MedVMAD: Medical imaging Anomaly Detection using Visual Language Models

Lucy Bodtman
University of Massachusetts Amherst
lbodtman@umass.edu

Deepti Guntur
University of Massachusetts Amherst
dguntur@umass.edu

Abstract

Accurate and efficient anomaly detection in medical imaging is essential to improve diagnostic reliability and reduce human error in healthcare settings. Identifying subtle abnormalities across various imaging modalities, such as MRI, CT, and X-rays, remains a significant challenge, often resulting in critical details being overlooked. We aim to develop a robust framework that enhances diagnostic precision using a visual language model such as CLIP [8]. Leveraging CLIP's visual-language alignment capabilities, our approach focuses on detecting minor abnormalities that may escape human observation. In this project, we will also try zero-shot anomaly detection. We will be utilizing learnable text prompts that broadly categorize images as 'normal' or 'abnormal' and learnable image feature token embeddings. During testing, comparison is performed between the adapted visual features and text prompt features, enabling the generation of multi-level anomaly score maps. This multi-level comparison enables the generation of comprehensive anomaly score maps, facilitating precise anomaly localization and classification across diverse medical imaging contexts.

https://github.com/deeptiguntur/MedVMAD

1. Introduction

In this project, we will explore the use of visual-language models, specifically CLIP, for medical anomaly detection.

1.1. Motivation

Medical imaging demands precise anomaly detection, as even the slightest oversight can lead to significant consequences in diagnosis and treatment. Traditional deep learning models often rely on domain-specific training data and struggle to generalize across varying imaging modalities, such as MRI, CT, and X-rays. Moreover, medical anomalies are inherently diverse, with no universal patterns, which makes anomaly detection a particularly challenging task.

This is where CLIP (Contrastive Language-Image Pretraining) provides a promising foundation. CLIP's ability to align visual and textual information enables the use of natural language descriptions to guide image understanding, making it adaptable to a wide range of medical scenarios without requiring extensive retraining. Leveraging CLIP allows us to harness its pre-trained knowledge and focus on improving anomaly detection by adapting it to the specific nuances of medical imaging.

To enhance CLIP's performance in this domain, we use both learnable text prompts and learnable image feature token embeddings.

- **Learnable Text Prompts:** Medical anomalies often require nuanced categorizations. Instead of relying solely on fixed textual descriptions, we incorporate learnable text prompts that can adapt to better describe ”normal” and ”abnormal” images based on the dataset and context. This flexibility allows the textual embeddings to align more closely with the visual features, improving classification and localization.
- **Learnable Image Feature Token Embeddings:** While CLIP's original visual representations are powerful, they are not tailored for fine-grained anomaly detection. By introducing learnable adapters, we refine the image feature tokens to better capture subtle variations and pixel-level abnormalities specific to medical imaging, enabling the model to detect even the smallest anomalies effectively.

By combining these two learnable components, our approach bridges the gap between general-purpose models like CLIP and the specialized demands of medical anomaly detection. This dual adaptation ensures both the visual and textual representations are fine-tuned for the task, resulting in a robust framework that enhances diagnostic precision, supports zero-shot detection, and generalizes well across diverse medical contexts.

1.2. Visual-Language Models

Vision-language models (VLMs) combine a text encoder with an image encoder, with each encoder processing its own media type to yield an embedding vector. These models are trained on image-description pairs, allowing them to learn and understand the relationship between words and images, as well as analyze and recognize visual concepts.

In the domain of medical Anomaly Detection (AD), this method could allow VLMs like CLIP to identify problems in medical images by comparing minor visual cues to predetermined textual categories, such as ”normal” or ”abnormal”. This method enables the model to assess and discover abnormalities by comparing adaptive visual features to descriptive text features, making it a useful tool in cases involving limited labeled data.
This approach leverages annotated training data to generate learnable visual patch feature embedding for classification and segmentation along with learnable text embedding to compute the similarity score using ground-truth, thereby enhancing the accuracy of comparisons between visual and textual inputs.

1.3. Dataset

The dataset used in BMAD [1] (BRaTS2021) 1 will be used as our training set for our model. This dataset is specifically designed for anomaly detection and localization in medical images and consists of various MRI scans. While the original paper encompasses six different medical datasets, our focus will be exclusively on the Brain MRI scans for training. For our training set, we have used 198 ”normal” images and 233 ”abnormal” images, a fraction of the original dataset. Our training set size is much smaller than our baseline models. For our testing set, we have used 83 images total (39 ”normal” and 44 ”abnormal”).

