When Adaptation Hurts: Connecting Representational Drift to OOD Failures in MedSAM Fine-Tuning.

Marko Haralović[0009−0004−1178−9964]1,2 ⋆, Sounic Akkaraju[0009−0003−7888−4176]1 ⋆⋆, Carlo Baretta[0009−0006−6680−1275]1, Vasil Zapryanov[0009−0002−1441−4039]1, and Alexia Briassouli[0000−0002−0545−3215]1

[1] University of Twente, Enschede, The Netherlands
[2] University of Zagreb, Zagreb, Croatia
{s.akkaraju,c.d.a.baretta,v.z.zapryanov}@student.utwente.nl; marko.haralovic@fer.hr; a.briassouli@utwente.nl

Abstract. Foundation models for medical image segmentation, like prompt-based MedSAM, generalize well across domains and modalities, often in zero or few-shot setups. However, their performance depends on the quality of prompts and the adaptation of the models to custom datasets. This work systematically examines how MedSAM generalizes across diverse medical imaging benchmarks, with six adaptation strategies: full-model and encoder-only LoRA, shallow and deep visual prompt tuning (VPT), and decoder-only and full fine-tuning. Models are trained on the International Skin Imaging Collaboration Challenge (ISIC 2018) dataset and evaluated under clean and increasingly noisy prompts on IN and Out-of-Distribution (OOD) datasets: close-OOD PH2 (dermoscopy), far-OOD BUSI (Breast Ultrasound Images Dataset) and CBIS-DDSM (Curated Breast Imaging Subset of the Digital Database for Screening Mammography). We show that adaptation improves performance on IN and close-OOD data but often reduces performance on far-OOD data. Full fine-tuning provides the best tradeoff, while encoder-only LoRA is the strongest parameter-efficient alternative, outperforming standard LoRA and VPT under far-OOD shifts. Using Centered Kernel Alignment (CKA), we show that far-OOD degradation is strongly associated with drift in decoder representations, whereas encoder similarity alone does not explain robustness. This suggests encoder-only LoRA provides stronger robustness than standard LoRA by adapting the encoder to distribution shift in visual features, while preserving the decoder pathway. We further show that random 0–100 pixel jitter on prompts produces more robust and better performing models. We thus conclude that robust MedSAM adaptation requires the combined consideration of prompt noise exposure, domain shift, and representation preservation. We release our code: https://github.com/ImSounic/medsam-vpt.

Keywords: Medical image segmentation · MedSAM · Visual prompt tuning · Out-of-distribution generalization · Representation similarity

⋆ Equal contribution.
⋆⋆ Equal contribution; corresponding author.

1 Introduction

Semantic segmentation is essential in many clinical applications, but requires costly manual annotation. While deep learning can reduce this burden, task-specific models often fail to generalize beyond their training domains. Segmentation foundation models (FMs) such as SAM [23] offer a prompt based alternative. Out-of-the-box SAM performance in medical imaging faces challenges, such as domain shift, weak boundaries, low contrast, and modality-specific appearance changes [17, 24]. This has motivated medical segmentation FMs [3, 15, 16] like MedSAM [15], which was trained on over one million medical images, across 150 tasks, but assumes region of interest (ROI) prompts at inference time.

The first challenge this approach faces is ROI sensitivity. Promptable models that require ROI prompts [15, 16] highly depend on the user-provided bounding box, which in clinical settings may be noisy, inconsistent, or unavailable, substantially affecting performance [3, 10, 26]. Previous work has therefore studied strategies to reduce ROI noise sensitivity, by identifying optimal prompting strategies [10, 26] or by introducing perturbed bounding-box prompts [21].

The second challenge is generalization under distribution shift, as models lose performance when applied beyond their training distribution [2, 7, 19]. Prior work has studied out-of-distribution (OOD) identification [2], robustness under distribution shift [19, 23], and generalization across domains [3, 17, 19].

