This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

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

Supplementary Material

A How hard of a task does each evaluation jitter represent?

In Table 1 we report that the same pixel-level jitter corresponds to very different effective prompt difficulty across datasets. For ISIC 2018 and PH^2, the increase in bounding-box area remains relatively moderate even at larger perturbations, whereas BUSI and especially CBIS-DDSM experience much larger relative ROI expansions, reflecting the smaller lesion sizes and more severe prompt degradation in those domains. This means that a fixed jitter value does not represent the same level of challenge across all datasets, and that far-OOD evaluation is substantially more demanding than in-domain evaluation. Taken together, these results support our use of jitter-aware evaluation as a necessary component for assessing robustness under noisy ROI prompts.

Table 1: Relative increase in tight bounding-box area under expansion due to evaluation jitter. Values report the mean area increase with respect to the original tight bounding box for each dataset and perturbation level.

Dataset	**pm=20**	**pm=50**	**pm=100**	**pm=200**
ISIC2018-test	11.70%	30.51%	66.40%	152.26%
PH2	6.57%	16.84%	35.11%	72.06%
BUSI	18.71%	51.36%	113.75%	270.96%
CBIS-DDSM	64.06%	193.49%	485.19%	1327.90%
Overall	**17.61%**	**48.85%**	**112.67%**	**280.14%**

B Boundary robustness under evaluation jitter

Table 2 reports boundary accuracy for clean bounding-box prompts. The adapted models obtain low median HD95 values on ISIC 2018 and PH^2. On the far-OOD datasets, the differences between methods are larger. Encoder-only LoRA, decoder-only fine-tuning, and full fine-tuning score relatively low boundary errors, while standard LoRA and the VPT variants have much wider interquartile ranges, particularly on BUSI and CBIS-DDSM.

Table 2: HD95 at zero evaluation jitter. Values are median [interquartile range] in pixels at 1024×1024; lower is better.

Method	ISIC 2018 (ID)	PH^2 (near-OOD)	BUSI (far-OOD)	CBIS-DDSM (far-OOD)
Zero-shot	6 [0, 15]	7 [1, 16]	14 [9, 25]	12 [7, 17]
Dec. only	0 [0, 3]	0 [0, 3]	4 [2, 17]	4 [3, 6]
VPT-S	0 [0, 3]	0 [0, 4]	21 [5, 387]	14 [8, 37]
VPT-D	0 [0, 3]	0 [0, 4]	25 [6, 394]	18 [10, 359]
LoRA	0 [0, 1]	0 [0, 1]	26 [6, 402]	47 [13, 438]
Enc-LoRA	0 [0, 0]	0 [0, 0]	3 [1, 11]	5 [3, 7]
Full FT	0 [0, 0]	0 [0, 0]	4 [1, 13]	4 [3, 6]

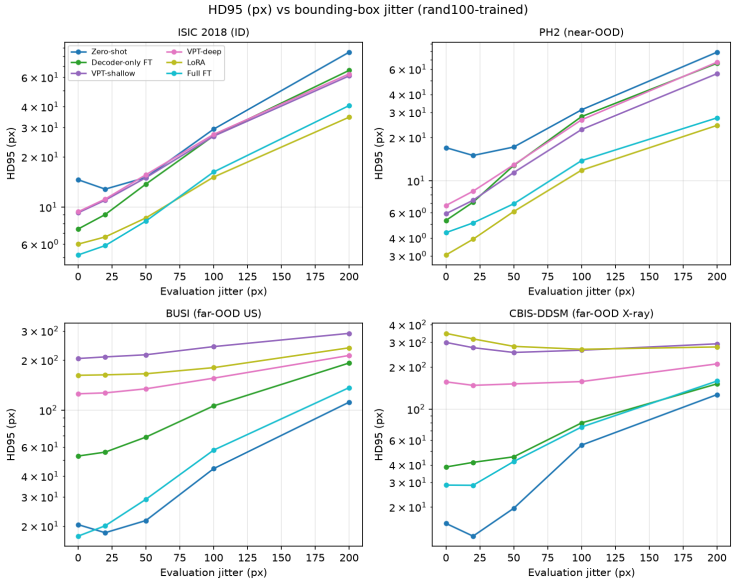


Fig. 1: HD95 across evaluation jitter levels for each adaptation method and dataset. Values are shown on a logarithmic scale in pixels at 1024×1024; lower is better.

