MCMIAD: Multi-Class Model for Medical Image Anomaly Detection

Shih-Chih Lin[1,2] LEOLIN65@GAPP.NTHU.EDU.TW
Jenq-Neng Hwang[2] HWANG@UW.EDU
Shang-Hong Lai[1] LAI@CS.NTHU.EDU.TW

[1] *International Intercollegiate Ph.D. Program and Department of Computer Science,*
National Tsing Hua University, Hsinchu, Taiwan
[2] *Department of Electrical and Computer Engineering,*
University of Washington, Seattle, WA, USA

Editors: Under Review for MIDL 2026

Abstract

Accurate anomaly detection in medical imaging is critical for clinical decision-making, yet many deployed systems still rely on disease-specific models and large labeled datasets. We present **MCMIAD**, a unified vision–language framework that couples a frozen EfficientNet image encoder and a CLIP text encoder with a shallow cross-modal fusion block and a denoising Transformer decoder. The framework is designed around three goals: *modality-agnostic deployment*, *prompt-guided explainability*, and *practical efficiency.*

MCMIAD keeps the vision backbone frozen and trains only a compact reconstruction head, making the method lightweight enough for typical clinical GPUs. On the BMAD benchmark, MCMIAD achieves strong image- and pixel-level AUROC across retina OCT, brain tumor MRI, and liver tumor CT, with particularly notable gains in one-shot settings where only a single normal example per category is available. Its anomaly heatmaps align with expected clinical regions of interest, supporting human-in-the-loop review.

We further analyze the contributions of CLIP-guided cross-attention, model size, and we discuss robustness, fairness, and deployment considerations relevant to real-world clinical workflows.

Keywords: Medical anomaly detection, vision–language models, CLIP

1. Introduction

Anomaly detection (AD) in medical imaging underpins early disease diagnosis and longitudinal follow-up. However, moving AD models from prototypes to real clinical systems remains challenging. Most supervised methods are trained per disease or per modality, require dense pixel-level annotations, and generalize poorly to unseen pathologies or new scanners. In many hospitals, maintaining a collection of disease-specific models is operationally infeasible.

Unsupervised anomaly detection (UAD) reduces annotation cost by learning only from normal images, but classical reconstruction- or memory-based methods still struggle with limited semantic interpretability, suboptimal localization, and sensitivity to domain shifts (Bao et al., 2024; Roth et al., 2022; Zavrtanik et al., 2021). At the same time, advances in multimodal vision–language models (VLMs) such as CLIP (Radford et al., 2021) have enabled cross-modal reasoning and open-vocabulary recognition. Applying these models directly to medical AD is, however, non-trivial: generic pre-training introduces domain gaps, and off-the-shelf VLMs provide weak pixel-level localization and are rarely optimized for unified multi-disease settings (Zhang et al., 2024; Jeong et al., 2023).

A central open question is therefore how to build a *unified*, *explainable*, and *efficient* medical AD framework that: (i) reuses strong vision–language backbones, (ii) scales across diseases and imaging modalities, and (iii) provides clinically meaningful heatmaps with modest computational overhead.

Our Approach. To address this, we propose **MCMIAD** (Multi-Class Model for Medical Image Anomaly Detection), a vision–language UAD framework that integrates: (i) an EfficientNet (Koonce, 2021) vision encoder for visual features, (ii) a CLIP text encoder with offline prompt embeddings distilled from a small set of carefully designed "normal" descriptions, and (iii) a shallow cross-attention fusion block followed by a denoising Transformer decoder. The vision encoders remain frozen, and only a compact reconstruction head is trained, which simplifies optimization and supports deployment on commodity GPUs. MCMIAD reconstructs normal features and derives anomaly maps from the discrepancy between the input and reconstructed features. Cross-modal fusion is controlled solely by a flag (`use_text_fusion`), enabling clean ablations that isolate the effect of incorporating semantic information.

Contributions. Our main contributions are summarized as follows:

- We introduce **MCMIAD**, a unified vision–language framework that combines a frozen, EfficientNet vision backbone with a compact reconstruction head for modality-agnostic medical anomaly detection and localization.

- We propose a **shallow cross-modal fusion module** for CLIP-guided reconstruction and a simple feature jittering strategy that jointly enhance robustness, localization, and generalization under domain shifts and limited supervision.

- We show that MCMIAD attains strong image- and pixel-level AUROC on three heterogeneous BMAD datasets, while using only a small number of trainable parameters in the task-specific head, highlighting a favorable trade-off between performance and efficiency.

- We provide ablations on cross-modal fusion, one-shot learning, and parameter efficiency make MCMIAD suitable for human-in-the-loop clinical use.

2. Related Work

Unsupervised Anomaly Detection in Medical Imaging. UAD in medical imaging has progressed from early reconstruction- and embedding-based approaches to modern unified and cross-modal frameworks. Feature-embedding methods often combine a pre-trained CNN backbone with nearest-neighbor matching or explicit density estimation in latent space (Roth et al., 2022; Rudolph et al., 2021). Memory-based designs such as PatchCore (Roth et al., 2022) achieve strong AUROC on industrial images, but require careful feature selection and may be fragile under domain shifts. Reconstruction-based methods,