The third challenge is the adaptation cost. Fine-tuning medical FMs can improve performance, but is computationally expensive. Parameter efficient methods such as LoRA [9] and visual prompt tuning (VPT) [11, 27] reduce this cost, but remain sensitive to dataset and prompt quality [10, 17, 22, 26]. As a result, adaptation may improve in-domain performance while degrading robustness to OOD data or noisy ROI prompts. Despite advances, prior work has not systematically connected fine-tuning strategy, prompt robustness, and domain generalization in medical segmentation FMs. Existing studies often focus on target domain adaptation, while separately acknowledging ROI noise and distribution shift. However, it remains unclear which fine-tuning strategies provide the best target-domain performance while preserving generalization, which lose most generalization capabilities, and which are most sensitive to noisy prompts.

Contributions This work studies how fine-tuning strategies affect segmentation performance under prompt and domain shift, and relates performance to the internal representations of MedSAM and its fine-tuned variants. In contrast to prior work that focuses mainly on adaptation, prompt refinement, or prompt automation, we jointly evaluate adaptation strategy, prompt robustness, and domain generalization. We compare zero-shot MedSAM with full and decoder-only fine-tuning, full-model and encoder-only LoRA, and shallow and deep VPT under controlled bounding-box perturbations in training and evaluation.

Our contributions are threefold. First, we systematically evaluate bounding-box perturbation, using models trained with clean boxes, fixed-noise boxes, and variable-noise boxes, then evaluated across multiple perturbation levels. Second, we compare full fine-tuning and parameter-efficient fine-tuning in the same

perturbation-aware setting to identify which methods remain robust when ROI prompts are imperfect, while test data shifts away from the target domain. Finally, we link layerwise representational similarity, measured with centered kernel alignment (CKA) [12], to in- and OOD segmentation performance, and test whether representation preservation is associated with OOD generalization.

2 Related Work

Recent progress in foundation models has led to segmentation systems that transfer well to medical imaging. MedSAM [15] is a notable example, adapting SAM [23] to medical images, achieving strong performance across modalities, with extended following versions [15, 16, 24, 25]. Medical datasets are often small and costly to annotate, so foundation models are often evaluated in zero-shot settings. However, out-of-the-box performance is often suboptimal, largely due to distribution shift between pretraining and target clinical data [3, 17, 23]. Fine-tuning on target domain data can improve performance, but it is computationally expensive, requires annotated data, and may reduce generalization ability.

Therefore, lightweight methods have been explored, like prompt tuning [6], visual prompt tuning [5, 8, 11, 20], and low rank adaptation (LoRA) adapters [6, 9]. These methods aim to preserve the benefits of large pre-trained models while reducing training cost, but remain sensitive to task design and prompt quality.

2.1 Efficient fine-tuning methods

MedSAM architecture MedSAM [15] is the backbone to be evaluated under zero-shot and fine-tuned settings. It consists of an image encoder, prompt encoder, and mask decoder. The image encoder is a Vision Transformer (ViT) with 12 transformer blocks that produce dense image embeddings. The prompt encoder embeds user prompts, such as bounding boxes, which the two-layer mask decoder fuses with image embeddings to produce segmentation masks.
Prompt tuning Prompt tuning freezes most of the network and learns only a small number of task specific parameters, reporting good performance in [6].
Visual prompt tuning (VPT) VPT extends prompt adaptation to vision models by learning a set of prompts while freezing the backbone [8, 11], improving performance in a low-data regime [27], supported by recent SAM variants [5, 20].
Low Rank Adaptation (LoRA) LoRA is a parameter-efficient fine-tuning strategy that updates a model through low-rank trainable weight matrices rather than modifying the full weight tensors, while keeping backbone fixed [6, 9].

2.2 Robustness and out-of-distribution (OOD) generalization

Generalization under domain shift remains a challenge in medical segmentation, since models often degrade when transferred across scanners, hospitals, or modalities [2, 7]. Related studies have focused on OOD identification [2], robustness under shift [23], and generalization across domains [3, 17].

2.3 Prompt sensitivity to noise

Prior work has shown that promptable segmentation models are sensitive to prompt type and quality [3, 10, 17, 24, 26], and that perturbing the box can significantly reduce segmentation accuracy, with [3], reporting strong variation across datasets and prompt settings [17, 24]. RobustMedSAM [14] claims image encoder preserves medical priors, while the mask decoder governs corruption robustness. Several works address this limitation by refining or augmenting the input prompt [10, 26], through bounding box correction [26], refining noisy prompts [10] or by training on perturbed bounding boxes [15, 21], where imprecise bounding boxes serve as an unsupervised localization task [22].