Figure 1 extends this comparison across evaluation jitter levels. Boundary error generally increases as the bounding box is expanded, as the quality of the prompts degrades, but the rate of degradation differs across adaptation methods and datasets. The largest errors occur for standard LoRA and VPT on the far-OOD datasets.

C How does CKA similarity correspond to segmentation performance

Figures 2 and 3 show that CKA similarity is related to segmentation performance, but the relationship depends strongly on which part of the network is compared. The scatter plots indicate that higher similarity in decoder-side and output-side representations is associated with better far-OOD performance, while encoder similarity alone does not track robustness as reliably. The heatmap makes the same pattern easier to see across datasets and jitter levels, where the strongest correlations appear in the far-OOD regime. Overall, these figures support the claim that preserving decoder and output representations is more important for robustness than preserving encoder features alone.

Figure 2 plots the performance gain over zero-shot against overall CKA similarity for all settings. Each method has it's own plot and each point in the scatter plot corresponds to one configuration of the domains (ID, near-OOD, and far-OOD) and training jitter for a given adaptation method. The figure displays how the CKA–performance relationship differs across adaptation strategies while still aggregating over all domain and jitter settings.

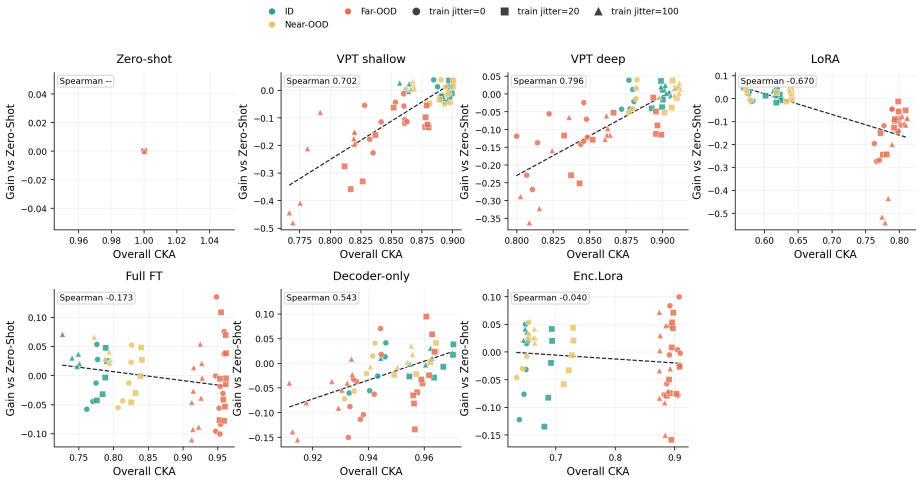


Fig. 2: CKA-performance relationship by method.

Figure 3 shows a plot per regime, where each plot shows the performance gain over zero-shot against overall CKA similarity for all settings. Each point in the scatter corresponds to one configuration of the adaptation method, training jitter, and evaluation jitter. Per method, the plots cover four datasets, three training jitters, and five evaluation jitters per method. The results shown in the four plots show how the relationship changes when moving from in-domain to far-OOD conditions.

