augmented framework with three innovations. Quality-aware expert retrieval (QAR) builds a dual-stream DINO/SigLIP expert bank with diversity-aware selection to ensure support-query compatibility. Anatomy-aware prompt mining (APM) performs style-aligned geometric matching and anatomy-guided point sampling from morphological priors. Hybrid multi-modal fusion (HMF) fuses three SAM branches (point, text, and box prompts) via training-free feature-consensus weighting. Experiments on three cross-domain benchmarks confirm state-of-the-art performance among training-free methods, with ablations validating each component's contribution.

Keywords: Cross-domain segmentation · Few-shot learning · Retrieval-augmented prompting · SAM · Training-free

Project page: https://trump0412.github.io/APEX-SAM/

1 Introduction

Medical image segmentation is central to diagnosis, intervention planning, and longitudinal assessment in routine clinical workflows [2]. Accurate organ and lesion delineation enables quantitative biomarker extraction, treatment response monitoring, and surgical planning. However, its main practical bottleneck is expensive dense annotation, especially for rare diseases and newly adopted protocols [22]. When combined with domain gaps across modalities (e.g., MRI vs.

* Equal contribution. † Corresponding author.

CT), scanners, and institutions, this makes cross-domain few-shot segmentation both clinically urgent and technically difficult.

Recent methods fall into two paradigms: those requiring training and those that are training-free. Training-based methods achieve strong accuracy via prototype matching, meta-learning, and domain adaptation [15,17,10,5,16,6]. These methods typically learn transferable representations from large labeled source datasets and adapt them to new targets via iterative optimization. However, they rely heavily on source domain supervision and repeated fine-tuning for each new target domain. Moreover, robustness often degrades under large clinical distribution shifts, and deployment requires substantial computational resources and domain expertise.

Within the training-free branch, support-driven prompting has become a promising direction for fast adaptation. Representative methods such as MAUP, SynPo, and OSAM build prompts from support-query correspondence without parameter updates, which is particularly attractive for high-frequency clinical environments where rapid deployment and minimal infrastructure are essential [25,11,18]. These methods leverage pre-trained foundation models (e.g., SAM [8]) and construct prompts from a small set of support exemplars at test time, enabling immediate inference on unseen anatomies without retraining.

Current support-driven pipelines still face three unresolved bottlenecks: inadequate support quality assurance, poor support-query alignment modeling, and underused multi-modal prompt signals [25,4,20,23]. These issues directly limit stability when support and query are weakly aligned in real cross-domain settings. First, support samples are often selected randomly without explicit quality control or diversity constraints. Second, direct mask transfer ignores geometric misalignment between domains, and existing point-prompting strategies do not leverage anatomical structure priors. Third, although modern foundation models such as SAM3 [4] support rich multi-modal prompts, current methods rely on single-modal prompts and lack principled fusion strategies.

Medical images offer rich exploitable priors that are often overlooked by existing methods. For instance, 3D scans exhibit strong inter-slice morphological continuity due to organ shape consistency, and cross-patient data share semantic consistency in anatomical structures despite appearance variations. Together with recent advances in promptable segmentation and retrieval-augmented generation, this opens a practical path for training-free inference without retraining [4]. The key insight is to unify quality-aware retrieval, geometry-guided adaptation, and multi-modal prompt orchestration in a single inference pipeline that exploits both data priors and model capabilities.

To this end, we propose **APEX-SAM**, a training-free retrieval-augmented framework for cross-domain few-shot medical image segmentation. Our contributions are threefold:

- **QAR**: a dual-stream DINO [13]/SigLIP [21] expert bank with diversity-aware quality scoring and two-level retrieval.
- **APM**: wavelet style normalisation, semantic gating, orientation-aware geometric alignment, and anatomy-guided prompt generation.