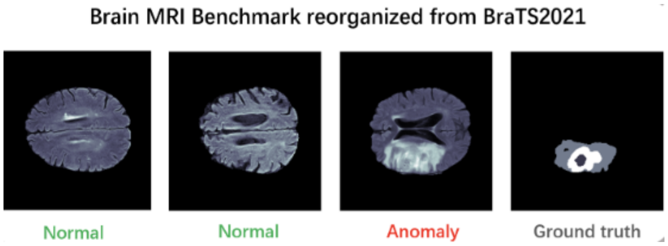


Figure 1. BRaTS 2021 dataset

The BMAD dataset provides valuable insights into our model’s performance and is used as a benchmark. We also use liver CTs Fig.2 from the BMAD dataset to view our model’s performance with zero-shot detection. Notably, anomalies in livers tend to be darker in the scans as opposed to brain MRI anomalies, which tend to be brighter. This dataset also presents a challenge as our model is only trained on brain MRI scans.

The Breast Cancer Screening 3 – Digital Breast Tomosynthesis (DBT)[3] is a curated dataset of images from 5,060 subjects, categorized into 4 groups: normal, actionable, biopsy-proven benign, and biopsy-proven cancer cases. This dataset presents a challenging yet realistic benchmark for future innovations of models aimed at detecting anomalies in DBT volumes. We tested on this dataset with normal and biopsy-proven cancer cases to see how well our model performs on unseen anomalies for zero-shot.

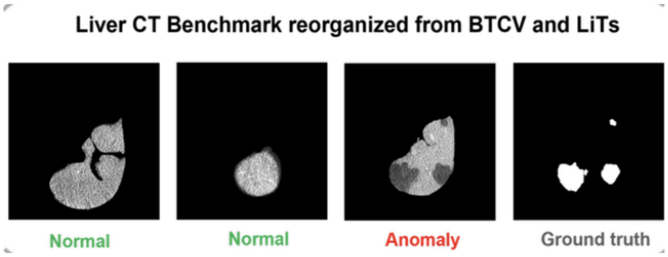


Figure 2. BCTV + LiTs dataset

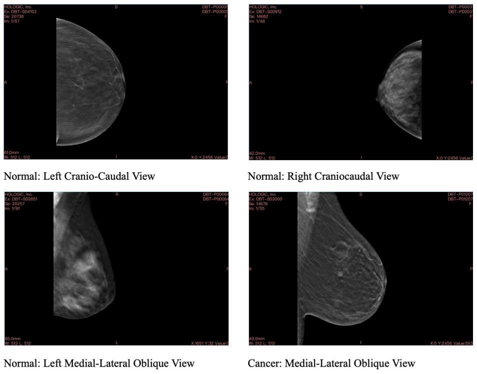

Normal: Left Cranio-Caudal View
Normal: Right Craniocaudal View
Normal: Left Medial-Lateral Oblique View
Cancer: Medial-Lateral Oblique View

Figure 3. Breast Cancer Tomosynthesis dataset

2. Related work

Anomaly Detection using VLMs - In the papers *”ANOMALYCLIP: OBJECT-AGNOSTIC PROMPT LEARNING FOR ZERO-SHOT ANOMALY DETECTION”* [10], *Anomaly Detection by Adapting a pre-trained Vision Language Model* [4], *A Zero-/Few-Shot Anomaly Classification and Segmentation Method* [5] explore anomaly detection using visual-language models (VLMs). These approaches utilize learnable prompts to obtain a unified representation of abnormal regions and incorporate a Region Refinement module for improved localization of anomalies. They extract image features and compare them with text prompts for effective anomaly classification. While these studies focus on multi-class anomaly detection across various domains, our project will specifically address anomaly detection in the medical domain. This area presents greater challenges than traditional anomaly detection due to the significant disparities between different data modalities, leading to complexities in accurately identifying medical data. Like AnomalyCLIP, we will be using the learnable text prompts approach in our project. AnomalyCLIP is one of our baselines.