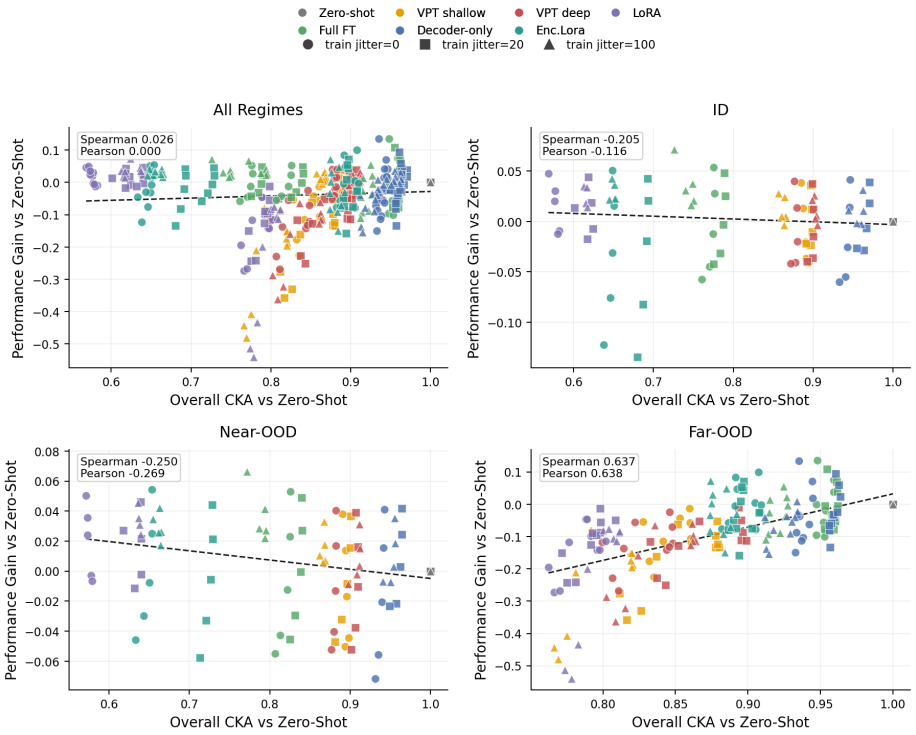


Fig. 3: Overview of the relationship between CKA similarity and segmentation performance.

Figure 4 shows a line plot split into 6 parts. The plot is split into encoder, image emb, prompt, decoder, tokens, and upscale sections. These are the different modules of the model, and the x-axis shows the different components within those modules that we fine-tune. The y-axis shows the linear CKA similarity values. Each line section represents a training method and shows how these different methods alter the internal state of a model compared to its original zero-shot state. The shaded regions around the curves denote variability across datasets and jitters. By showing how fine-tuning affects the similarity scores per method, per component of the model, this plot shows which components are stable and

which are sensitive to the fine-tuning task.
In the encoder (E00–E11), most adaptation strategies largely preserve the representation of the reference MedSAM model. The zero-shot model, by definition, remains at a similarity of 1.0, while adapted variants show modest but method-dependent drift, with approaches such as LoRA and encoder-only LoRA showing the strongest deviations.
In contrast, similarity drops much more sharply in the decoder and token components, particularly for VPT-shallow, VPT deep, and decoder-only tuning, indicating that these parts of the network undergo the largest representational changes during adaptation. The token outputs are substantially reshaped across methods, whereas LoRA variants and full fine-tuning tend to better preserve decoder-side structure, which is consistent with their stronger far-OOD robustness in the main experiments.

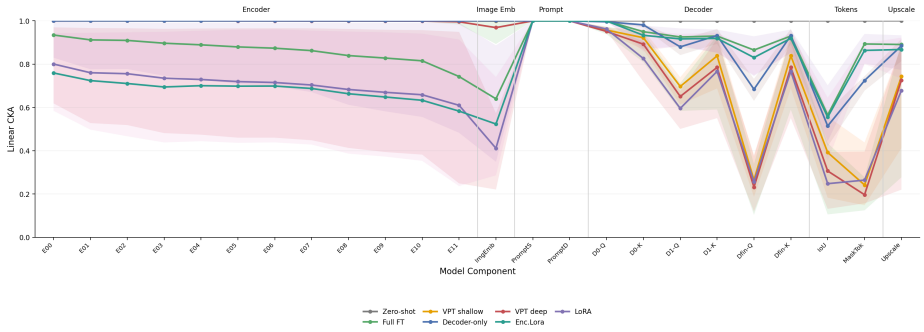


Fig. 4: CKA similarity across datasets and jitters using perfect bounding boxes.