2.4 Feature representation comparison

Comparing representations is a well-studied problem in machine learning [12]. Comparing only final embeddings can miss layer-specific changes, originating from fine tuning strategies that adapt different parts of MedSAM. We therefore compare internal feature representations across layers using Centered Kernel Alignment (CKA) [12]. For each model, we compute CKA between its layer representations and those of zero shot MedSAM, using the same input images.

3 Methodology

3.1 Perturbed bounding box prompts

All images are resized to 1024×1024. The bounding box prompts undergo evaluation perturbations of $p \in \{0, 20, 50, 100, 200\}$, corresponding to mean side length increases of 0%, 1.95%, 4.88%, 9.77%, and 19.53% of image size, and mean box area increases of 0%, 17.61%, 48.85%, 112.67%, and 280.14%, reaching as high as 1327.9% for small lesions in CBIS-DDSM.

During training: We train models under three prompt settings: clean bounding boxes, bounding boxes perturbed by a fixed 20 pixels, and bounding boxes randomly perturbed between 0 and 100 pixels. This allows us to test whether exposure to imperfect prompts improves localization robustness, and whether training with a fixed perturbation causes overfitting to a specific noise level.

During evaluation We evaluate each trained model across five bounding box perturbation levels: 0, 20, 50, 100, and 200 pixels. This tests robustness under increasing prompt noise and evaluates whether models trained with perturbations up to 100 pixels generalize to more severe 200 pixel perturbations in inference.

3.2 Datasets and in- vs out-of-domain separation

We use ISIC 2018 [4] dermoscopic skin lesion images [CC BY-NC] for training and in-domain evaluation. PH^2 [18] [public research use], another dermoscopic skin lesion dataset, is used as a close out-of-domain dataset because it shares the same imaging modality and target structure but differs in acquisition source

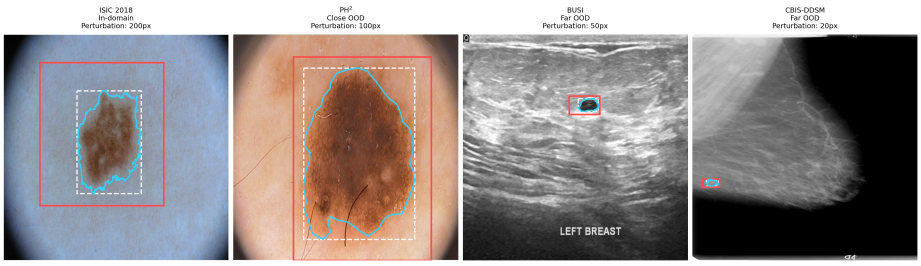

Fig. 1: Prompt perturbation examples.

and dataset distribution. To evaluate stronger domain shift, we use BUSI breast ultrasound images [1] [CC BY 4.0] and CBIS-DDSM mammography images [13] [TCIA data usage policy] as far out-of-domain datasets. This separation allows testing whether adaptation improves performance only for target domain or preserves generalization across different clinical imaging domains.

3.3 Representation similarity measurement

We measure representation similarity using linear Centered Kernel Alignment (CKA). For each adapted checkpoint, we compare layerwise representations against zero shot MedSAM using the same images and prompts. CKA is computed across the image encoder, prompt encoder, and mask decoder, with higher values indicating stronger preservation of the original MedSAM representation.

4 Experiments and results

4.1 Implementation details

Training paradigms implementation All methods use the MedSAM backbone with a frozen prompt encoder, so bounding-box prompts are encoded identically across experiments. Full fine-tuning updates the image encoder and mask decoder, while decoder-only fine-tuning updates only the mask decoder. Full LoRA adds trainable adapters to both the image encoder and decoder, whereas encoder-only LoRA adds adapters only to the image encoder. VPT freezes the image encoder and adds learnable visual prompt tokens either once at the encoder input for shallow VPT or at every encoder block for deep VPT; in both VPT settings, the mask decoder is trained.

