MedSAD-CLIP: Supervised CLIP with Token-Patch Cross-Attention for Medical Anomaly Detection and Segmentation

Supplementary Material

6. Experimental Details

Implementation Details. All experiments are implemented in *PyTorch* [29] and conducted on a workstation equipped with a single NVIDIA L40S GPU (48 GB memory). The proposed model is trained using the Adam optimizer [21] with a learning rate of 1×10^{-4} and a batch size of 32 for 50 epochs. Input images are resized to a spatial resolution of 240×240 before being fed into the model, and the learnable prompt length is 10 for our experiments. We set the number of learnable tokens to 10, and the margin parameter τ is fixed at 0.4 by default. We adopt the CLIP ViT-L/14 [30] as the baseline architecture, upon which our adapter and learnable prompt modules are integrated. All training and evaluation procedures strictly follow the same preprocessing and data partitioning strategy as in the baseline CLIP configuration to ensure fair comparison. During training, we employ standard data augmentations such as random horizontal flipping and normalization consistent with CLIP preprocessing.

Baseline Setups. To ensure a fair and comprehensive evaluation, we compare our proposed framework against several strong CLIP-based baselines under two configurations: (1) zero-/few-shot learning and (2) supervised learning. In the zero-/few-shot configuration, we follow prior medical anomaly detection studies that adapt CLIP pretrained model. Specifically, Adaclip [5], Aaclip [27], and Anomalyclip [49] serve as zero-shot baselines, while MVFA [17], MadCLIP [3], BGAD [44] represent few-shot variants. These models leverage the advantages of CLIP through prompt tuning or lightweight finetuning using either no data or a limited number of samples from the training set. In contrast, the supervised configuration employs the entire labeled dataset to finetune the CLIP model. For this setup, all the aforementioned baselines are trained in a fully supervised manner. This dual-configuration design enables a systematic comparison between zero-/few-shot adaptation and full supervision, highlighting differences in anomaly detection and segmentation performance across diverse medical datasets.

7. Comparison of Dice and AUC score

AUC (or pixel-level pAUC) is a widely used metric for anomaly detection because it evaluates the model's ability to correctly rank abnormal and normal samples across all thresholds. It is particularly useful when assessing global discriminative power or when the primary goal is to verify whether a model can reliably distinguish between healthy and abnormal cases. However, AUC has notable limitations that it does not reflect spatial localization quality, it remains insensitive to boundary errors, and high AUC values can arise even when the predicted anomaly maps are coarse. In clinical workflows, this can be problematic, as doctors must still manually examine the entire image to identify the exact lesion, increasing review time despite strong AUC scores. In contrast, the Dice coefficient directly measures spatial overlap between predictions and ground-truth lesions, making it far more indicative of precise localization. High Dice scores correspond to accurate, well-defined lesion boundaries, which substantially reduces the burden on clinicians by minimizing the amount of manual inspection required during image interpretation.

As shown in Tab. 4, both AdaCLIP[5] and MVFA[17] frequently achieve very high pAUC scores, often exceeding 90% and even reaching 99% in several datasets. However, their Dice scores remain substantially lower, with many cases falling below 50% and even dropping to single digits on challenging datasets such as Lung Infection. This discrepancy indicates that these models can correctly rank abnormal pixels above normal pixels on average, yet fail to produce spatially coherent lesion masks. In other words, pAUC reflects the model's ability to highlight anomalous regions in a coarse, score-based manner, while the low Dice values reveal that such signals lack the precise localization needed for reliable segmentation. Consequently, although high pAUC may suggest good anomaly discrimination, the absence of accurate boundaries forces clinicians to visually re-identify lesion extents, limiting the practical usefulness of these methods in real diagnostic workflows.

The qualitative results in Fig. 4 further illustrate this limitation. For recent zero/few-shot works [5, 17, 27], the predicted score maps often exhibit broad, diffused activations that roughly highlight anomalous areas but fail to correspond to the exact lesion geometry. These models tend to generate large blobs of high response or scattered noisy activations, indicating that they capture global anomaly cues but lack the spatial precision required for clean segmentation. As a result, although their score maps contain some degree of anomaly awareness, the predicted masks derived from these scores are either overly coarse or incomplete, with boundaries that deviate significantly from the ground truth. This behavior explains the consistently low Dice scores that the models can rank abnormal regions correctly, reflected in high pAUC, but cannot accurately delineate the lesion contours, leading to poor boundary quality and unre-

Table 4. Comparison of Dice, Accuracy, and pAUC across four datasets.