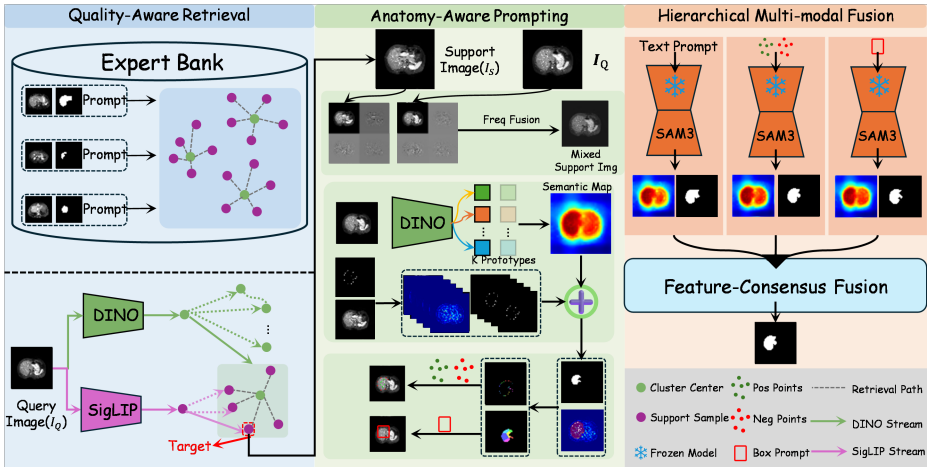

Fig. 1. Overview of APEX-SAM. QAR (left) builds a hierarchical expert bank with DINO/SigLIP dual-stream indexing. APM (middle) performs style alignment, semantic gating, orientation-aware geometric alignment, and anatomy-aware point sampling. HMF (right) fuses three SAM3 prompt branches via feature-consensus weighting.

- **HMF**: point, text, and box prompts as three independent SAM branches, fused by reliability-weighted feature consensus.

APEX-SAM outperforms the strongest training-free baseline by an average of 21.1 pp mean Dice across three cross-domain benchmarks. Code and project materials are available on the project page.

2 Method

Problem setting. We study training-free cross-domain few-shot medical image segmentation with a frozen foundation model [8,4]. Given a query image $\mathbf{I}_q$ and a textual target description t_q, the goal is to predict a binary segmentation mask without any gradient update. Formally, we require a mapping

$$\hat{\mathbf{M}}_q = \mathcal{F}(\mathbf{I}_q, t_q, \mathcal{D}), \tag{1}$$

where $\mathcal{D}$ is an external expert bank and $\mathcal{F}$ is entirely executed at inference time. Each entry $e_i = (I_i, M_i, t_i, z_i^D, z_i^S, q_i)$ stores the support image, mask, text label, a DINO structural key z_i^D, a SigLIP multi-modal key z_i^S, and a quality score q_i. Our method instantiates three principles as **QAR**, **APM**, and **HMF**: retrieval before prompting, geometry-guided alignment before transfer, and multi-branch prompt fusion. Figure 1 provides an end-to-end overview.

2.1 Quality-Aware Expert Retrieval (QAR)

Expert bank construction. Naively pooling all annotated images creates a redundant bank dominated by easy, similar cases; we instead cluster supports

by DINO [13] structural keys $\{z_i^D\}$ and greedily select entries per cluster using a quality-coverage-diversity score. The SigLIP [21] retrieval key for each candidate is formed as

$$z_i^S = \mathrm{Norm}(\alpha\, v_i + (1-\alpha)\, u_i)\,, \tag{2}$$

where $v_i = \mathrm{SigLIP}_{\mathrm{img}}(I_i)$ captures visual appearance and $u_i = \mathrm{SigLIP}_{\mathrm{txt}}(t_i)$ encodes the anatomy label, ensuring cross-modal alignment in the retrieval index. Inside each cluster k with selected subset $\mathcal{S}_k$ and centroid μ_k^S, a candidate i is scored by

$$\pi(i \mid \mathcal{S}_k) = q_i + \cos\big(z_i^S, \mu_k^S\big) + \min_{j\in\mathcal{S}_k}\big(1 - \cos\big(z_i^S, z_j^S\big)\big)\,, \tag{3}$$

where the three equally-weighted terms reward annotation quality, cluster representativeness, and diversity relative to already-selected supports, respectively. This ensures the bank is compact yet covers the intra-class appearance variation arising from different imaging protocols and scanner vendors.

Two-level retrieval at inference.

Given a query, QAR routes it to the top-L clusters by DINO structural similarity r_k. QAR then re-ranks the candidate supports by multi-modal similarity

$$s_{qi} = \cos\big(z_q^S,\, z_i^S\big)\,, \quad \mathcal{S}_q = \mathrm{TopM}(\{s_{qi}\})\,, \tag{4}$$

which fuses structural and semantic cues into a single retrieval score; the two-level index keeps cost sublinear in bank size while multi-modal keys prevent structural matches from overriding semantic misalignment. For append-only bank growth—built offline from source-domain ground-truth masks—a new entry is inserted only when its maximum cosine similarity to existing cluster members falls below β, with the cluster centre updated online in $\mathcal{O}(1)$ time; at inference the append path uses the predicted mask and SAM confidence as quality proxy, requiring no target-domain annotation.