Training details All models were trained for 5 epochs on ISIC 2018 at 1024×1024 resolution using AdamW, mixed precision, and an equally weighted Dice + cross-entropy loss. The 2,594 ISIC images were split 80/10/10 into train, validation, and test sets. Experiments used 8 GPUs with 16–22 GB memory, while training lasted 8 hours. Full fine-tuning used batch size 1 and learning rate 10^{-5}, while parameter-efficient methods used batch size 4 and learning

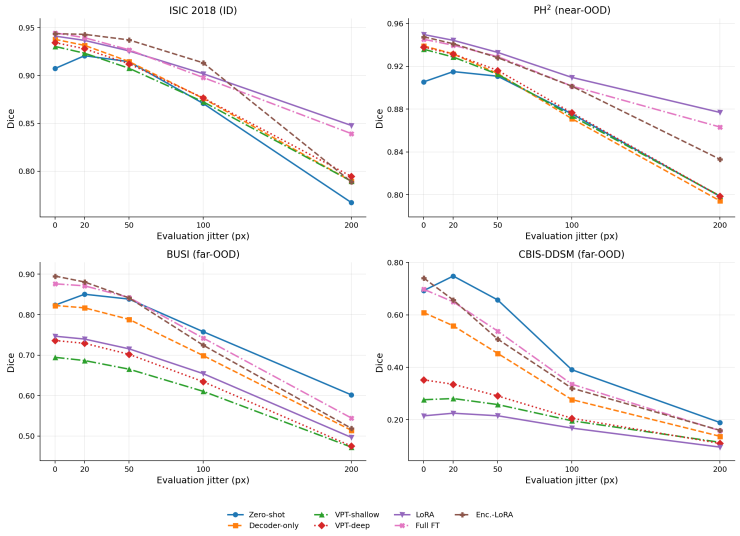

Fig. 2: Robustness to bounding-box perturbation under variable training jitter. Dice is reported across evaluation jitter levels for all adaptation strategies on ID, near-OOD, and far-OOD datasets.

rate 10^{-4}. Hyperparameters were fixed across datasets, Dice was reported for readability, and each configuration was repeated across three random seeds. For LoRA, we used rank (r=8), scaling factor ($\alpha = 16$), and dropout (0.0); for VPT, we used 10 prompt tokens per insertion point. The number of trainable parameters was 93.7M for full fine-tuning, 4.4M for full LoRA, 4.1M for encoder-only LoRA, 4.1M for decoder-only fine-tuning, and 4.1M/4.2M for shallow/deep VPT. CKA and OOD evaluation were performed on 200 PH2, 647 BUSI, and 362 CBIS-DDSM images. For ISIC, CKA was computed on all 2,594 images.

4.2 Per-method performance across variable evaluation jitter

Figure 2 summarizes the robustness of each method across jitter levels. For each method, we average performance across the three training prompt perturbations: 0, fixed 20, and random 0–100 pixels. Overall, performance degrades substantially as evaluation jitter increases. The results also show variability across adaptation methods. Full fine-tuning and encoder-only LoRA are the strongest and most robust methods overall. Standard LoRA, drops substantially on OOD datasets, despite strong in-domain performance, as reported in the literature [2]. At higher jitter levels, zero-shot outperforms all adapted models on BUSI and CBIS-DDSM. Only full fine-tuning and encoder-LoRA maintain far OOD performance close to zero-shot while still improving in-domain performance.

Table 1: Mean Dice gain over zero-shot by regime, method, and training jitter. Values are averaged over three runs; the Mean column reports mean ± standard deviation. **Bold** and underline indicate the best and second-best methods.

