
SD-MAD: Sign-Driven Few-shot Multi-Anomaly Detection in Medical Images

Kaiyu Guo*

Shanghai Academy of AI for Science
University of Queensland
Brisbane, Australia

Tan Pan*

Fudan University
Shanghai Academy of AI for Science
Shanghai, China

Chen Jiang[†]

Shanghai Academy of AI for Science
Shanghai, China

Zijian Wang

University of Queensland
Brisbane, Australia

Brian C. Lovell

University of Queensland
Brisbane, Australia

Limei Han

Fudan University
Shanghai Academy of AI for Science
Shanghai, China

Yuan Cheng[†]

Fudan University
Shanghai Academy of AI for Science
Shanghai, China

Mahsa Baktashmotlagh

University of Queensland
Brisbane, Australia

Abstract

Medical anomaly detection (AD) is crucial for early clinical intervention, yet it faces challenges due to limited access to high-quality medical imaging data, caused by privacy concerns and data silos. Few-shot learning has emerged as a promising approach to alleviate these limitations by leveraging the large-scale prior knowledge embedded in vision-language models (VLMs). Recent advancements in few-shot medical AD have treated normal and abnormal cases as a one-class classification problem, often overlooking the distinction among multiple anomaly categories. Thus, in this paper, we propose a framework tailored for few-shot medical anomaly detection in the scenario where the identification of multiple anomaly categories is required. To capture the detailed radiological signs of medical anomaly categories, our framework incorporates diverse textual descriptions for each category generated by a Large-Language model, under the assumption that different anomalies in medical images may share common radiological signs in each category. Specifically, we introduce SD-MAD, a two-stage **Sign-Driven** few-shot **Multi-Anomaly Detection** framework: (i) Radiological signs are aligned with anomaly categories by amplifying inter-anomaly discrepancy; (ii) Aligned signs are selected further to mitigate the effect of the under-fitting and uncertain-sample issue caused by limited medical data, employing an automatic sign selection strategy at inference. Moreover, we propose three protocols to comprehensively quantify the performance of multi-anomaly detection. Extensive experiments illustrate the effectiveness of our method.

*Equal contribution. This research was conducted during an internship at the Shanghai Academy of Artificial Intelligence for Science.

[†]Corresponding author

1 Introduction

Medical anomaly detection (AD) has emerged as a critical area of research within the healthcare domain [15]. The detection of anomalies, such as tumors [1] and lesions [12], is essential for prompt clinical intervention. However, access to high-quality medical imaging data remains a significant challenge due to privacy concerns and institutional data silos, thereby highlighting the importance of few-shot learning approaches in medical anomaly detection.

Traditional few-shot anomaly detection [37, 22] often struggles to generalize the model from the limited data to a universal situation because of the limited prior knowledge scale of the model. Recently, many works [23, 9, 17] utilize the large-scale vision-language model (VLM), such as CLIP [35, 40], to help improve the generalization ability of the model in medical anomaly detection. Similar to traditional anomaly detection methods, these approaches identify anomalies by designing a score function that determines whether a given input is normal or abnormal (one-class classification). However, in real-world scenarios, especially in medical imaging, it is crucial to distinguish between different categories of anomalies, as they may correspond to varying pathological conditions and require distinct clinical responses. For example, distinguishing between a lung tumor and pneumonia in chest X-rays is crucial, as they require different treatment approaches: surgery or chemotherapy for cancer [33], and antibiotics for infection [4]. Thus, this paper aims to investigate scenarios involving the presence of diverse anomaly types by few-shot learning. The difference between the existing setting and our work is illustrated in Figure 1(a) and 1(b).

We hypothesize that different anomalies in medical images may share common radiological signs (*e.g.,*) in each category, such as abnormal density or shape, while also exhibiting unique signs that are specific to each anomaly category. These distinct features can provide valuable diagnostic information, enabling more accurate classification and treatment planning. By leveraging both shared and unique patterns, we aim to improve the detection and differentiation of various anomalies in medical imaging. Based on this hypothesis, firstly, we introduce a CLIP-based framework that explicitly (i) **links each anomaly class to a small set of textual “symptom” (signs) descriptions** and measures their similarity to image features. For each anomaly, we enumerate radiologic signs (*e.g.,* “brain with craniotomy defect”, “brain with unclear focal abnormality”) as prompts. As shown in Figure 1(d), aligning visual embeddings with these sign prompts allows the model to learn fine-grained inter-anomaly distinctions. However, recent work [41, 40] reveals that prompt-based alignment in medical vision–language models can be uncertain: not all signs contribute equally, and some may even introduce noise in intra-class matching. To address this, at inference time, we (ii) **automatically select the most informative prompts for each few-shot example** [38], thereby mitigating misleading matches within the same anomaly class. By addressing both inter-anomaly and intra-anomaly challenges, our approach delivers more accurate and reliable multi-anomaly detection under few-shot conditions.

We structure the evaluation protocol for the multi-category medical AD task around three layers to capture the full spectrum of multi-anomaly detection performance: (1) assessing the model’s ability to distinguish between normal and abnormal instances; (2) evaluating the model’s ability to perform multi-label prediction across distinct anomaly types; and (3) assessing the model’s ability to correctly identify the specific types of anomalies. Existing methods face challenges in adapting to the last two protocols, primarily because their scoring functions are not designed to generalize to these task settings.