2.2 Anatomy-Aware Prompt Mining (APM)

Cross-domain style normalisation. Cross-modality intensity mismatches corrupt feature comparisons, so we apply a one-level Haar DWT and swap the support's low-frequency sub-band with the query's:

$$\tilde{I}_s = \mathrm{IDWT}(LL_q,\, LH_s,\, HL_s,\, HH_s)\,. \tag{5}$$

This preserves organ-boundary detail in the high-frequency sub-bands while removing global intensity shift.

Semantic gating. We cluster support foreground features into K_r=8 DINO prototypes $\{c_k\}$ and build a binary gate

$$G(x) = \mathbb{1}\big[\max_k \cos\big(F_q(x),\, c_k\big) > \eta\big], \tag{6}$$

masking out query pixels with low semantic affinity to the support, suppressing false-positive activations before geometric alignment.

Orientation-aware geometric alignment.

Even after style normalisation, organ scale, orientation, and boundary curvature may still differ across patients. We find the best-fitting transform T^* by minimising a directional chamfer cost over the support boundary:

$$T^* = \arg\min_{T\in\mathcal{T}} \frac{1}{|B_s|} \sum_{(x,\theta)\in B_s} \mathrm{DT}_{k(\theta+r)}(T(x)), \tag{7}$$

where $\mathrm{DT}_{k(\cdot)}$ is the oriented distance transform binned by edge direction. Unlike isotropic Chamfer matching, the directional penalty yields finer shape correspondence on thin structures such as myocardium and kidney cortex. The aligned pre-mask is then $M_{\mathrm{pre}} = T^*(M_s) \cap G$, confining warped support shape to semantically gated query regions.

Prompt generation. From M_{pre} we extract a bounding box b, positive points at Voronoi cell centroids [1] of the foreground for spatial diversity, and negative points at the lowest-affinity pixel in each boundary sector—placing negative cues in the anatomically adjacent tissue most prone to false-positive activation.

2.3 Hybrid Multi-Modal Fusion (HMF)

Three-branch inference.

We exploit the multi-prompt interface by running SAM once per modality: Branch A uses geometry-only points $\{P^+, P^-\}$; Branch B supplies the anatomy text label t_s; Branch C uses the bounding box b from M_{pre}. Each branch produces last-layer logits M^b and feature maps F^b ($b \in \{po, tp, bx\}$).

Training-free reliability-weighted fusion. Rather than treating all branches equally, we score each by its cosine agreement with the per-pixel mean feature $\bar{F}$: $q_b = |\Omega|^{-1}\sum_x \cos(F^b(x), \bar{F}(x))$, then normalised into weights w_b via softmax with temperature γ. The final prediction is

$$\hat{M}_q = \sigma\left(\sum_{b\in\{po,tp,bx\}} w_b \cdot \mathrm{logit}\big(\mathrm{clip}(M^b, \epsilon, 1-\epsilon)\big)\right), \tag{8}$$

where ϵ clips probabilities away from $\{0, 1\}$ to prevent log-odds divergence, and no learned parameters are required.

3 Experiments

3.1 Experimental Setup

We evaluate APEX-SAM on Abd-MRI, Abd-CT, and Card-MRI [7,9,26] under a cross-domain 5-shot protocol: a unified expert bank is built offline from all three source-domain splits (patient-level leave-one-out); at inference, retrieval