Medical Anomaly Detection using VLMs - The paper ”Adapting Visual-Language Models for Generalizable Anomaly Detection in Medical Images” [6], *MediCLIP: Adapting CLIP for Few-shot Medical Image Anomaly Detection* [9] introduced a lightweight multi-level adaptation and comparison framework to re-purpose the CLIP model for medical anomaly detection. MVFA-AD [6] approach incorporates multiple residual adapters within the pre-trained visual encoder to enhance feature extraction progressively across multiple levels. MFVA-AD [6] approach generates fixed text embeddings using textual prompts. In our project, we are using learnable textual embeddings instead of fixed embeddings, allowing the model to better capture the nuances between images and descriptions. Additionally, their model performs both anomaly segmentation and anomaly classification, depending on the availability of ground truth in the dataset. For this project, we primarily focus on anomaly segmentation with ground truth. MVFA-AD is another one of our baselines. MediCLIP model uses zero-shot with a support set, but it primarily focuses on few-shot, which differs from our model as we do not use a support set for zero-shot detection.

Medical Anomaly Detection - There are several papers such as [2], [7] for medical anomaly detection using auto-encoders. This research introduces Reversed Auto-Encoders (RA), an unsupervised anomaly detection method designed to enhance the identification of diverse pathologies in medical imaging. By reconstructing pseudo-healthy versions of pathological inputs, RA detects a broader range of anomalies across various imaging modalities. But for this project, we will focus on AD using visual-language models.

Unlike other approaches, we utilize multiple feature patch tokens to compare the textual, image scores, and ground-truth segmentation scores, enhancing the model’s ability to process and analyze medical imaging data effectively.

3. Method

In this section, we outline the methodology we employed for anomaly detection in medical imaging using visual-language models, specifically focusing on CLIP [8]. Our approach involves several key components that leverage CLIP’s capabilities to improve the accuracy and reliability of anomaly detection in diverse medical contexts. We are using CLIP as our primary visual-language model.

3.1. Approach

Text prompts: In this project, we are using learnable text prompts for the images. We are using only two learnable text prompts: one representing ”normal” and another representing ”anomalous” images. For CLIP the textual prompts for normal are structured as ”normal [cls]” and for anomalous ”’abnormal [cls] with lesion’”. As it is difficult to list all possible anomaly types in the text prompts for each image, we are using a generic prompt template that will cover all of the scenarios. Existing models utilize multiple distinct prompts for each of the normal and abnormal categories. In our approach, we aim to evaluate the model’s performance when limited to using only two prompts. These prompts are processed by a text-encoder that transforms the text into high-dimensional textual embeddings for both the normal and abnormal categories. These embeddings provide a semantic representation that enables effective comparison with visual features extracted from the image.

Image encoding: To generate the visual embedding for the image, the input image is passed through a visual encoder of CLIP, where each layer processes it to generate a local visual embedding. Some encoder layers are frozen, preserving the pre-trained weights from the VLM, while we use adapters for learnable patch embeddings to adapt to medical imaging characteristics. Our approach integrates multiple residual adapters into the pre-trained visual encoder, enabling a stepwise enhancement of visual features across different levels. The dual-adapters outputs two sets of patch feature embeddings at each level, one is used for classification and another is used for segmentation. This allows the visual encoder to be sensitive to subtle anomalies present in medical images, where abnormal features are often small and difficult to distinguish from normal patterns.

Training: During training, we have both image and text prompts as input to the model, and we get textual and visual embeddings after passing through the text and vision encoders and adapters of CLIP. We are calculating similarity score using binary cross-entropy of each feature patch embedding with the text embedding, this will be our image(global) loss. Instead of relying on a single visual embedding for the similarity score, we utilize separate pixel-level feature embeddings, enabling more accurate anomaly classification. For segmentation of anomaly, we are computing a local similarity map of the segmentation feature embedding with the ground truth masks. This map helps identify specific regions within the image that deviate from normal patterns, enhancing the model’s ability to localize anomalies. We calculate local loss of the similarity map with respect to ground truth. Both the local and global loss are used to compute the total loss.
Training parameters:

- **Epochs:** 10
- **Batch size:** 8
- **Image size:** 336

- **Learning rate:** 0.001
- **Feature list:** [6, 12, 18, 24]
- **Optimizer:** Adam
- **Loss:** Dice, Focal, BCE

  Loss function $L_{total} = L_{\text{global}} + \lambda \sum_{M_k \in M} L_{M_k}^{\text{local}}$

Here, M is the numbers of layers of the model, λ is a trade-off hyper-parameters to balance the local and global loss.