D Degradation of performance depending on bounding box jitter

Figures 5 and 6 show that all methods degrade as bounding-box jitter increases, but they do not degrade at the same rate. Full fine-tuning and encoder-only LoRA have the shallowest degradation curves, especially when they are trained with variable jitter, while standard LoRA and VPT drop more quickly under stronger perturbation. The heatmap makes the same trend visible across datasets, showing that far-OOD performance is much more sensitive to prompt noise than in-domain performance. Specifically for CBIS-DDSM the Dice scores collapse to values below 0.2 for all methods when heavy jitter is introduced to the bounding box dimensions. These results support our claim that exposure to variable prompt noise during training improves robustness under realistic prompt perturbations.

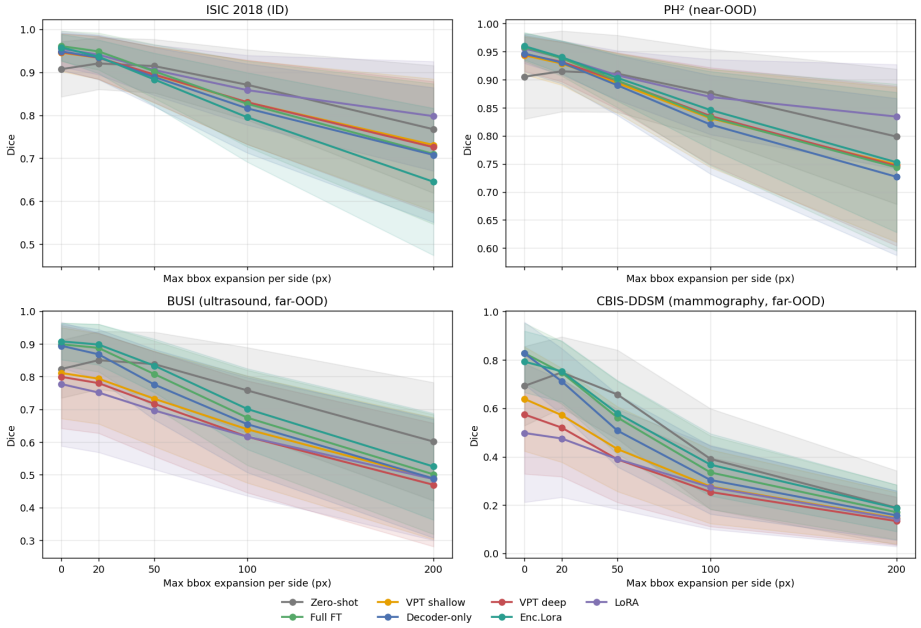


Fig. 5: Dice degradation with increasing bounding-box jitter by dataset and method.

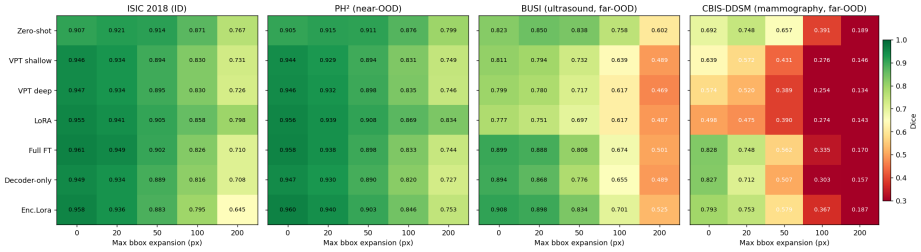


Fig. 6: Heatmap of the degradation in Dice with increasing bounding-box jitter, by datasets across models

E Segmentation performance breakdown across jitter levels

Table 3: Performance and CKA under bounding-box jitter. Dice performance and CKA are reported at zero jitter, averaged across evaluation jitters, and at 200 px jitter. For performance columns, underline marks the best method within each dataset and training-jitter block, while bold marks the best value for each dataset and evaluation column across all training-jitter settings. For CKA columns, bold and underline mark the best and second-best adapted methods, respectively, within each dataset and training-jitter block