Regime	Method	0	20	r100	Mean	Best train	Best method
ID	Dec. only	-0.016	-0.007	0.014	-0.003 ± 0.002	r100	
	VPT-S	-0.008	-0.001	0.009	-0.000 ± 0.003	r100	
	VPT-D	-0.010	-0.005	0.013	-0.001 ± 0.002	r100	
	LoRA	0.012	0.016	0.035	$\mathbf{0.021 \pm 0.001}$	r100	**LoRA**
	Enc-LoRA	-0.027	-0.028	0.053	-0.001 ± 0.002	r100	
	Full FT	-0.006	-0.005	0.034	$\underline{0.007 \pm 0.002}$	r100	
Close OOD	Dec. only	-0.016	-0.007	0.009	-0.005 ± 0.004	r100	
	VPT-S	-0.009	-0.004	0.009	-0.001 ± 0.003	r100	
	VPT-D	-0.011	-0.007	0.011	-0.002 ± 0.001	r100	
	LoRA	0.015	0.020	0.042	$\mathbf{0.026 \pm 0.001}$	r100	**LoRA**
	Enc-LoRA	-0.017	-0.022	0.023	-0.005 ± 0.002	r100	
	Full FT	-0.007	-0.003	0.035	$\underline{0.008 \pm 0.002}$	r100	
Far OOD	Dec. only	-0.036	-0.032	-0.088	-0.052 ± 0.010	20	
	VPT-S	-0.127	-0.161	-0.230	-0.172 ± 0.002	0	
	VPT-D	-0.126	-0.139	-0.198	-0.154 ± 0.015	0	
	LoRA	-0.142	-0.157	-0.228	-0.176 ± 0.017	0	
	Enc-LoRA	-0.024	-0.052	-0.054	-0.043 ± 0.010	0	
	Full FT	-0.014	-0.011	-0.029	$\mathbf{-0.018 \pm 0.005}$	20	**Full FT**
Overall	Dec. only	-0.023	-0.015	-0.022	-0.020 ± 0.005	20	
	VPT-S	-0.048	-0.055	-0.071	-0.058 ± 0.003	0	
	VPT-D	-0.049	-0.050	-0.058	-0.052 ± 0.004	0	
	LoRA	-0.038	-0.040	-0.051	-0.043 ± 0.006	0	
	Enc-LoRA	-0.023	-0.034	0.007	$\underline{-0.016 \pm 0.004}$	r100	
	Full FT	-0.009	-0.006	0.013	$\mathbf{-0.001 \pm 0.003}$	r100	**Full FT**

4.3 Performance regarding method and training jitter

Table 1 summarizes Dice gain relative to zero-shot MedSAM across adaptation methods and training jitter. Overall, zero-shot performance is difficult to improve upon, especially on far-OOD datasets, while in-domain and close-OOD gains often come with reduced generalization.

Standard LoRA achieves the largest ID and close-OOD gains, but drops sharply under far-OOD shift, with a mean gain of -0.176. In contrast, encoder-only LoRA is more robust, with a far-OOD mean of -0.043 and the second-best overall mean of -0.016. Better generalization by restricting LoRA to the encoder is consistent with the CKA analysis (Section 4.4) that shows preserving decoder representations is associated with far-OOD robustness. Visual prompt tuning does not provide a favorable tradeoff: shallow and deep VPT show negligible ID and close-OOD gains, with large far-OOD drops. Training with random 0–100-pixel jitter generally produces the best adapted models overall, especially for full fine-tuning and encoder-only LoRA.

Across all regimes, full fine-tuning remains the strongest method, achieving the best far-OOD mean gain (-0.018) and best overall mean gain (-0.001). Encoder-only LoRA is the strongest parameter-efficient alternative, clearly outperforming standard LoRA and VPT under OOD shift.

4.4 CKA correlation with OOD performance

The following CKA analysis links performance to decoder representational drift, as each training paradigm inevitably shifts internal model representations across

Table 2: Spearman/Pearson correlations between CKA and Dice gain over zero-shot. Stars denote false discovery rate corrected significance across 30 tests: * $q < 0.05$, ** $q < 0.01$, *** $q < 0.001$. Total sample size is $n = 3000$.

CKA feature	ID (ISIC 2018)	Close-OOD (PH^2)	Far-OOD (BUSI/CBIS)
IoU token layer	0.097 / 0.106	-0.018 / -0.021	0.760^{***} / 0.692^{***}
Output layer	0.060 / 0.026	0.004 / -0.017	0.755^{***} / 0.718^{***}
Decoder layers	-0.075 / -0.107	-0.022 / -0.079	0.726^{***} / 0.687^{***}
Upscaled emb.	0.195 / 0.050	0.294^{*} / 0.213	0.713^{***} / 0.791^{***}
Encoder layers	-0.282^{*} / -0.299^{*}	-0.281^{*} / -0.352^{**}	0.091 / 0.059

layers. Here, we connect which MedSAM internal representations correlate with performance gains. Table 2 shows that CKA is weakly correlated with performance gains in ID and close-OOD regimes, but strongly correlated with far-OOD gains for decoder-side representations. The IoU token output, output layer, decoder layers, and upscaled embedding show high positive far-OOD correlations. In contrast, encoder-layer CKA is weakly negative across regimes, indicating that encoder similarity to zero-shot MedSAM alone does not explain robustness.