As summarized below, our contributions are threefold:

1. **Framework for few-shot multi-anomaly detection.** We introduce a few-shot anomaly detector that natively handles multiple anomaly classes within a single model, based on learning the alignment of radiological signs and anomaly categories.
2. **Inter- and intra-anomaly alignment.** We align image embeddings with sets of anomaly-specific prompts during training and, at inference, automatically select the most informative prompts to mitigate the uncertain-sample issue in the vision–language alignment.
3. **Rigorous evaluation protocol.** We assess our approach on seven medical imaging datasets across three evaluation settings, covering both single-class and multi-anomaly scenarios, and demonstrate consistent improvements over state-of-the-art baselines.

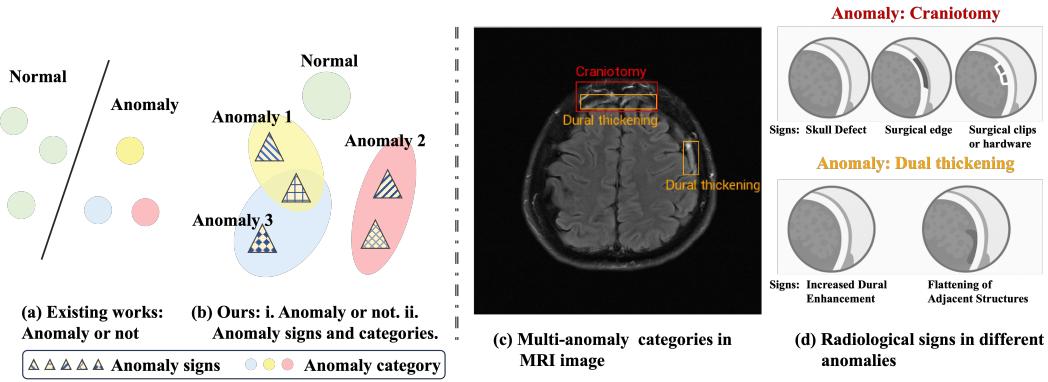


Figure 1: Figures (a) and (b) visualize the difference between our task and previous tasks. Figures (c) and (d) explain a multi-anomaly scenario, and radiological signs of different medical anomalies in the Brain MRI.

2 Related Work

Medical Anomaly detection. Traditional medical anomaly detection methods rely on well-curated anomaly datasets, training on normal images and evaluating on abnormal ones [3, 7, 47, 49, 42, 18, 30, 16]. These approaches model the normal data distribution and identify anomalies as deviations from this distribution, achieving impressive performance. Many of these methods are designed for specific anatomical regions [13, 43] and treat anomaly detection (AD) as a one-class classification problem [3, 7, 26]. However, in real-world scenarios, the same individual may experience multiple diseases affecting the same organ. Recently, the open-set AD method [50] has shifted focus to detecting multiple anomalies instead of relying on one-class classification. These methods require enough training data to formulate the expected distributions, which can be hard to adapt to few-shot setting. To address the challenge of limited large-scale labeled datasets, some approaches have explored few-shot anomaly detection techniques as follows.

Few-shot Anomaly detection. Few-shot anomaly detection has gained significant attention in recent years due to its ability to identify rare or unseen anomalies with limited labeled data. Previous models utilized disentangled representations of anomalies [10] or contrastive learning mechanisms [44] to alleviate the bias, accounting for unseen anomalies. MVFA [24] utilized multi-level adaptation and a contrastive framework to improve generalization across various medical datasets. UniVAD [17] proposed a general framework to detect anomalies across different domains with a training-free unified model. AA-CLIP [31] advanced CLIP model in a two-stage approach to enhance CLIP’s anomaly discrimination ability. Although those methods perform well in various datasets, there is still a lack of few-shot multi-anomaly detection for medical data.

Vision-language model. Vision-language models have demonstrated significant potential across a range of tasks. CLIP [35] excels in image-text alignment and has been successfully applied to various applications, such as classification and text-image retrieval. To expand CLIP’s capabilities to medical data, MedCLIP [40] was introduced as a foundation for medical image-text alignment. Based on those pre-trained foundation models, recent studies [21, 27, 8] in anomaly detection have leveraged pre-trained CLIP models for language-guided anomaly detection and segmentation, achieving impressive results and highlighting the promising potential of these models in this domain.

3 Methodology

In this section, we first formulate the problem of few-shot anomaly detection and few-shot multi-anomaly detection in medical images. Then we propose our methods within two parts: In section 3.2, we propose a training method with a tailored adapter for vision-language models and an inter-anomaly representation learning loss function; In Section 3.3, we propose an inference strategy to filter the outlier prompts, which aims to handle the intra-anomaly uncertain samples. Figure 2 shows the overall pipeline of our model.