Dataset	Training	Method	Dice@0	mDice	Dice@200	CKA@0	mCKA	CKA@200
ISIC 2018 *ID*	0	Zero-shot	0.9072	0.9019	0.7673	1.0000	1.0000	1.0000
		Decoder-only	0.9487	0.8798	0.7075	**0.9462**	**0.9432**	**0.9330**
		VPT-shallow	0.9457	0.8858	0.7306	0.8845	0.8908	0.8913
		VPT-deep	0.9472	0.8863	0.7256	0.8767	0.8791	0.8718
		LoRA	0.9545	0.9015	0.7978	0.5685	0.5794	0.5771
		LoRA Encoder-only	0.9580	0.8435	0.6450	0.6484	0.6462	0.6379
		Full FT	0.9609	0.8924	0.7098	0.7755	0.7738	0.7609
	20	Zero-shot	0.9072	0.9019	0.7673	1.0000	1.0000	1.0000
		Decoder-only	0.9462	0.8961	0.7408	**0.9703**	**0.9651**	**0.9546**
		VPT-shallow	0.9429	0.8948	0.7455	0.8969	0.8965	0.8914
		VPT-deep	0.9444	0.8890	0.7277	0.8987	0.8981	0.8924
		LoRA	0.9510	0.8998	0.7808	0.6184	0.6168	0.6012
		LoRA Encoder-only	0.9496	0.8414	0.6330	0.6925	0.6888	0.6802
		Full FT	0.9553	0.8987	0.7252	0.7886	0.7851	0.7754
	rand100	Zero-shot	0.9072	0.9019	0.7673	1.0000	1.0000	1.0000
		Decoder-only	0.9384	0.9042	0.7782	**0.9623**	**0.9569**	**0.9444**
		VPT-shallow	0.9316	0.9040	0.7951	0.8645	0.8623	0.8568
		VPT-deep	0.9330	0.9037	0.7886	0.9017	0.9026	0.8989
		LoRA	0.9421	0.9240	**0.8494**	0.6125	0.6136	0.6064
		LoRA Encoder-only	0.9436	0.9052	0.7890	0.6511	0.6498	0.6449
		Full FT	0.9446	**0.9244**	0.8386	0.7488	0.7425	0.7260
PH^2 *near-OOD*	0	Zero-shot	0.9054	0.9004	0.7986	1.0000	1.0000	1.0000
		Decoder-only	0.9466	0.8802	0.7269	**0.9423**	**0.9385**	**0.9314**
		VPT-shallow	0.9436	0.8846	0.7485	0.8905	0.8958	0.8933
		VPT-deep	0.9458	0.8883	0.7463	0.8825	0.8810	0.8765
		LoRA	0.9557	0.9053	0.8342	0.5716	0.5766	0.5733
		LoRA Encoder-only	**0.9598**	0.8804	0.7528	0.6535	0.6466	0.6331
		Full FT	0.9583	0.8898	0.7437	0.8258	0.8201	0.8064
	20	Zero-shot	0.9054	0.9004	0.7986	1.0000	1.0000	1.0000
		Decoder-only	0.9473	0.9024	0.7753	**0.9642**	**0.9593**	**0.9492**
		VPT-shallow	0.9421	0.8922	0.7516	0.9004	0.8934	0.8808
		VPT-deep	0.9444	0.8898	0.7465	0.9067	0.9066	0.9013
		LoRA	0.9516	0.9032	0.8258	0.6391	0.6337	0.6178
		LoRA Encoder-only	0.9497	0.8750	0.7408	0.7282	0.7234	0.7129
		Full FT	0.9543	0.8996	0.7533	0.8398	0.8350	0.8251
	rand100	Zero-shot	0.9054	0.9004	0.7986	1.0000	1.0000	1.0000
		Decoder-only	0.9408	0.9053	0.7916	**0.9552**	**0.9512**	**0.9406**