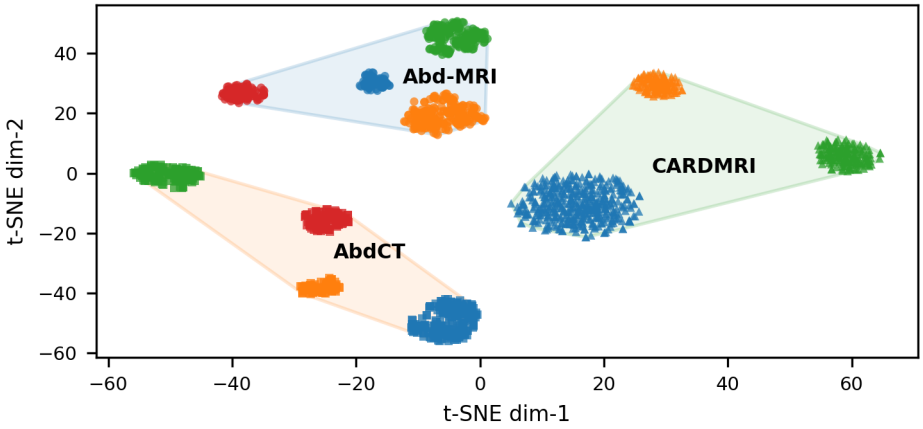

Fig. 2. Dataset visualization. Samples and DINO-feature t-SNE (one point per slice) for Abd-MRI, Abd-CT, and Card-MRI, illustrating cross-domain appearance variation.

Table 1. Quantitative comparison (Dice %) on Abd-MRI and Abd-CT. Best in **bold**, second underlined.

Method	Ref.	Abd-MRI					Abd-CT				
		Liver	LK	RK	Spleen	Mean	Liver	LK	RK	Spleen	Mean
PANet	ICCV'19	39.24	26.47	37.35	26.79	32.46	40.29	30.61	26.66	30.21	31.94
SSL-ALP	TMI'22	70.74	55.49	67.43	58.39	63.01	71.38	34.48	32.32	51.67	47.46
RPT	MICCAI'23	49.22	42.45	47.14	48.84	46.91	65.87	40.07	35.97	51.22	48.28
PATNet	ECCV'22	57.01	50.23	53.01	51.63	52.97	75.94	46.62	42.68	63.94	57.29
IFA	CVPR'24	50.22	35.99	34.00	42.21	40.61	46.62	25.13	26.56	24.85	30.79
FAMNet	AAAI'25	73.01	57.28	74.68	58.21	65.79	73.57	57.79	61.89	65.78	64.75
MAUP	MICCAI'25	78.16	58.23	72.34	59.65	67.09	78.25	59.41	71.80	60.38	67.46
APEX-SAM	Ours	**95.10**	**96.88**	**96.01**	**95.23**	**95.81**	**93.47**	**90.28**	**91.83**	**92.06**	**91.91**

is label-filtered to entries matching the queried anatomy, and no target-domain labels are used. Dice is reported as the per-class mean over 1000 episodes.

Our implementation uses DINO/SigLIP retrieval keys, SAM3 decoding, and rigid/similarity geometric alignment; q_i is mean SAM mask-confidence over foreground pixels; query/support slices are patient-disjoint.

Hyperparameters: L=3, M=5, α=0.6, η=0.94, γ=0.1, N_v=6, N_s=8.

3.2 Main Results

APEX-SAM achieves consistent gains over all baselines on three benchmarks; Tables 1 and 2 compare against PANet [17], SSL-ALP [14], RPT [24], PATNet [19], IFA [12], FAMNet [3], and MAUP [25].

Table 1 and Table 2 report per-class Dice on all three benchmarks; abdominal gains are larger because cross-modality appearance divergence (MRI vs. CT)

Table 2. Quantitative comparison (Dice %) on Card-MRI. Best in **bold**, second underlined.

Method	Ref.	LV-BP	LV-MYO	RV	Mean
PANet	ICCV'19	51.42	25.75	25.75	36.66
SSL-ALP	TMI'22	83.47	22.73	66.21	57.47
RPT	MICCAI'23	60.84	42.28	57.30	53.47
PATNet	ECCV'22	65.35	50.63	68.34	61.44
IFA	CVPR'24	50.43	31.32	30.74	37.50
FAMNet	AAAI'25	86.64	51.82	76.26	71.58
MAUP	MICCAI'25	88.36	52.74	78.29	73.13
APEX-SAM	Ours	**92.75**	**68.41**	**88.23**	**83.13**