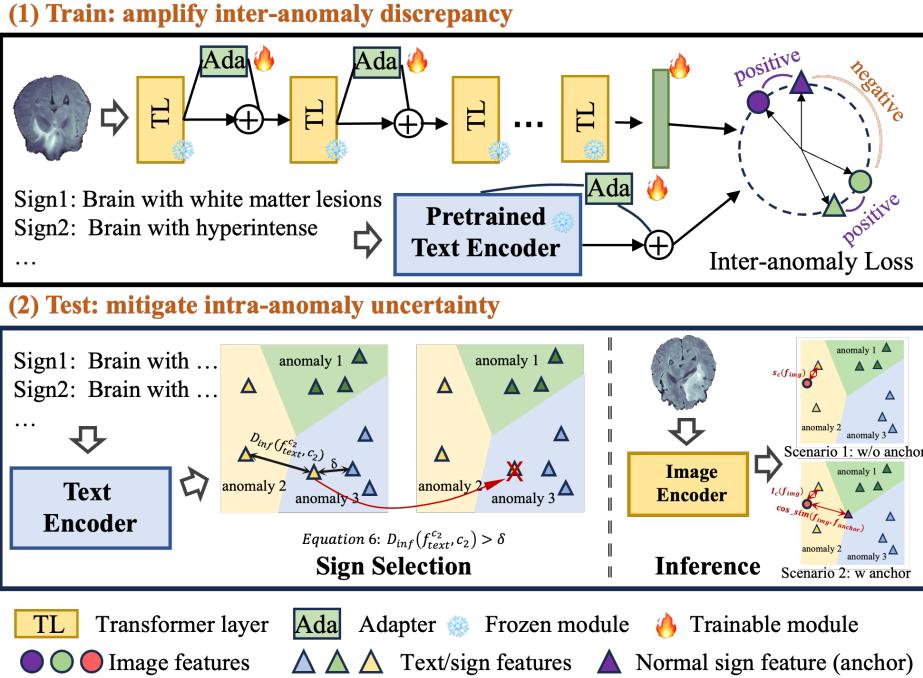


Figure 2: The pipeline of SD-MAD. In the framework, the training phase is designed to amplify inter-anomaly discrepancies, and the inference stage aims to handle the uncertain-sample problem in each anomaly category.

3.1 Problem Formulation

Few-shot medical anomaly detection: Following the setting of previous work [23] on few-shot medical anomaly detection, the few-shot training samples can be presented as $\mathcal{D}_{few} = \{(x_i, c_i, s_i)\}_i^K$,

where K is the number of samples, x_i is the i -th image, the corresponding image-level label $c_i \in \{0, 1\}$, and the pixel-level label $s_i \in \{0, 1\}^{h \times w}$ is a binary mask with the same size $h \times w$ as the image x_i . For a given test image x_{test} , image-level and pixel-level medical anomaly detection are evaluated with the corresponding image labels c_{test} and pixel labels s_{test} .

Few-shot medical anomaly detection with multiple anomaly categories: Similar to the setting of few-shot medical anomaly detection, few-shot training samples can be presented as $\hat{\mathcal{D}}_{few} = \{(x_i, \mathbf{c}_i)\}_i^K$, where $\mathbf{c}_i \in \{0, 1\}^d$ is a d -dimensional label. Since it is hard to access the pixel-level labels for the multi-anomaly medical datasets, we do not consider the pixel-level label in this setting. Thus, given a test image x_{test} , only image-level medical anomaly detection is evaluated with the corresponding image labels \mathbf{c}_{test} in the scenarios where multiple anomaly categories exist.

3.2 Training: Amplify Inter-anomaly Discrepancy

Shift Adapter. To preserve the large-scale prior knowledge encoded in CLIP, we propose a shift adapter designed to effectively aggregate learning signals from few-shot samples while retaining the original prior information. The shift adapter is used for both image and text encoders, which is shown as our pipeline in Figure 2.

Considering the feature \hat{f}_i^{in} is input of the adapter, which is also the output of the i -th transformer layer, the output of the adapter at the i -th transformer layer is

$$\hat{f}_i^{ada} = \alpha(W_i^2 \alpha(W_i^1 \hat{f}_i^{in})), \quad (1)$$

where W_i^1 and W_i^2 are trainable linear weights of the adapter at the i -th transformer layer, α is the activation function.

Inspired from residual learning methods [19], we integrate the output of the original transformer layer \hat{f}_i^{out} with \hat{f}_i^{ada} by inner interpolation as follows:

$$f_i^{out} = \lambda \hat{f}_i^{out} + (1 - \lambda) \hat{f}_i^{ada}, \quad (2)$$

where λ is the hyperparameter to control the interpolation ratio. To avoid the overfitting caused by the limited number of few-shot samples, we restrict the application of the adapter to four layers in the image encoder and one layer in the text encoder.

Inter-anomaly Loss. The text-vision alignment in CLIP depends on this contrastive learning insight with the cosine similarity. From the view of contrastive learning [34, 36], the distance of the positive image-text pairs should be smaller than the distance of negative text-image pairs. Towards this end, existing work directly minimize the cosine distance between the positive image-text pairs to align the text and image features as follows:

$$L_{img-text} = \min_{\theta} \sum_{i \in [1, N_c]} d(f_{img}^c, f_{text,i}^c). \quad (3)$$

Here, $d(\cdot, \cdot)$ denotes the cosine distance between two input vectors, θ is the trainable parameters, image feature f_{img}^c and detailed-description text features $f_{text,i}^c$ belong to anomaly category $c \in \mathcal{C}$ of the given image, N_c is the number of text prompts corresponding to category c .

It is important to note that Equation 3 does not account for the distances of negative pairs. This is because simply increasing the distance between negative pairs provides limited utility in enabling the model to accurately identify the anomaly categories. For instance, given an abnormal image exhibiting only the anomaly of a lesion, the prediction may still fail despite strong alignment of positive pairs, as the model may erroneously assign high similarity scores to irrelevant categories, resulting in false positives. To handle this issue, we introduce an anchor feature f_{anchor} that serves to define the boundary between normal and abnormal images. Thus, the following relationship should be satisfied.