Method	Brain			Retina			Lung			Breast		
	Dice	Acc	pAUC	Dice	Acc	pAUC	Dice	Acc	pAUC	Dice	Acc	pAUC
Adaclip (zero) [5]	43.52	56.63	90.89	38.62	61.74	92.92	9.09	56.25	60.76	36.19	23.60	86.53
Adaclip (full) [5]	46.99	48.19	98.66	89.47	96.52	**99.68**	84.93	92.19	**99.63**	79.87	87.64	**93.18**
MVFA (zero) [17]	48.34	54.87	95.06	66.93	79.82	97.10	68.54	76.56	97.19	52.09	70.22	89.51
MVFA (full) [17]	64.71	92.68	99.18	91.18	97.12	99.53	73.19	94.53	98.07	76.57	88.76	90.79
MedSAD-CLIP	**89.47**	**96.34**	**99.54**	**93.18**	**97.37**	98.97	**87.16**	**99.22**	98.47	**84.96**	**91.01**	87.48

liable localization.

8. Additional Ablation Study on Hyperparameter Settings

8.1. Length of Learnable Tokens and Margin for Margin-based image-text Contrastive Loss

Table 5. Effect of the number of learnable tokens on Dice (%) and Accuracy (%) across four datasets.

#Tokens	Brain		Retina		Lung		Breast	
	Dice	Acc	Dice	Acc	Dice	Acc	Dice	Acc
10	**89.47**	**96.34**	**93.18**	97.37	**87.16**	**99.22**	**84.96**	**91.01**
20	86.37	93.90	89.68	97.36	86.71	97.65	75.28	89.88
35	85.42	89.02	90.73	**98.24**	86.44	99.21	76.41	89.89

Table 6. Effect of different margins in MC-Loss on Dice (%) and Accuracy (%) across four datasets.

Margin	Brain		Retina		Lung		Breast	
	Dice	Acc	Dice	Acc	Dice	Acc	Dice	Acc
0.2	84.87	89.02	92.18	97.36	84.64	98.43	76.40	**94.38**
0.4	**89.47**	**96.34**	**93.18**	**97.37**	**87.16**	**99.22**	**84.96**	91.01
0.6	84.66	93.90	89.69	96.49	86.21	98.45	76.38	90.44
0.8	81.17	82.92	62.92	96.49	85.53	98.43	75.28	92.69

Tab. 5 and Tab. 6 present two key ablations on the number of learnable prompt tokens and the margin value in the margin-contrastive loss. For learnable tokens, using 10 tokens consistently achieves the best performance across Brain, Retina, and Lung, with the highest Dice scores (89.47%, 93.18%, and 87.16%, respectively). Increasing the prompt length to 20 or 35 leads to clear degradation, indicating that excessive learnable tokens may introduce redundancy and degrade image-text alignment. For the contrastive margin, a margin of 0.4 provides the most stable and discriminative learning signal, producing the top Dice scores on Brain and Retina while avoiding the severe performance collapse observed at larger margins (e.g., Dice drops to 62.92% on Retina when margin = 0.8). These results collectively indicate that a compact prompt representation (10 tokens) combined with a moderate contrastive separation (margin = 0.4) yields the most robust alignment and segmentation performance.

8.2. Different Threshold for Dice score of Baselines

Table 7. Comparison of threshold-based Dice (%) across baseline models on four datasets.

	Aaclip[27]				**Anomalyclip**[49]			
Threshold	Brain	Retina	Lung	Breast	Brain	Retina	Lung	Breast
0.3	14.04	38.03	9.54	10.74	27.53	32.08	7.92	10.85
0.4	18.59	46.94	12.06	20.81	33.17	33.04	7.71	12.42
0.5	22.68	47.98	21.73	41.54	43.25	37.25	7.47	13.52
0.6	35.48	43.94	27.32	56.92	48.76	35.97	7.42	14.02
0.7	41.59	41.88	34.79	67.21	51.54	35.87	7.51	13.89
	MadCLIP[36]				**Adaclip**[5]			
Threshold	Brain	Retina	Lung	Breast	Brain	Retina	Lung	Breast
0.3	0.34	0.78	0.70	0.35	42.08	39.05	6.45	32.54
0.4	0.38	0.81	0.71	0.41	42.66	39.96	7.38	35.19
0.5	0.38	0.82	0.67	0.43	43.52	38.62	9.13	36.19
0.6	0.36	0.81	0.58	0.41	44.58	39.41	13.47	36.03
0.7	0.33	0.78	0.45	0.37	45.78	39.63	18.03	36.89

Tab. 7 presents an ablation study on threshold selection for converting raw anomaly score maps (ranging from 0 to 1) into binary segmentation masks for baseline models, ensuring a fair and objective comparison against our method. Since these raw models originally produced anomaly score with pixel value range from 0 to 1, and a decision threshold must be chosen to evaluate segmentation performance. While a default threshold of 0.5 is commonly used, we vary thresholds from 0.3 to 0.7 to examine the model performance variance. We do not use extremely low thresholds effectively classify nearly all pixels as anomalous (over-segmentation), whereas extremely high thresholds suppress most anomalies (under-segmentation).