		VPT-shallow	0.9382	0.9106	0.8092	0.8674	0.8665	0.8619
		VPT-deep	0.9368	0.9060	0.7899	0.9114	0.9080	0.8975
		LoRA	0.9510	**0.9306**	**0.8724**	0.6363	0.6349	0.6241
		LoRA Encoder-only	0.9475	0.9102	0.8332	0.6631	0.6616	0.6544
		Full FT	0.9468	0.9265	0.8651	0.7936	0.7882	0.7718
CBIS-DDSM *far-OOD*	0	Zero-shot	0.6924	**0.5986**	**0.1889**	1.0000	1.0000	1.0000
		Decoder-only	0.8273	0.5076	0.1570	0.9353	0.9339	0.9345
		VPT-shallow	0.6388	0.4264	0.1461	0.8278	0.8361	0.8531
		VPT-deep	0.5744	0.3876	0.1336	0.7996	0.8104	0.8221
		LoRA	0.4979	0.3795	0.1433	0.7624	0.7710	0.7882
		LoRA Encoder-only	0.7926	0.5358	0.1872	0.9073	0.9070	0.9092
		Full FT	**0.8280**	0.5480	0.1704	**0.9479**	**0.9487**	**0.9533**
	20	Zero-shot	0.6924	**0.5986**	**0.1889**	1.0000	1.0000	1.0000
		Decoder-only	0.7872	0.5136	0.1565	**0.9607**	**0.9584**	0.9589
		VPT-shallow	0.4167	0.3157	0.1273	0.8116	0.8285	0.8515
		VPT-deep	0.5767	0.3965	0.1367	0.8324	0.8462	0.8663
		LoRA	0.5422	0.3951	0.1392	0.7718	0.7857	0.8088
		LoRA Encoder-only	0.7467	0.4796	0.1628	0.8957	0.8984	0.9078
		Full FT	0.8015	0.5577	0.1744	0.9547	0.9560	**0.9613**
	rand100	Zero-shot	0.6924	**0.5986**	**0.1889**	1.0000	1.0000	1.0000
		Decoder-only	0.6527	0.4749	0.1528	0.9115	**0.9166**	**0.9268**
		VPT-shallow	0.2496	0.2327	0.1101	0.7664	0.7767	0.7917
		VPT-deep	0.4049	0.3178	0.1241	0.8023	0.8172	0.8358
		LoRA	0.1785	0.2076	0.1073	0.7738	0.7849	0.8003
		LoRA Encoder-only	0.7410	0.4769	0.1592	0.8850	0.8864	0.8920
		Full FT	0.7064	0.5159	0.1626	**0.9152**	0.9137	0.9146
BUSI *far-OOD*	0	Zero-shot	0.8234	**0.8156**	**0.6017**	1.0000	1.0000	1.0000
		Decoder-only	0.8944	0.7663	0.4888	0.9442	0.9415	0.9370
		VPT-shallow	0.8112	0.7218	0.4889	0.8599	0.8605	0.8625
		VPT-deep	0.7995	0.7047	0.4695	0.8459	0.8466	0.8458
		LoRA	0.7775	0.6882	0.4870	0.7888	0.7952	0.8018
		LoRA Encoder-only	**0.9078**	0.7732	0.5253	0.8915	0.8887	0.8821
		Full FT	0.8991	0.7901	0.5014	**0.9591**	**0.9568**	**0.9528**
	20	Zero-shot	0.8234	**0.8156**	**0.6017**	1.0000	1.0000	1.0000
		Decoder-only	0.8828	0.7963	0.5370	**0.9628**	0.9602	0.9559
		VPT-shallow	0.7701	0.6944	0.4781	0.8787	0.8787	0.8792
		VPT-deep	0.8136	0.7329	0.4884	0.8935	0.8959	0.8990
		LoRA	0.8109	0.7362	0.5126	0.7981	0.7965	0.7927
		LoRA Encoder-only	0.8949	0.7734	0.5227	0.8972	0.8931	0.8864
		Full FT	0.8935	0.8094	0.5241	0.9611	**0.9608**	**0.9591**
	rand100	Zero-shot	0.8234	**0.8156**	**0.6017**	1.0000	1.0000	1.0000
		Decoder-only	0.8322	0.7771	0.5114	**0.9338**	**0.9319**	**0.9291**
		VPT-shallow	0.6723	0.6303	0.4509	0.8193	0.8197	0.8198
		VPT-deep	0.7631	0.7070	0.4747	0.8614	0.8620	0.8604
		LoRA	0.7474	0.7132	0.5179	0.8039	0.8086	0.8114
		LoRA Encoder-only	0.8954	0.7722	0.5190	0.8740	0.8744	0.8738

		Full FT	0.8783	0.8085	0.5135	0.9247	0.9253	0.9223

Table 4: Dice under bounding-box jitter. Models are trained with 0 px, 20 px, or random 0–100 px jitter and evaluated at 0–200 px jitter. Values are means over 3 seeds; underline denotes the best method within each training block and bold the best per dataset/evaluation jitter.