Remark 3.1 Given an image feature belonging to category c , we have

$$\sup_{i \in [1, N_c]} d(f_{img}^c, f_{text,i}^c) \leq d(f_{img}^c, f_{anchor}) \leq \inf_{k \neq c, j \in [1, N_k]} d(f_{img}^c, f_{text,j}^k)$$

Given the image feature f_{img}^c and category c , Remark 3.1 indicates that f_{anchor} serves as the hyperplane to separate the subspace of category c and other categories. To distinguish the difference between the normal category and other anomalies simultaneously, we set the f_{anchor} as the feature of the text prompt corresponding to normal images. According to Remark 3.1, we propose the following loss:

$$\begin{aligned} \hat{d}_{positive,i}^c &= \max(0, d(f_{img}^c, f_{text,i}^c) - d(f_{img}^c, f_{anchor})) \\ \hat{d}_{negative,j}^{c,k} &= \max(0, d(f_{img}^c, f_{text,j}^k) - d(f_{img}^c, f_{text,j}^k)) \\ L_{anchor} &= \min_{\theta} \sum_{i \in [1, N_c]} \hat{d}_{positive,i}^c + \sum_{k \neq c, j \in [1, N_k]} \hat{d}_{negative,j}^{c,k} \end{aligned} \quad (4)$$

As discussed above, the overall loss for amplifying the inter-anomaly discrepancy is

$$L = L_{img-text} + L_{anchor} \quad (5)$$

3.3 Inference: Mitigate Intra-anomaly Uncertain-sample Issue

During the inference stage, image features from the test set are evaluated against the text prompt features corresponding to each anomaly category. However, the limited number of few-shot training samples, combined with uncertainty in medical vision-language [41], may cause under-fitted features that fail to capture anomaly characteristic-specific information. Thus, to address this issue, we divided our inference stage into two parts as follows.

Sign Selection. As we discussed above, each anomaly category contains several prompt features corresponding to the anomaly signs. Thus, there should be a labeling function $h_{text}(\cdot)$ satisfied $h_{text}(f_{text}^c) = c$. Therefore, we have the definition of the distance between given text feature f_{text} and category c in the following.

Definition 3.2 Given a text feature f_{text} , the distance between the prompt feature f'_{text} and the decision region of the anomaly category c is

$$D_{inf}(f_{text}, c) = \inf_{\{f'_{text} | h(f'_{text}) = c, f'_{text} \neq f_{text}\}} d(f'_{text}, f_{text})$$

Definition 3.2 provides a definition of distance between the prompt feature f'_{text} and the decision region $\{f'_{text} | h(f'_{text}) = c\}$. For the ideal situation, we have the following relation.

Remark 3.3 Given a text feature f_{text}^c belonging to category c , we have

$$D_{inf}(f_{text}^c, c) < \delta$$

Where $\delta \triangleq \inf_{k \neq c, k \in C} D_{inf}(f_{text}^c, k)$

As shown in Figure 2, the outlier text features in each category may break the relation in Remark 3.3. However, in the inference time, the text features are fixed. Thus, we propose to modify the labeling function $h(\cdot)$ to mitigate this problem.

Given a text feature f_{text}^c which satisfies $h(f_{text}^c) = c$, the new labeling function is defined as

$$h_{new}(f_{text}^c) = \begin{cases} c & \text{if } D_{inf}(f_{text}^c, c) < \delta \\ -1 & \text{else} \end{cases} \quad (6)$$

The Equation 6 indicates that the new labeling function $h_{new}(\cdot)$ discards the distorted features that break the relation in Remark 3.3 for each anomaly category. The sign selection process can be viewed in Fig. 2 (2). This labeling function is used for the score function design, which we will discuss in the following.

Inference. Unlike previous methods [23, 25], which focus solely on evaluating the Area Under the Receiver Operating Characteristic curve (AUROC) using a continuous scoring function, we additionally consider scenarios that require binary predictions. For the binary prediction, the anchor feature is required for the evaluation.

Scenario 1: Continuous scoring function without anchor feature. Without anchor feature, for the given category c and the image feature f_{img} , the score function corresponding to c is

$$s_c(f_{img}) = \sup_{h_{new}(f_{text})=c} \text{cosine_similarity}(f_{img}, f_{text}). \quad (7)$$

As we discussed in Section 3.1, the label of image x is a vector \mathbf{c} . Thus, the score vector corresponding to \mathbf{c} is $\mathbf{s}_c = \{s_{c_i}(f_{img})\}_{i=1}^K$, where K is the number of anomaly categories.

Scenario 2: Binary prediction with anchor feature. With the anchor feature, we can achieve the binary prediction for each anomaly category. The prediction p_c for category c with a give image feature f_{img} is

$$p_c(f_{img}) = \begin{cases} 1 & \text{if } I_c(f_{img}) > \text{cosine_similarity}(f_{img}, f_{anchor}) \\ 0 & \text{else} \end{cases} \quad (8)$$

, where $I_c(f_{img}) \triangleq \inf_{h_{new}(f_{text})=c} \text{cosine_sim}(f_{img}, f_{text})$. The precision vector corresponding to \mathbf{c} is $\mathbf{p}_c = \{p_{c_i}(f_{img})\}_{i=1}^K$. This prediction can be used for the evaluation with the Hamming score and the subset accuracy score.