For local loss for image segmentation we will be using Dice, Focal loss and global loss is binary cross-entropy loss. And S is the ground-truth segmentation mask

$$L_{local} = \text{Dice}(\text{softmax}(F_{seg}, F_{text}), S) + \text{Focal}(\text{softmax}(F_{seg}, F_{text}), S)$$

$$L_{global} = \text{BCE}(\text{softmax}(F_{cls}, F_{text}), S)$$

Testing: At test time, the model uses learned representations for images and prompts to detect anomalies in unseen images. First, the test image is passed through the visual encoder to extract multi-level visual features, capturing both high-level and fine-grained details. Simultaneously, text prompts representing ”normal” and ”anomalous” categories are processed by the text encoder, generating corresponding textual embeddings. The model then calculates similarity scores between these text embeddings and the image features at each layer using cosine similarity. These scores are combined into a global anomaly score for overall classification and local similarity maps for pixel-level anomaly localization.

4. Results

Figure 4: Visual Results for Anomalous Image Detection: This figure demonstrates the effectiveness of our proposed method in detecting anomalies in a set of brain MRI images. The results are compared with ground truth and our baseline models.

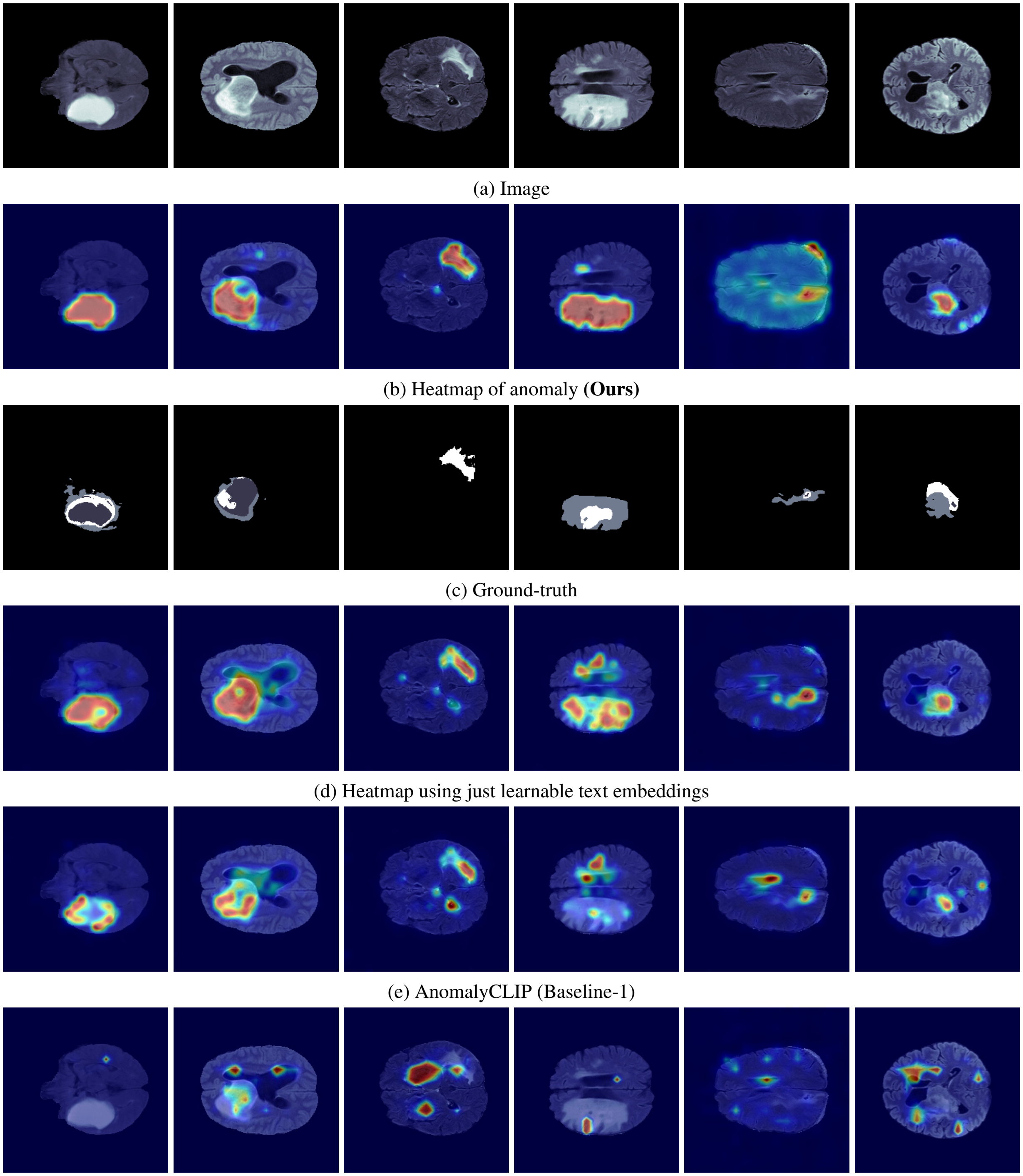
- **(a) Image**: The original input brain MRI images used for evaluation.
- **(b) Heatmap of anomaly (Ours)**: The anomaly localization heatmaps generated by our proposed method. These heatmaps highlight region of detected anomaly, demonstrating precise detection of anomalous areas.
- **(c) Ground-truth**: The ground truth segmentation masks indicating the actual locations of anomalies in the input images.
- **(d) Heatmap using just learnable text embeddings**: Results obtained when the dataset is trained exclusively with learnable text prompts without learnable adapter. This approach tests the model’s ability to leverage textual information for anomaly detection.
- **(e) AnomalyCLIP**: The results for anomalous images tested using the AnomalyCLIP model [10], this serves as one of the baseline for our model.
- **(f) MVFA-AD**: The results for anomalous images tested using the MVFA-AD model [6], this serves as one of the baselines for our model.