Dataset	**Train**	**Method**	**E0**	**E20**	**E50**	**E100**	**E200**
ISIC 2018 *trained on* rand100	pm=0	Zero-shot	0.9072	0.9206	0.9144	0.8708	0.7673
		Decoder-only	0.9487	0.9349	0.8900	0.8176	0.7102
		VPT-shallow	0.9449	0.9327	0.8952	0.8335	0.7352
		VPT-deep	0.9470	0.9343	0.8944	0.8281	0.7251
		LoRA	0.9556	0.9416	0.9040	0.8522	0.7867
		LoRA Encoder-only	0.9580	0.9364	0.8832	0.7951	0.6450
		Full FT	**0.9610**	**0.9483**	0.9024	0.8259	0.7115
	pm=20	Zero-shot	0.9072	0.9206	0.9144	0.8708	0.7673
		Decoder-only	0.9462	0.9365	0.9021	0.8331	0.7271
		VPT-shallow	0.9428	0.9326	0.9042	0.8476	0.7478
		VPT-deep	0.9432	0.9324	0.9014	0.8407	0.7380
		LoRA	0.9492	0.9392	0.9114	0.8629	0.7974
		LoRA Encoder-only	0.9496	0.9411	0.8950	0.7885	0.6330
		Full FT	0.9556	0.9438	0.9050	0.8314	0.7201
	rand100	Zero-shot	0.9072	0.9206	0.9144	0.8708	0.7673
		Decoder-only	0.9377	0.9314	0.9145	0.8761	0.7896
		VPT-shallow	0.9303	0.9232	0.9075	0.8725	0.7896
		VPT-deep	0.9343	0.9279	0.9121	0.8766	0.7946
		LoRA	0.9411	0.9366	0.9256	0.9018	**0.8478**
		LoRA Encoder-only	0.9436	0.9429	**0.9371**	**0.9132**	0.7890
		Full FT	0.9445	0.9394	0.9268	0.8979	0.8393
PH^2 *near-OOD* rand100	pm=0	Zero-shot	0.9054	0.9150	0.9107	0.8756	0.7986
		Decoder-only	0.9467	0.9312	0.8917	0.8229	0.7309
		VPT-shallow	0.9433	0.9298	0.8976	0.8360	0.7542
		VPT-deep	0.9456	0.9315	0.8961	0.8324	0.7425
		LoRA	0.9570	0.9384	0.9045	0.8600	0.8197
		LoRA Encoder-only	0.9598	0.9402	0.9032	0.8459	0.7528
		Full FT	**0.9583**	0.9385	0.8981	0.8320	0.7434
	pm=20	Zero-shot	0.9054	0.9150	0.9107	0.8756	0.7986
		Decoder-only	0.9449	0.9329	0.9019	0.8382	0.7513
		VPT-shallow	0.9426	0.9310	0.9040	0.8482	0.7591
		VPT-deep	0.9428	0.9308	0.9022	0.8431	0.7533
		LoRA	0.9511	0.9379	0.9128	0.8717	0.8338
		LoRA Encoder-only	0.9497	0.9363	0.9051	0.8429	0.7408
		Full FT	0.9537	0.9398	0.9044	0.8400	0.7516
	rand100	Zero-shot	0.9054	0.9150	0.9107	0.8756	0.7986
		Decoder-only	0.9390	0.9315	0.9132	0.8711	0.7944