4 Experiments

4.1 Experimental Setup

Evaluation protocols We introduce three evaluation protocols: 1) for general anomaly detection, following previous works [23, 25], we quantify the performance with area under the receiver operating curve (AUROC) metric on image- and pixel-level; 2) We introduce Hamming score and subset accuracy to evaluate the performance on multi-label prediction on the task of multi-anomaly detection; 3) We exploit the AUROC metric for each class to evaluate the performance on the specific types of

Table 1: Comparison on general anomaly detection. "Avg." is short for "average".

	Dataset	DRA	BGAD	MVFA	Ours
Img-level (AUROC(%))	BrainMRI	80.6	83.6	92.4	91.4
	LiverCT	59.6	72.5	81.2	86.9
	RESC	90.9	86.2	96.2	95.2
	HIS	68.7	-	82.7	81.6
	ChestXRay	75.8	-	82.0	82.7
Pixel-level (AUROC(%))	OCT	99.0	-	99.4	99.8
	BrainMRI	74.8	92.7	97.3	96.5
	LiverCT	71.8	98.9	99.7	99.5
	RESC	77.3	93.8	99.0	99.0
Avg.		77.6	88.0	92.2	92.5

anomalies. Specifically, given the anomaly type c , the binary label is set as 1 for the images belonging to type c , and 0 for the others.

Dataset We evaluate the methods with 7 datasets. For general medical anomaly detection, we follow the BMAD benchmark [3], which includes 6 datasets: Brain MRI [1, 2, 32], Liver CT [6, 29], retinal OCT [28, 20], Chest X-ray [39], and Digital Histopathology [5]. Among these datasets, both image- and pixel-level metrics are evaluated for BrainMRI [1, 2, 32], LiverCT [6, 29], and RESC [20]. For the other datasets, namely OCT17 [28], ChestXray [39] and HIS [5], only image-level scores are evaluated.

The experiments for multi-anomaly detection are built from the brain MRI dataset in fastMRI+ [48, 46]. We select 6 anomaly categories and the same slice-level images, namely slice 0, 5 and 10, for the multi-anomaly detection tasks. More details can be viewed in the Appendix.

Training details We select CLIP with ViT-L/14 [14] as the backbone model with the size of input as 240×240 . We employ our shift adapter to the 6-, 8-, 18- and 24-th layers in the transformer of the CLIP image encoder and to the last layer to the transformer of the CLIP text encoder. Every training process is conducted in 50 epochs. The training process requires 4000 Mib GPU memory for the model. The experiments are conducted on an A100 GPU.

4.2 General Few-shot Medical Anomaly Detection

We first evaluate our method under the setting of general few-shot anomaly detection. We conduct the 4-shot experiments with state-of-the-art few-shot medical anomaly detection methods, MVFA [23], and other few-shot anomaly detection methods, namely BRA [11] and BGAD [45]. To adapt our method to pixel-level score, we combine our method with MVFA. Specifically, we aggregate our inter-anomaly loss with the losses of MVFA.

As Table 1 shows, even though our methods are not designed for the general few-shot anomaly detection, we still average outperform other methods. In addition, for LiverCT, our method can significantly improve the anomaly detection performance.

4.3 Multi-category Few-shot Medical Anomaly detection

As discussed above, we introduce two evaluation protocols for multi-category few-shot medical anomaly detection: evaluating the model’s ability to perform multi-label prediction across distinct anomaly types and assessing the model’s ability to correctly identify the specific types of anomalies. As previous few-shot anomaly detection methods can not handle the multi-category scenarios, we only compare the baseline model CLIP [35] and the vision-language model tailored for medical image MedCLIP [40]. In the following, we present the performance of our experiments in the two settings.

4.3.1 Multi-label Prediction

For multi-label prediction, we utilize two evaluation metrics to quantify the performance, namely Hamming score and subset accuracy. We provide the details of the two evaluation metrics in the appendix.

Table 2: The 1-shot results of the experiments on multi-label prediction. The evaluation metrics are Hamming score and subset accuracy. "SS" is short for "Sign Selection"

		Clip	MedClip	Ours (no SS)	Ours (full model)
slice 0	Hamming(%)↑	80.2	77.1	85.8	87.2
	Subset acc.(%)↑	0.4	18.5	34.6	60.8
slice 5	Hamming(%)↑	77.6	63.5	72.7	76.5
	Subset acc.(%)	0	0	29.0	27.3
slice10	Hamming(%)↑	78.3	73.2	73.8	79.2
	Subset acc.(%)↑	0	1.9	19.8	21.7

Table 2 shows the 1-shot results with the two evaluation metrics. As shown in the table, our full model demonstrates superior performance compared to the other methods. Also, the comparison between the methods with and without sign selection illustrates the effectiveness of our inference strategy. From the table, we can tell that the pretrained model with vanilla training process can solve the multi-label prediction task for the medical anomaly detection. The comparison between our method without sign selection and vanilla CLIP illustrates the effectiveness of our training process. In addition, the comparison of the performance with and without sign selection also demonstrates the validity of our inference stage.