The comparison highlights that our method outperforms AnomalyCLIP and MVFA-AD in terms of anomaly localization, as evident from the more accurate and focused heatmaps in row (b), which closely align with the ground-truth masks in row (c) for Fig. 4. This demonstrates the capability of this method to improve anomaly detection accuracy in medical imaging tasks. Our model exhibits fewer irrelevant or noisy highlights in non-anomalous regions. This indicates that our method is more focused and avoids over estimating anomalies. MVFA-AD (f) compared to the other models fails to capture some of the anomalies as shown in the results of Fig 4.

In the Fig. 5, we show the results when the models are tested on ”normal” images (a). Our model (b) still detects some anomalies when there are none. However, in comparison, to text learnable (c) and AnomalyCLIP (d), ours visually is more accurate as the noise in the heatmap is significantly less compared to the (c) and (d) models. The MVFA-AD model performs slightly better in the case of non-anomalous images.

The heatmaps produced by our model show a sharper and more distinct contrast between anomalous and normal regions. This improves interpretability, making it easier for health-care workers to identify affected areas. Our model closely corresponds to the ground truth images, indicating that the model effectively captures the true anomalies.

We observed that a model utilizing only learnable text embeddings trained on the dataset outperforms the baseline AnomalyCLIP model[10], which was not trained on this specific dataset. However, incorporating both learnable image embeddings and learnable text embeddings significantly enhances the precision of anomaly detection compared to using learnable text embeddings alone. Interestingly, the MVFA-AD model [6], although trained using learnable image features, did not achieve a comparable performance, highlighting its limitations in this context. This observation underscores the importance of leveraging both image and text learnable parameters together, as relying solely on one modality—either image or text—results in suboptimal performance. The combination of both modalities allows for a more comprehensive understanding of the data, leading to a better anomaly detection accuracy.