Most models such as Anomalyclip, Aaclip, Adaclip exhibit noticeable fluctuations. For instance, Aaclip achieves its best Dice on Brain at threshold 0.7, yet performs worse at intermediate values, highlighting the instability of selecting a "best" threshold universally. Across the broader results, threshold 0.5 consistently provides either the highest or a competitive Dice score for most methods and datasets, without the extreme behavior observed at 0.3 or 0.7. In conclusion, 0.5 is adopted as a fair and balanced threshold for evaluating raw anomaly-score-based CLIP models, offering a stable compromise across models and datasets without favoring over- or under-segmentation.

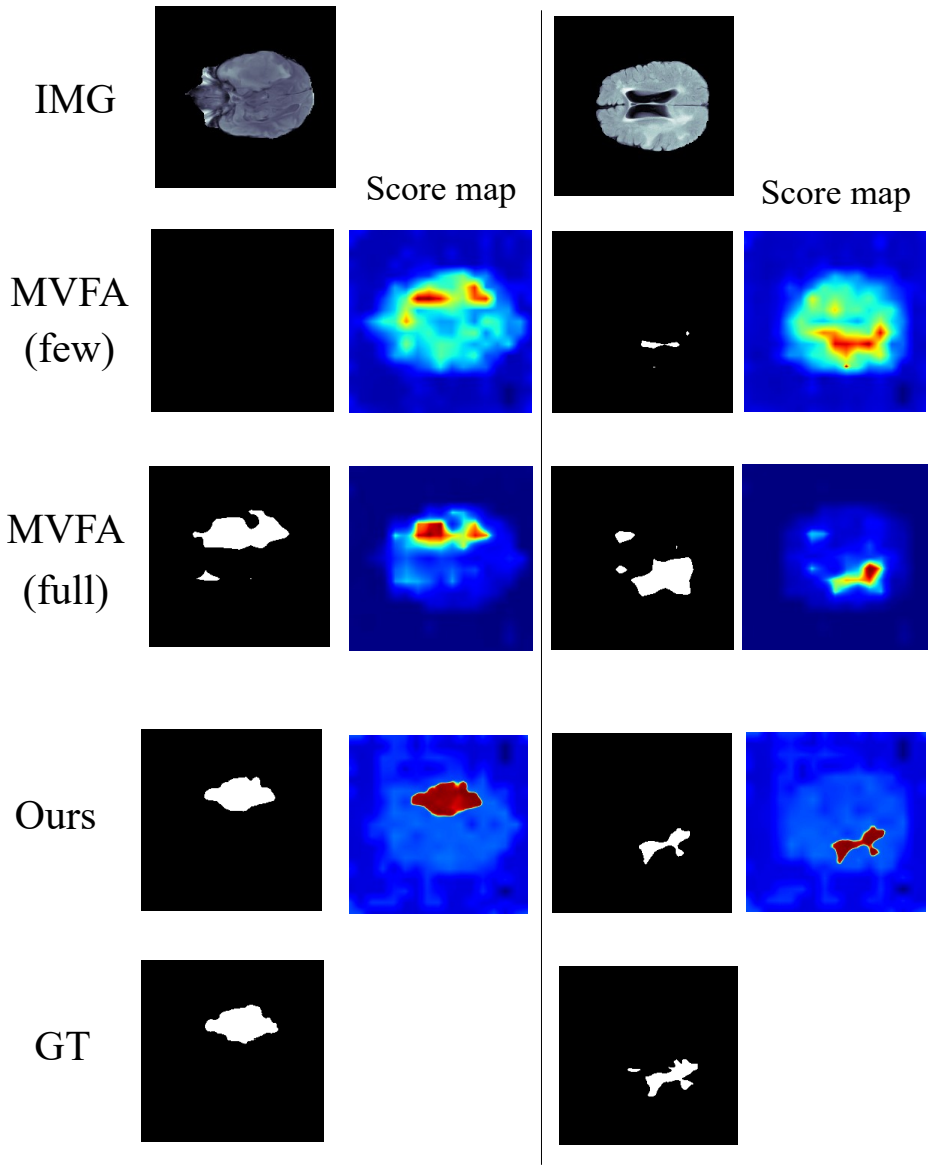

Figure 4. Qualitative results comparing MVFA (few/full) and our method on brain anomaly segmentation, showing both binary masks and score maps.