4.3.2 Category-wise AUROC

To evaluate the ability to recognize the specific anomaly type, we introduce the category-wise AUROC metric. This metric indicates the performance of the model in each category. Since the label *Small vessel chronic white matter ischemic change* can not be achieved in slices 5 and 10, we only evaluate 5 categories within these two slices.

As Table 3 shows, our methods significantly improve the average performance for every slice. Sign selection may not be able to improve category-wise performance. We assume that the reason is that the outlier prompt may fit some images in the test samples. Even if during the training stage, the few-shot samples may underfit these outlier prompts, the prior information may still fit them well to the corresponding anomalies. Thus, the sign selection may not perform well for category-wise settings. The performance on *Enlarged ventricles* can illustrate this issue clearly.

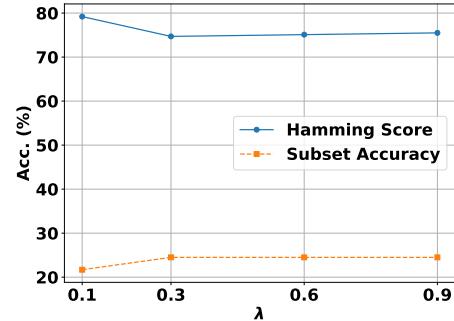


Figure 3: The ablation study on λ . We conduct the experiments on the multi-label prediction task with two metrics, namely Hamming score and Subset accuracy.

4.4 Visualization of Image-Text Similarity

To evaluate the alignment of our method, we visualize the alignment in Figure 4. As the figure shows, our training method promotes the alignment between abnormal images and corresponding prompts. However, some prompts may exhibit overconfidence in a false category. For instance, in Figure 4b, the characteristic *surgical scarring* has equally high similarities between both *Craniotomy* and *Posttreatment change* images. This phenomenon may result in uncertain samples for the prediction, which leads us to propose the sign selection method.

Table 3: The 1-shot results of the experiments on category-wise AUROC. The reported results are AUROC score (%). We also report average (Avg.) results for each slice. "Small vessel ischemic change" corresponds to the label " Small vessel chronic white matter ischemic change" in the FastMRI+ dataset. "SS" and "Avg." are short for "Sign Selection" and "average" respectively.

		CLIP	MedCLIP	Ours(no SS)	Ours(full model)
slice 0	Craniotomy	42.7	50.0	68.2	70.9
	Posttreatment change	73.5	51.1	67.3	71.7
	Nonspecific lesion	56.8	44.3	65.1	56.7
	Dural thickening	44.6	48.5	58.9	57.5
	Enlarged ventricles	65.3	68.7	62.5	71.9
	Small vessel ischemic change	62.1	39.4	81.7	80.3
Avg.		57.5	50.3	67.3	68.2
slice 5	Craniotomy	67.2	46.9	55.0	55.4
	Posttreatment change	59.4	51.7	63.9	62.4
	Nonspecific lesion	47.9	44.2	64.9	64.9
	Dural thickening	51.1	63.9	64.5	57.4
	Enlarged ventricles	76.1	51.3	66.6	79.3
	Avg.	60.3	51.6	63.0	63.9
slice 10	Craniotomy	48.3	43.0	51.6	62.4
	Posttreatment change	61.1	58.2	40.6	46.6
	Nonspecific lesion	37.5	45.6	71.3	58.0
	Dural thickening	57.7	48.8	72.2	69.9
	Enlarged ventricles	98.1	40.1	100	70.5
	Avg.	60.5	47.1	67.1	61.5

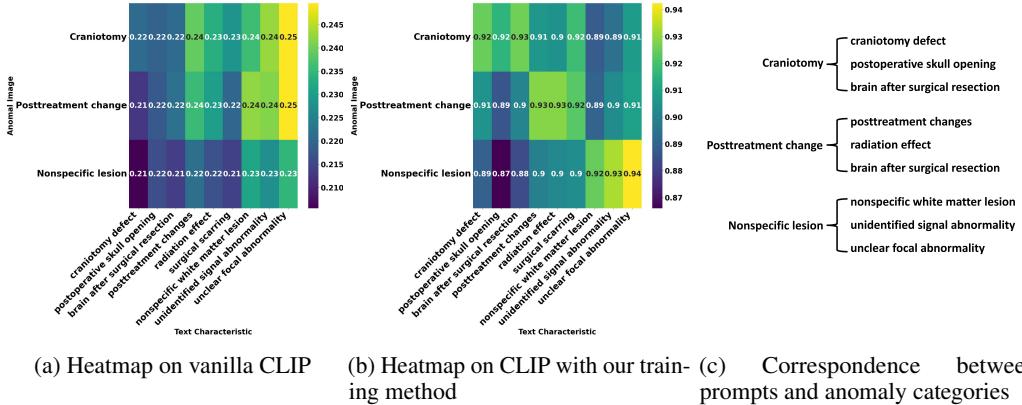


Figure 4: Visualization of image-text similarity heatmaps. (a) visualizes the heatmap on vanilla CLIP, (b) visualizes the heatmap on our trained model. The correspondence between prompts and anomaly categories is provided in (c).