Dataset	Train	Method	E0	E20	E50	E100	E200
		VPT-shallow	0.9359	0.9286	0.9125	0.8736	0.7990
		VPT-deep	0.9379	0.9314	0.9161	0.8766	0.7988
		LoRA	0.9496	**0.9442**	**0.9329**	**0.9096**	**0.8769**
		LoRA Encoder-only	0.9475	0.9410	0.9280	0.9014	0.8332
		Full FT	0.9452	0.9396	0.9290	0.9015	0.8633
BUSI *far-OOD*	pm=0	Zero-shot	0.8234	0.8504	0.8384	**0.7579**	**0.6017**
		Decoder-only	0.8935	0.8693	0.7785	0.6587	0.4915
		VPT-shallow	0.7921	0.7758	0.7190	0.6315	0.4868
		VPT-deep	0.8017	0.7842	0.7243	0.6288	0.4824
		LoRA	0.7801	0.7540	0.6975	0.6186	0.4903
		LoRA Encoder-only	**0.9078**	**0.8983**	0.8336	0.7012	0.5253
		Full FT	0.9007	0.8860	0.8051	0.6722	0.4996
	pm=20	Zero-shot	0.8234	0.8504	0.8384	0.7579	0.6017
		Decoder-only	0.8807	0.8717	0.8088	0.6879	0.5184
		VPT-shallow	0.7623	0.7509	0.7091	0.6261	0.4790
		VPT-deep	0.7702	0.7624	0.7224	0.6360	0.4853
		LoRA	0.7745	0.7586	0.7162	0.6347	0.5010
		LoRA Encoder-only	0.8949	0.8937	**0.8470**	0.7089	0.5227
		Full FT	0.8996	0.8889	0.8236	0.6902	0.5096
	rand100	Zero-shot	0.8234	0.8504	0.8384	0.7579	0.6017
		Decoder-only	0.8226	0.8168	0.7883	0.6989	0.5142
		VPT-shallow	0.6945	0.6865	0.6652	0.6105	0.4725
		VPT-deep	0.7360	0.7291	0.7019	0.6343	0.4751
		LoRA	0.7467	0.7396	0.7155	0.6542	0.4965
		LoRA Encoder-only	0.8954	0.8807	0.8414	0.7246	0.5190
		Full FT	0.8762	0.8712	0.8424	0.7422	0.5445
CBIS-DDSM *far-OOD*	pm=0	Zero-shot	0.6924	0.7483	**0.6570**	**0.3905**	**0.1889**
		Decoder-only	0.8280	0.7059	0.5039	0.3026	0.1574
		VPT-shallow	0.5662	0.5199	0.3961	0.2565	0.1374
		VPT-deep	0.5725	0.5102	0.3904	0.2582	0.1368
		LoRA	0.5094	0.4782	0.3883	0.2703	0.1467
		LoRA Encoder-only	0.7926	**0.7530**	0.5792	0.3668	0.1872
		Full FT	**0.8285**	0.7438	0.5618	0.3364	0.1713
	pm=20	Zero-shot	0.6924	0.7483	0.6570	0.3905	0.1889
		Decoder-only	0.7824	0.7012	0.5186	0.3040	0.1532
		VPT-shallow	0.4555	0.4369	0.3532	0.2380	0.1313
		VPT-deep	0.5117	0.4810	0.3910	0.2603	0.1369
		LoRA	0.4370	0.4193	0.3580	0.2472	0.1327
		LoRA Encoder-only	0.7467	0.6743	0.4987	0.3153	0.1628
		Full FT	0.8051	0.7398	0.5705	0.3414	0.1714
	rand100	Zero-shot	0.6924	0.7483	0.6570	0.3905	0.1889
		Decoder-only	0.6090	0.5583	0.4528	0.2764	0.1362
		VPT-shallow	0.2759	0.2804	0.2574	0.1956	0.1132
		VPT-deep	0.3518	0.3349	0.2903	0.2053	0.1091
		LoRA	0.2143	0.2244	0.2146	0.1676	0.0949

Dataset	Train	Method	E0	E20	E50	E100	E200
		LoRA Encoder-only	0.7410	0.6569	0.5074	0.3201	0.1592
		Full FT	0.6980	0.6497	0.5380	0.3353	0.1575