(a) Image

(b) Heatmap of anomaly **(Ours)**

(c) Ground-truth

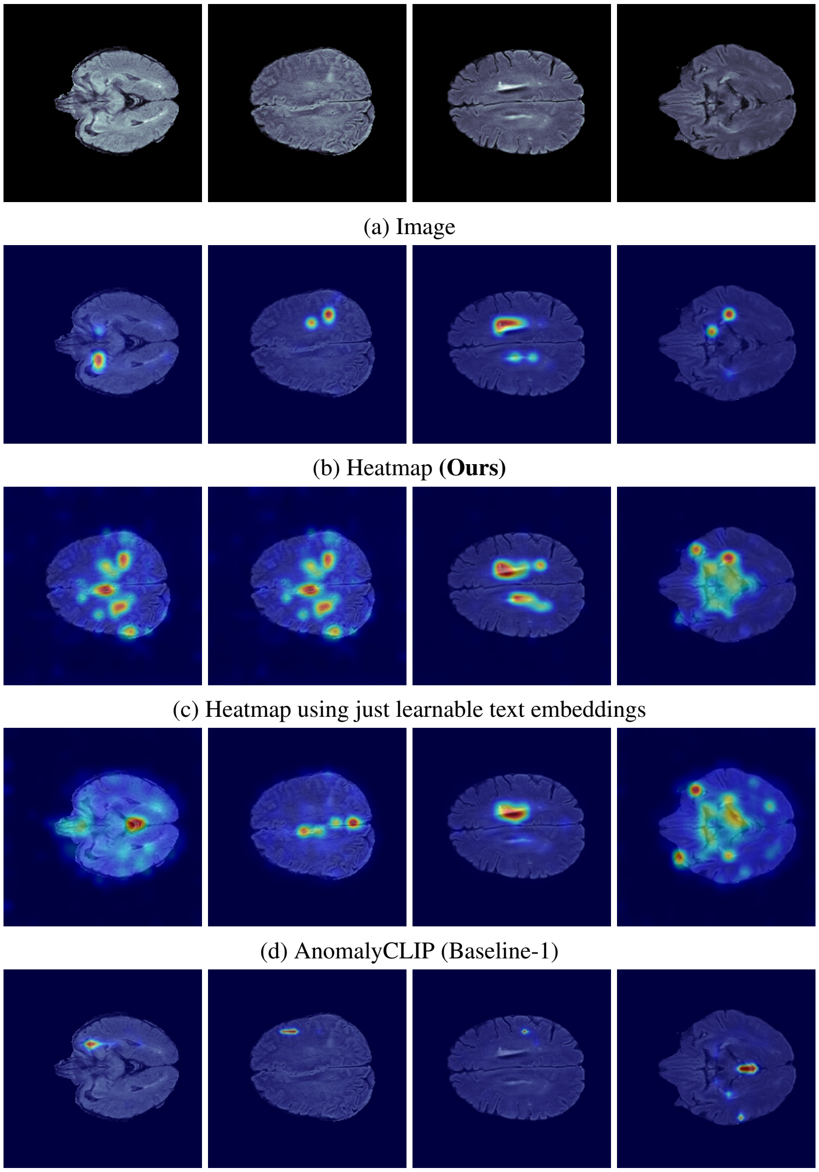
(d) Heatmap using just learnable text embeddings

(e) AnomalyCLIP (Baseline-1)

(f) MVFA-AD (Baseline-2)

Figure 4. Results for anomalous images using ours and baseline models

To test the model's capability for zero-shot learning, we also evaluated our model's performance on an unseen breast cancer dataset [3] Fig. 6, which presents a significantly more challenging task due to the lack of accurate anomaly ground truth images. Despite these challenges, our model demonstrated good performance in identifying anomalies, but not as accurate if presented with non-anomalous



(a) Image

(b) Heatmap **(Ours)**

(c) Heatmap using just learnable text embeddings

(d) AnomalyCLIP (Baseline-1)

(e) MVFA-AD (Baseline-2)

Figure 5. Results for non-anomalous images using ours and baseline models

images. This dataset did not include ground truth, hence why we did not compute AUC in our evaluation. Visually, our model does well with detecting anomalies in cancerous breast tosmosynthesis images.

For liver CT-scans, we see that MVFA-AD (e) model performs the best. Our model (c) does a good job in reducing noise as opposed to AnomalyCLIP (d). However, as stated before, our model is only trained on brain MRI scans, hence why we do not see a high accuracy in detecting liver anomalies.

Furthermore, we note that training the model on a broader range of anomaly classes has the potential to significantly improve its performance for a wider range of medical anomaly detection tasks. By exposing the model to diverse anomaly types during training, we can enhance its ability to generalize across different datasets and medical imaging modalities, making it a more robust solution for real-world applications.