4.5 Ablation study on λ

To evaluate the effect of the hyperparameter in the Shift Adapter, we conducted ablation studies for λ on the multi-label prediction with the 5th slice. Figure 3 shows the experimental results. The figure does not exhibit significant impacts on the performance when λ changes, which shows the robustness of the learnable adapter. In addition, we find that there is a trade-off between Hamming score and Subset accuracy when λ increases. We assume that it is because the increase of λ may cause slight overfitting to the few-shot samples. Therefore, the model may produce fewer predictions in the presence of intra-class variation within the same anomaly type. While this may lead to a reduction in Hamming score, it could potentially enhance the overall prediction accuracy.

5 Conclusion and Limitation

In this paper, we introduce a novel setting for medical anomaly detection, termed multi-anomaly detection. Unlike previous settings that typically assume a single anomaly per image, multi-anomaly detection is designed to address scenarios where multiple anomalies co-exist within the same clinical image. Building on this new task, we propose a method based on a vision-language model (VLM) for both inter- and intra-anomaly alignment. Specifically, we propose an inter-anomaly loss to amplify the inter-anomaly discrepancy and update the CLIP model with trainable Shift Adapters. In addition, we design a sign selection method to mitigate the intra-anomaly uncertainty at the inference stage. To thoroughly evaluate the performance of our method in the task of multi-anomaly detection, besides the general setting in anomaly detection, we propose two more evaluation protocols, namely multi-label prediction and category-wise AUROC. The extensive experiments illustrate the effectiveness of our method.

Limitation Even if our proposed method can effectively address the multi-anomaly detection task, there are still limitations, which mainly rely on the correspondence between the prompt and the anomaly categories. Some prompts may correspond to more than one anomaly type, which may result in false predictions if we ignore this nature. Addressing this ambiguity in prompt-anomaly correspondence will be the focus of our future work.

References

- [1] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F. C. Kitamura, S. Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021.
- [2] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.
- [3] J. Bao, H. Sun, H. Deng, Y. He, Z. Zhang, and X. Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053, 2024.
- [4] M. Bassetti, F. Magnè, D. R. Giacobbe, L. Bini, and A. Vena. New antibiotics for gram-negative pneumonia. *European Respiratory Review*, 31(166), 2022.
- [5] B. E. Bejnordi, M. Veta, P. J. Van Diest, B. Van Ginneken, N. Karssemeijer, G. Litjens, J. A. Van Der Laak, M. Hermsen, Q. F. Manson, M. Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Jama*, 318(22):2199–2210, 2017.
- [6] P. Bilic, P. Christ, H. B. Li, E. Vorontsov, A. Ben-Cohen, G. Kaassis, A. Szeskin, C. Jacobs, G. E. H. Mamani, G. Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023.
- [7] Y. Cai, H. Chen, X. Yang, Y. Zhou, and K.-T. Cheng. Dual-distribution discrepancy with self-supervised refinement for anomaly detection in medical images. *Medical image analysis*, 86:102794, 2023.
- [8] Y. Cao, X. Xu, Y. Cheng, C. Sun, Z. Du, L. Gao, and W. Shen. Personalizing vision-language models with hybrid prompts for zero-shot anomaly detection. *IEEE Transactions on Cybernetics*, 2025.
- [9] Y. Cao, J. Zhang, L. Frittoli, Y. Cheng, W. Shen, and G. Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. In *European Conference on Computer Vision*, 2024.
- [10] C. Ding, G. Pang, and C. Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7388–7398, 2022.

- [11] C. Ding, G. Pang, and C. Shen. Catching both gray and black swans: Open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7388–7398, 2022.
- [12] Z. Ding, Q. Dong, H. Xu, C. Li, X. Ding, and Y. Huang. Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In *International Conference on Neural Information Processing*, pages 133–144. Springer, 2022.
- [13] Z. Ding, Q. Dong, H. Xu, C. Li, X. Ding, and Y. Huang. Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In *International Conference on Neural Information Processing*, pages 133–144. Springer, 2022.
- [14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
- [15] T. Fernando, H. Gammulle, S. Denman, S. Sridharan, and C. Fookes. Deep learning for medical anomaly detection—a survey. *ACM Computing Surveys (CSUR)*, 54(7):1–37, 2021.
- [16] M. S. Graham, P.-D. Tudosi, P. Wright, W. H. L. Pinaya, P. Teikari, A. Patel, J.-M. U-King-Im, Y. H. Mah, J. T. Teo, H. R. Jäger, et al. Latent transformer models for out-of-distribution detection. *Medical Image Analysis*, 90:102967, 2023.
- [17] Z. Gu, B. Zhu, G. Zhu, Y. Chen, M. Tang, and J. Wang. Univad: A training-free unified model for few-shot visual anomaly detection. *arXiv preprint arXiv:2412.03342*, 2024.
- [18] R. Hassanaly, C. Brianceau, M. Solal, O. Colliot, and N. Burgos. Evaluation of pseudo-healthy image reconstruction for anomaly detection with deep generative models: Application to brain fdg pet. *arXiv preprint arXiv:2401.16363*, 2024.
- [19] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [20] J. Hu, Y. Chen, and Z. Yi. Automated segmentation of macular edema in oct using deep neural networks. *Medical image analysis*, 55:216–227, 2019.
- [21] L. Hua, X. Su, Y. Luo, S. You, and J. Long. Hieclip: Hierarchical clip with explicit alignment for zero-shot anomaly detection. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.
- [22] C. Huang, H. Guan, A. Jiang, Y. Zhang, M. Spratling, and Y.-F. Wang. Registration based few-shot anomaly detection. In *European conference on computer vision*, pages 303–319. Springer, 2022.
- [23] C. Huang, A. Jiang, J. Feng, Y. Zhang, X. Wang, and Y. Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024.
- [24] C. Huang, A. Jiang, J. Feng, Y. Zhang, X. Wang, and Y. Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11375–11385, 2024.
- [25] J. Jeong, Y. Zou, T. Kim, D. Zhang, A. Ravichandran, and O. Dabeer. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19606–19616, 2023.
- [26] A. Jiang, C. Huang, Q. Cao, S. Wu, Z. Zeng, K. Chen, Y. Zhang, and Y. Wang. Multi-scale cross-restoration framework for electrocardiogram anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 87–97. Springer, 2023.
- [27] E. Jin, Q. Feng, Y. Mou, G. Lakemeyer, S. Decker, O. Simons, and J. Stegmaier. Logicad: Explainable anomaly detection via vlm-based text feature extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4129–4137, 2025.