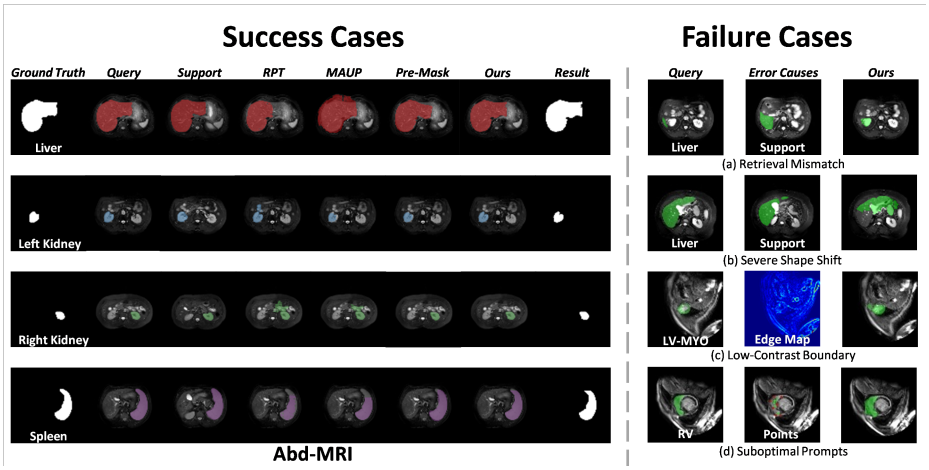

Fig. 3. Qualitative and failure cases. Left: successful cross-domain masks. Right: representative failure modes.

degrades direct feature matching, while cardiac structures such as LV-MYO have thin, irregular morphology that limits the benefit of geometric alignment.

Figure 3 visualises successful cross-domain masks alongside four failure modes: low-contrast boundaries (27%), severe shape shift (35%), retrieval mismatch (23%), and suboptimal prompts (15%).

3.3 Ablation Study

Module-wise gains are +7.8 pp for QAR, +6.1 pp for APM, and +5.5 pp for HMF; the thresholded append-only memory update contributes an additional +4.0 pp by continuously enriching the bank with high-confidence inference samples.

Table 3. Ablation on core designs and memory update strategy (Dice %).

Configuration	Enabled Modules			Memory Rule	Mean Dice
	QAR	APM	HMF		
Prompt-only baseline	✗	✗	✗	–	72.4
+ QAR	✓	✗	✗	Fixed	80.2
+ QAR + APM	✓	✓	✗	Fixed	86.3
+ QAR + APM + HMF	✓	✓	✓	Fixed	91.8
+ Full + thresholded append-only (ours)	✓	✓	✓	Thresholded append	**95.81**

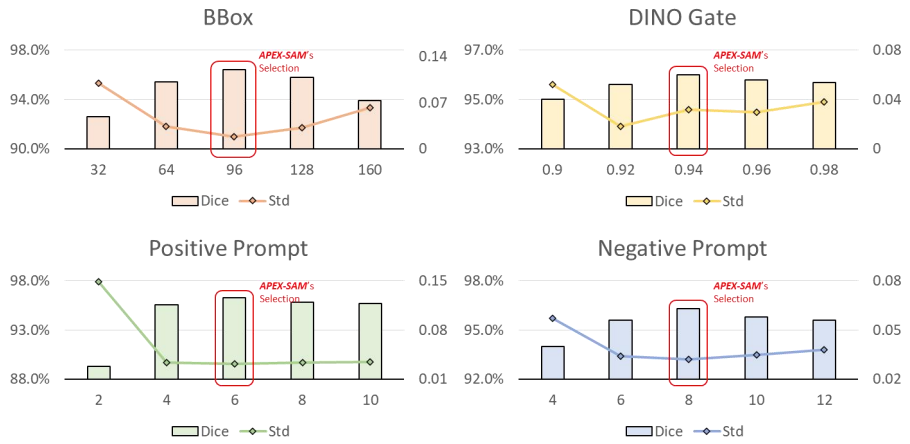

Fig. 4. Hyperparameter sensitivity. Mean Dice (bars, left axis) and standard deviation (lines, right axis) on Abd-MRI; each parameter is swept independently. Red boxes mark the selected values.

3.4 Hyperparameter Sensitivity

Figure 4 sweeps key hyperparameters on Abd-MRI; Dice remains stable within ±one step of each selected value, confirming robust deployment across varied clinical scenarios.

4 Conclusion

APEX-SAM is a training-free framework for cross-domain few-shot medical image segmentation that chains quality-aware expert retrieval, anatomy-aware prompt mining, and hybrid multi-modal fusion without any gradient update. Experiments on three benchmarks confirm that reliable support selection, geometry-aware alignment, and complementary prompt fusion are key to robust cross-domain adaptation, while the append-only bank and frozen parameters reduce deployment barriers and facilitate clinical adoption without infrastructure overhead. The current implementation still assumes a known target anatomy label and invokes multiple SAM prompt branches, which increases inference cost