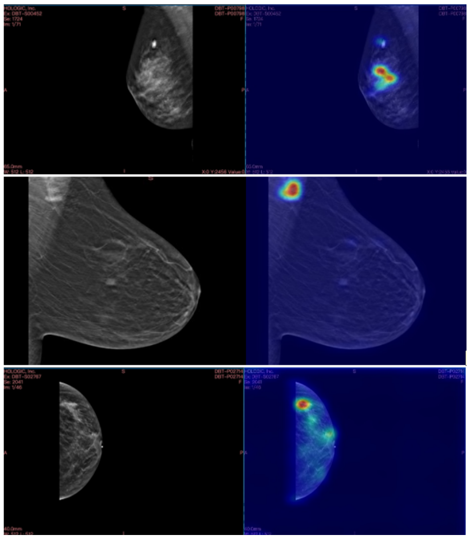


Figure 6. Test results on breast-cancer dataset

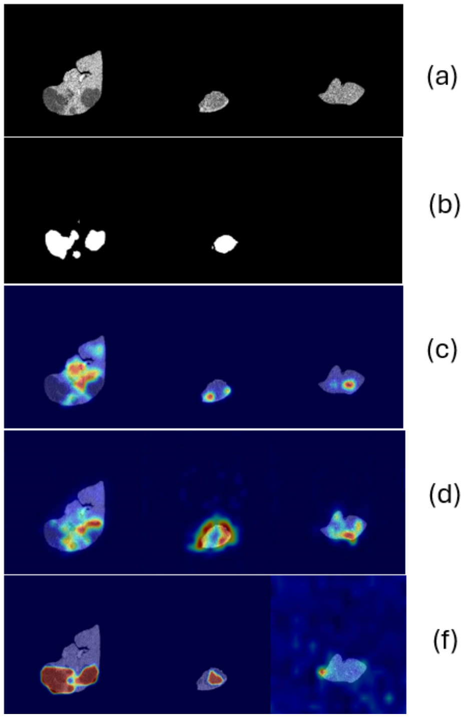


Figure 7. Zero-shot test results on Liver-CT dataset (a) Image, (b) Ground-truth, (c) MedVMAD (Ours), (d) AnomalyCLIP, (e) MVFA-AD. First two are anomalous images, and third one is a nomal image

5. Evaluation

The evaluation results of the proposed MedVMAD model are presented in Table 1, alongside comparisons with AnomalyCLIP [10] and MVFA-AD [6]. The performance metrics used for comparison include Pixel AUROC, Pixel AUPRO, Image AUROC, and Image AP. These metrics provide a comprehensive assessment of the models' capabilities in anomaly detection tasks.

Model	**Pixel AUROC**	**Pixel AUPRO**	**Image AUROC**	**Image AP**
AnomalyCLIP	96.5	**77**	70.3	76
MVFA-AD	89.1	57.4	79.6	84.5
MedVMAD **(Ours)**	**96.7**	56.9	**84.6**	**87.7**

Table 1. Performance Metrics (in %) with BRaTS2021 Dataset

Pixel AUROC reflects the model's ability to distinguish anomalous pixels from normal ones. MedVMAD achieved the highest score of 96.7, surpassing AnomalyCLIP (96.5) and MVFA-AD (89.1), demonstrating higher precision in pixel-level anomaly detection.

Pixel AUPRO, which measures the area under the precision-recall curve for pixel-level anomalies, shows that AnomalyCLIP performed the best with 77, while MedVMAD achieved 56.9, which is slightly lower than MVFA-AD's 57.4.

Image AUROC assesses the model's ability to classify entire images as anomalous or normal. MedVMAD performed better with a score of 84.6, significantly outperforming both AnomalyCLIP (70.3) and MVFA-AD (79.6). This highlights the model's ability in capturing holistic anomalies across the image.

Image AP, the average precision for image-level anomalies, further confirms MedVMAD's effectiveness, achieving the highest score of 87.7, followed by MVFA-AD (84.5) and AnomalyCLIP (76).

MediCLIP [6] used a different dataset for their evaluations and achieved an Image-AUROC of 91.3 ±1.0 using zero-shot anomaly detection with a support set size of k=16.

Model	**Pixel AUROC**	**Pixel AUPRO**	**Image AUROC**	**Image AP**
AnomalyCLIP	95.1	**82.8**	60.4	58.3
MVFA-AD	**99.9**	60.3	**84.5**	**87.3**
MedVMAD **(Ours)**	95.2	58	56.7	51.2

Table 2. Zero-shot Performance Metrics with BTCV + LiTs (Liver CT) Dataset

As shown in table 2, even though our model does not perform as well as AnomalyCLIP and MVFA-AD for zero-shot detection, our model was only trained on one class of medical anomalies, so it is as expected. However, despite only being trained on one class, our model visually is comparable to AnomalyCLIP as shown in the results. In contrast to MediCLIP, our model does not use a support set for zero-shot detection, highlighting the difference in methodology.