- [28] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *cell*, 172(5):1122–1131, 2018.
- [29] B. Landman, Z. Xu, J. Igelsias, M. Styner, T. Langerak, and A. Klein. Miccai multi-atlas labeling beyond the cranial vault—workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, page 12. Munich, Germany, 2015.
- [30] J. Linmans, G. Raya, J. van der Laak, and G. Litjens. Diffusion models for out-of-distribution detection in digital pathology. *Medical Image Analysis*, 93:103088, 2024.
- [31] W. Ma, X. Zhang, Q. Yao, F. Tang, C. Wu, Y. Li, R. Yan, Z. Jiang, and S. K. Zhou. Aa-clip: Enhancing zero-shot anomaly detection via anomaly-aware clip. *arXiv preprint arXiv:2503.06661*, 2025.
- [32] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
- [33] F. Montagne, F. Guisier, N. Venissac, and J.-M. Baste. The role of surgery in lung cancer treatment: present indications and future perspectives—state of the art. *Cancers*, 13(15):3711, 2021.
- [34] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.
- [35] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.
- [36] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.
- [37] S. Sheynin, S. Benaim, and L. Wolf. A hierarchical transformation-discriminating generative model for few shot anomaly detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8495–8504, 2021.
- [38] K. Shum, S. Diao, and T. Zhang. Automatic prompt augmentation and selection with chain-of-thought from labeled data. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12113–12139, Singapore, Dec. 2023. Association for Computational Linguistics.
- [39] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. Summers. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, volume 7, page 46. sn, 2017.
- [40] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2022, page 3876, 2022.
- [41] P. Xia, Z. Chen, J. Tian, G. Yangrui, R. Hou, Y. Xu, Z. Wu, Z. Fan, Y. Zhou, K. Zhu, W. Zheng, Z. Wang, X. Wang, X. Zhang, C. Bansal, M. Niethammer, J. Huang, H. Zhu, Y. Li, J. Sun, Z. Ge, G. Li, J. Zou, and H. Yao. CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.
- [42] T. Xiang, Y. Zhang, Y. Lu, A. L. Yuille, C. Zhang, W. Cai, and Z. Zhou. Squid: Deep feature in-painting for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23890–23901, 2023.
- [43] H. Xu, Y. Zhang, X. Chen, C. Jing, L. Sun, Y. Huang, and X. Ding. Afsc: Adaptive fourier space compression for anomaly detection. *IEEE Transactions on Industrial Informatics*, 2024.

- [44] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24490–24499, 2023.
- [45] X. Yao, R. Li, J. Zhang, J. Sun, and C. Zhang. Explicit boundary guided semi-push-pull contrastive learning for supervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24490–24499, 2023.
- [46] J. Zbontar, F. Knoll, A. Sriram, T. Murrell, Z. Huang, M. J. Muckley, A. Defazio, R. Stern, P. Johnson, M. Bruno, M. Parente, K. J. Geras, J. Katsnelson, H. Chandarana, Z. Zhang, M. Drozdal, A. Romero, M. Rabbat, P. Vincent, N. Yakubova, J. Pinkerton, D. Wang, E. Owens, C. L. Zitnick, M. P. Recht, D. K. Sodickson, and Y. W. Lui. fastMRI: An open dataset and benchmarks for accelerated MRI, 2018.
- [47] J. Zhang, Y. Xie, G. Pang, Z. Liao, J. Verjans, W. Li, Z. Sun, J. He, Y. Li, C. Shen, et al. Viral pneumonia screening on chest x-rays using confidence-aware anomaly detection. *IEEE transactions on medical imaging*, 40(3):879–890, 2020.
- [48] R. Zhao, B. Yaman, Y. Zhang, R. Stewart, A. Dixon, F. Knoll, Z. Huang, Y. W. Lui, M. S. Hansen, and M. P. Lungren. fastmri+, clinical pathology annotations for knee and brain fully sampled magnetic resonance imaging data. *Scientific Data*, 9(1):152, 2022.
- [49] K. Zhou, J. Li, W. Luo, Z. Li, J. Yang, H. Fu, J. Cheng, J. Liu, and S. Gao. Proxy-bridged image reconstruction network for anomaly detection in medical images. *IEEE Transactions on Medical Imaging*, 41(3):582–594, 2021.
- [50] J. Zhu, C. Ding, Y. Tian, and G. Pang. Anomaly heterogeneity learning for open-set supervised anomaly detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17616–17626, 2024.