relative to single-pass prompting. Future work will explore multi-class and 3D extensions with uncertainty-aware retrieval.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (No. 62176242) and Hubei Provincial Natural Science Foundation of China (No. 2025AFB832).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aurenhammer, F.: Voronoi diagrams: a survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR) **23**(3), 345–405 (1991)
2. Azad, R., Aghdam, E.K., Rauland, A., Jia, Y., Avval, A.H., Bozorgpour, A., Karimijafarbigloo, S., Cohen, J.P., Adeli, E., Merhof, D.: Medical image segmentation review: The success of u-net. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10076–10095 (2024)
3. Bo, Y., Zhu, Y., Li, L., Zhang, H.: Famnet: Frequency-aware matching network for cross-domain few-shot medical image segmentation. In: AAAI Conference on Artificial Intelligence (2025)
4. Carion, N., et al.: SAM 3: Segment anything with concepts. In: International Conference on Learning Representations (2026), https://openreview.net/forum?id=r35clVtGzw
5. Cheng, Z., Wang, S., Xin, T., Zhou, T., Zhang, H., Shao, L.: Few-shot medical image segmentation via generating multiple representative descriptors. IEEE Transactions on Medical Imaging **43**(6), 2202–2214 (2024)
6. Cheng, Z., Zhu, Y., Wang, S., Xin, T., Zhang, H.: Cascaded alternating refinement transformer for few-shot medical image segmentation. ACM Transactions on Intelligent Systems and Technology (2025)
7. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
9. Landman, B., Xu, Z., Igelsias, J., Styner, M., Langerak, T., Klein, A.: Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In: Proceedings of the MICCAI—Workshop Challenge. vol. 5, p. 12 (2015)
10. Lei, S., Zhang, X., He, J., Chen, F., Du, B., Lu, C.T.: Cross-domain few-shot semantic segmentation. In: European Conference on Computer Vision. pp. 73–90 (2022)
11. Liu, Y., Xiao, H., Chai, J., Zhang, Y., Wang, R., Meng, Z., Luo, Z.: Synpo: Boosting training-free few-shot medical segmentation via high-quality negative prompts (2025), https://arxiv.org/abs/2506.15153
12. Nie, J., Xing, Y., Zhang, G., Yan, P., Xiao, A., Tan, Y.P., Kot, A.C., Lu, S.: Cross-domain few-shot segmentation via iterative support-query correspondence mining. IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 3380–3390 (2024)

13. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
14. Ouyang, C., Biffi, C., Chen, C., Kart, T., Qiu, H., Rueckert, D.: Self-supervised learning for few-shot medical image segmentation. IEEE Transactions on Medical Imaging **41**(7), 1837–1848 (2022)
15. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
16. Tang, S., Yan, S., Qi, X., Gao, J., Ye, M., Zhang, J., Zhu, X.: Few-shot medical image segmentation with high-fidelity prototypes. Medical Image Analysis **100**, 103412 (2025)
17. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 9197–9206 (2019)
18. Wang, R., Yang, Z., Song, Y.: Osam-fundus: A training-free, one-shot segmentation framework for optic disc and cup in fundus images. Biomedical Signal Processing and Control **100**, 107069 (2025)
19. Wu, H., Xiao, F., Liang, C.: Dual contrastive learning with anatomical auxiliary supervision for few-shot medical image segmentation. In: European Conference on Computer Vision. pp. 417–434 (2022)
20. Yang, Y., Fang, X., Li, X., Han, Y., Yu, Z.: 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)
21. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
22. Zhang, S., Metaxas, D.: On the challenges and perspectives of foundation models for medical image analysis. Medical Image Analysis **91**, 102996 (2024)
23. Zhu, Y., Li, M., Ye, Q., Wang, S., Xin, T., Zhang, H.: Robustemd: Domain robust matching for cross-domain few-shot medical image segmentation (2025), https://arxiv.org/abs/2410.01110
24. Zhu, Y., Wang, S., Xin, T., Zhang, H.: Few-shot medical image segmentation via a region-enhanced prototypical transformer. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 271–280. Springer (2023)
25. Zhu, Y., Zhang, H.: MAUP: Training-free Multi-center Adaptive Uncertainty-aware Prompting for Cross-domain Few-shot Medical Image Segmentation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. vol. LNCS 15966. Springer Nature Switzerland (September 2025)
26. Zhuang, X.: Multivariate mixture model for myocardial segmentation combining multi-source images. IEEE transactions on pattern analysis and machine intelligence **41**(12), 2933–2946 (2018)