Fig. 8 shows a steady decline in loss as training progresses, indicating that the model is effectively learning the image-text alignment task over time. In the first few epochs, there is a significant reduction in loss, suggesting that the model quickly captures the fundamental relationships between the image and text representations. We observed that

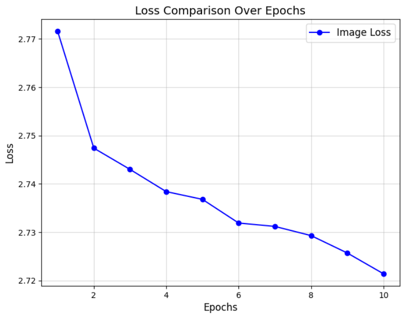


Figure 8. Image loss (Image and text similarity loss) loss over epochs

training beyond ten epochs was unnecessary, as the rate of decline in the model's loss significantly slowed and due to limited computational power to train for more epochs. Therefore, we chose to limit training to 10 epochs.

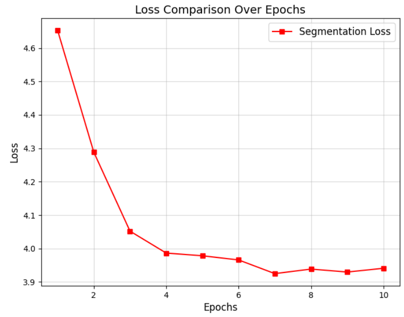


Figure 9. Segmentation loss over epochs

In fig. 9, during the first few epochs, there is a significant reduction in segmentation loss, indicating that the model quickly learns to segment and identify the relevant features effectively. Around epoch 5, the rate of segmentation loss reduction slows, thus showing that the model was approaching convergence. Segmentation loss converges and beyond 10 epochs, we did not see further reduction, hence why we selected 10 epochs.

6. Future Work

Since our model was only trained on one class, our results for zero-shot anomaly detection were not as accurate as they could have been. For further improving the model, we could train it on multiple classes, including the entirety of the BMAD dataset [1], as well as other diverse medical image datasets, rather than just on brain MRIs. Expanding the range of anomaly classes would enhance the model's ability to generalize across different medical imaging modalities. Additionally, incorporating a larger variety of medical conditions could improve the model's robustness in real-world clinical settings, where diverse anomaly types are encountered. However, despite the fact that our model is only trained on one class, our model does reasonably well for the breast cancer dataset.

Zero-shot anomaly detection also holds promise for medical anomalies where data is limited. In many cases, medical imaging data is limited, due to privacy concerns, data sharing restrictions, or due to the high costs associated with collecting large, labeled datasets. Zero-shot anomaly detection may lessen this challenge by enabling the model to detect abnormalities without needing extensive training data for each new condition, which makes it a valuable tool for healthcare systems with limited data.

7. Conclusion

Our model, MedVMAD, which leverages both learnable text and image embeddings, in general, outperforms existing models in terms of anomaly localization and detection accuracy. The results demonstrate that combining these two modalities allows the model to capture a more comprehensive understanding of the data. The positive performance on the challenging dataset with breast cancer images, further supports its potential for use in clinical applications. Future work will explore expanding the range of anomaly types during training to further improve the model's performance across different classes.

References

[1] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection, 2024. 2, 8

[2] Cosmin I. Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A. Schnabel. Towards universal unsupervised anomaly detection in medical imaging, 2024. 3

[3] M. Buda, A. Saha, R. Walsh, S. Ghate, N. Li, A. Swiecicki, J. Y. Lo, J. Yang, and M. Mazurowski. Breast cancer screening – digital breast tomosynthesis (bcs-dbt) (version 5), 2020. Dataset. 2, 5

[4] Yuxuan Cai, Xinwei He, Dingkang Liang, Ao Tong, and Xiang Bai. Anomaly detection by adapting a pre-trained vision language model, 2024. 2

[5] Xuhai Chen, Yue Han, and Jiangning Zhang. April-gan: A zero-/few-shot anomaly classification and segmentation method for cvpr 2023 vand workshop challenge tracks 12: 1st place on zero-shot ad and 4th place on few-shot ad, 2023. 2

[6] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images, 2024. 3, 4, 7

[7] Shuai Lu, Weihang Zhang, Jia Guo, Hanruo Liu, Huiqi Li, and Ningli Wang. Patchcl-ae: Anomaly detection for medical images using patch-wise contrastive learning-based auto-

encoder. *Computerized Medical Imaging and Graphics*, 114: 102366, 2024. 3

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 3

[9] Ximiao Zhang, Min Xu, Dehui Qiu, Ruixin Yan, Ning Lang, and Xiuzhuang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection, 2024. 3

[10] Qihang Zhou, Guansong Pang, Yu Tian, Shibo He, and Jiming Chen. Anomalyclip: Object-agnostic prompt learning for zero-shot anomaly detection, 2024. 2, 4, 7