Overall, far-OOD degradation is tied to drift in the mask decoder and output representations, suggesting that generalization is harmed when output-side representations are disrupted.. Therefore we include encoder-only LoRA fine-tuning, which performs better both overall and on far-OOD data than standard LoRA, while remaining competitive with full FT. This analysis, often ignored by works focusing mainly on Dice optimization, suggests encoder adaptation of MedSAM can result in a much more robust model overall.

Because the pooled CKA observations share images and backbones across methods, seeds, and perturbation levels, they are not independent and the effective sample size is smaller than the nominal n. The reported p-values should therefore be interpreted as descriptive rather than strict hypothesis tests. This does not change the qualitative conclusion that decoder-side similarity tracks far-OOD performance more closely than encoder similarity.

5 Conclusion

We systematically evaluated MedSAM adaptation under joint domain shift and bounding-box prompt perturbation, both during training and evaluation. By training with clean boxes, fixed 20-pixel, and random 0–100 pixel perturbations, and evaluating across increasing prompt noise, we show that prompt robustness depends strongly on both the adaptation strategy and the training jitter regime. Zero-shot MedSAM remains difficult to improve upon when far-OOD robustness is considered, especially under severe prompt perturbation.

Across adaptation methods, full fine-tuning provides the strongest overall tradeoff, while encoder-only LoRA is the best parameter-efficient alternative. Compared with standard LoRA, decoder-only fine-tuning, and shallow or deep

visual prompt tuning, encoder-only LoRA better preserves far-OOD performance while maintaining strong in-domain and close-OOD performance.

Our CKA analysis links performance differences to decoder representational drift. Far-OOD generalization is strongly associated with preserving decoder and output representations, whereas encoder similarity does not explain robustness. This supports encoder-only LoRA as a practical CKA-informed adaptation strategy: adapting the image encoder while avoiding decoder/output drift leads to a stronger robustness tradeoff than standard LoRA. Overall, variable 0–100 pixel jitter training produces the most reliable adapted models.

Deployment and limitations. Full fine-tuning is strongest for far-OOD transfer, while encoder-only LoRA is the strongest parameter-efficient method. Because training is limited to dermoscopy, the far-OOD results reflect combined modality and target-structure shift and should be interpreted primarily through method rankings.

Disclosure of Interests. The authors have no competing interests in the paper.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020). https://doi.org/10.1016/j.dib.2019.104863
2. Arega, T.W., Bricq, S., Meriaudeau, F.: Post-hoc out-of-distribution detection for cardiac mri segmentation. Computerized Medical Imaging and Graphics **119**, 102476 (2025). https://doi.org/10.1016/j.compmedimag.2024.102476
3. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K.: Sam on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035 (2023). https://doi.org/10.48550/arXiv.2305.00035
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., Kittler, H., Halpern, A.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
5. Cui, R., Liu, L., Zou, J., Hu, X., Pei, J., Qin, J.: Taming large vision model for medical image segmentation via dual visual prompt tuning. Computerized Medical Imaging and Graphics **124**, 102608 (2025). https://doi.org/https://doi.org/10.1016/j.compmedimag.2025.102608, `https://www.sciencedirect.com/science/article/pii/S089561112500117X`
6. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S.A., Hospedales, T.: Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In: Burgos, N., Petitjean, C., Vakalopoulou, M., Christodoulidis, S., Coupe, P., Delingette, H., Lartizien, C., Mateus, D. (eds.) Proceedings of The 7nd International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 250, pp. 406–425. PMLR (03–05 Jul 2024), `https://proceedings.mlr.press/v250/dutt24a.html`
7. Gutbrod, M., Rauber, D., Weber Nunes, D., Palm, C.: Openmibood: Open medical imaging benchmarks for out-of-distribution detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
8. He, A., Wu, Y., Wang, Z., Li, T., Fu, H.: Dvpt: Dynamic visual prompt tuning of large pre-trained models for medical image analysis. Neural Networks **185**, 107168

(2025). https://doi.org/https://doi.org/10.1016/j.neunet.2025.107168, https://www.sciencedirect.com/science/article/pii/S0893608025000474

9. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
10. Huang, Y., Yang, X., Zhou, H., Cao, Y., Dou, H., Dong, F., Ni, D.: Robust box prompt based sam for medical image segmentation. In: Machine Learning in Medical Imaging. pp. 1–11. Lecture Notes in Computer Science, Springer (2024). https://doi.org/10.1007/978-3-031-73290-4_1
11. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision. pp. 709–727. Springer (2022)
12. Kornblith, S., Norouzi, M., Lee, H., Hinton, G.: Similarity of neural network representations revisited. In: International Conference on Machine Learning. pp. 3519–3529. PMLR (2019)
13. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific Data **4**, 170177 (2017). https://doi.org/10.1038/sdata.2017.177
14. Li, J., Chen, M., Nnamdi, M.C., Tamo, J.B., Marteau, B.L., Wang, M.D.: Robustmedsam: Degradation-resilient medical image segmentation via robust foundation model adaptation. In: Proceedings of the CVPR 2026 Workshop on Computer Vision for Real-world Clinical Translation (2026), cV4Clinic Workshop
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature **15**, 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
16. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025)
17. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023). https://doi.org/10.1016/j.media.2023.102918
18. Mendonça, T., Ferreira, P.M., Marques, J.S., Marçal, A.R.S., Rozeira, J.: Ph2: A dermoscopic image database for research and benchmarking. Annual International Conference of the IEEE Engineering in Medicine and Biology Society pp. 5437–5440 (2013). https://doi.org/10.1109/EMBC.2013.6610779
19. Nguyen, D.M.H., et al.: On the out of distribution robustness of foundation models in medical image segmentation. In: International Workshop on Machine Learning in Medical Imaging (2023)
20. Piater, T., Barz, B., Freytag, A.: Prompt-tuning sam: From generalist to specialist with only 2,048 parameters and 16 training images. pp. 4688–4698 (06 2025). https://doi.org/10.1109/CVPRW67362.2025.00455
21. Rahman, M.M., Munir, M., Jha, D., Bagci, U., Marculescu, R.: Pp-sam: Perturbed prompts for robust adaption of segment anything model for polyp segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 4989–4995 (2024). https://doi.org/10.1109/CVPRW63382.2024.00504
22. Shi, Y., Ma, J., Yang, J., Wang, S., Zhang, Y.: Beyond pixel-wise supervision for medical image segmentation: From traditional models to foundation models. arXiv preprint arXiv:2404.13239 (2024). https://doi.org/10.48550/arXiv.2404.13239

23. Wang, A., Islam, M., Xu, M., Zhang, Y., Ren, H.: Sam meets robotic surgery: An empirical study on generalization, robustness and adaptation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops. Lecture Notes in Computer Science, vol. 14393, pp. 234–244. Springer (2023). https://doi.org/10.1007/978-3-031-47401-9_23
24. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., Qiao, Y.: Sam-med3d: Towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision Workshops (2024)
25. Wen, C., He, L.: Psam: Prompt-based segment anything model adaption for medical image segmentation. In: 2024 IEEE International Conference on Systems, Man, and Cybernetics. pp. 3324–3329 (2024). https://doi.org/10.1109/SMC54092.2024.10831390
26. Yao, X., Liu, H., Hu, D., Lu, D., Lou, A., Li, H., Deng, R., Arenas, G., Oguz, B., Schwartz, N., Byram, B.C., Oguz, I.: Fnpc-sam: Uncertainty-guided false negative/positive control for sam on noisy medical images. In: Medical Imaging 2024: Image Processing. Proceedings of SPIE, vol. 12926, p. 1292602 (2024). https://doi.org/10.1117/12.3006867
27. Zhao, M., Zhu, Z., Shi, J., Wang, Z., Chen, J., An, H., Yan, B.: Promptseg: Learning to segment medical image via visual prompts. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 1–5 (2025)