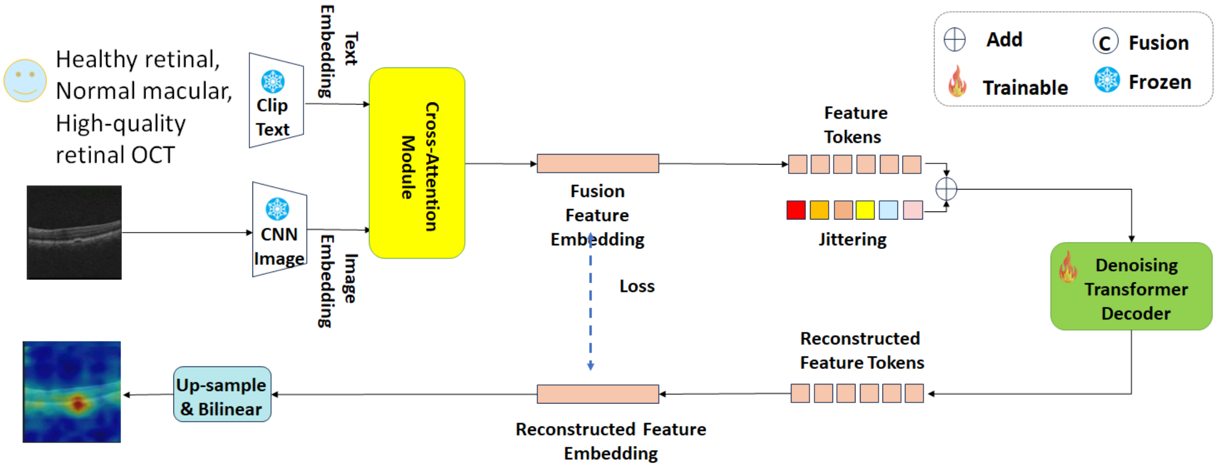

Figure 1: Overall architecture of the proposed MCMIAD framework. A frozen CNN based image encoder and CLIP text encoder provide visual and semantic features that are aligned via a shallow cross-attention block, followed by feature jittering and a denoising Transformer decoder that reconstructs normal features and yields anomaly heatmaps.

including autoencoders, GANs, and variational models, flag anomalies via reconstruction error, but the reconstruction objective can inadvertently model abnormal structures, reducing contrast between normal and anomalous regions.

Synthesis- and Reconstruction-Based Approaches. Synthesis-based methods like DRAEM (Zavrtanik et al., 2021) generate pseudo-anomalies via noise or texture mixing to better regularize the detector. Recent reconstruction-based models focus on multi-scale structure and edges (Liu et al., 2022; Deng and Li, 2022), or combine quantization with Transformer decoders. These approaches improve localization, yet often remain modality-specific and computationally heavy, limiting their applicability to multi-disease clinical workflows.

Vision–Language Models and Cross-Modal AD. CLIP (Radford et al., 2021) and related VLMs have inspired several anomaly detection methods that leverage text prompts to guide localization and scoring. WinCLIP (Jeong et al., 2023) and CLIP Surgery (Li et al., 2023) adapt CLIP for zero-shot or few-shot industrial AD. Medical extensions (Zhang et al., 2024) fine-tune CLIP on paired images and reports, but typically focus on classification or segmentation rather than unified UAD. Moreover, many VLM-based methods rely on cosine similarity in the embedding space and do not explicitly reconstruct normal features, which can limit pixel-level interpretability.

Unified and Transformer-Based Anomaly Detection. Unified architectures such as UniAD (You et al., 2022), HVQ-Trans (Lu et al., 2023), and DiAD (He et al., 2024) use Transformer decoders and reconstruction to support multi-class or cross-domain AD. While they report strong AUROC, some models introduce substantial computational cost (e.g., diffusion samplers) or large task-specific heads. MCMIAD provides a complementary design by leveraging a frozen vision backbone and a minimal reconstruction head with explicit text-guided fusion, specifically targeting deployment settings that prioritize parameter efficiency and clinically meaningful interpretability.

3. Proposed Framework: MCMIAD

3.1. Architecture Overview

MCMIAD is a unified vision–language framework for robust anomaly detection across heterogeneous modalities. As shown in Fig. 1, it consists of:

1. a frozen CNN based vision encoder,
2. a frozen CLIP text encoder with offline prompt embeddings,
3. a shallow cross-modal fusion block implemented as a single Multi-Head Cross-Attention layer,
4. a feature jittering module, and
5. a denoising Transformer decoder that reconstructs normal features.

Anomaly heatmaps are derived from the discrepancy between input and reconstructed features. The reconstruction head is the only trainable component during MCMIAD training, totaling **2.97M** parameters.

3.2. Frozen CLIP Encoders and Offline Prompt Embeddings

We adopt an EfficientNet (Koonce, 2021) as the visual backbone and its paired text encoder for prompts. Before integrating them into MCMIAD, we perform two lightweight steps.

Light domain adaptation. We first fine-tune CLIP on a small set of in-domain image–text pairs (brain MRI, retina OCT, and liver CT slices with radiology-style descriptions) using a contrastive loss. The adapted weights are then frozen for all MCMIAD experiments, mitigating domain gap while preserving the generality of CLIP.

Offline "normal" text embedding. To stabilize training and avoid prompt engineering at inference, we precompute a single dataset-level text embedding that summarizes multiple clinically realistic descriptions of normal anatomy. Concretely, we construct 15 prompts (five per modality) describing normal brain MRI, retina OCT, and liver CT appearance. Each prompt is encoded using the adapted CLIP text encoder, L_2-normalized, and then averaged:

$$\mathbf{e}_{\text{txt}} = \frac{1}{15} \sum_{i=1}^{15} \frac{\mathbf{t}_i}{\|\mathbf{t}_i\|_2}.$$

The resulting embedding $\mathbf{e}_{\text{txt}} \in \mathbb{R}^{D_{\text{text}}}$ serves as a shared semantic prior and is loaded once during model initialization.

For a given image $\mathbf{I}$, the CLIP image encoder produces a feature map $\mathbf{F}_{\text{in}} \in \mathbb{R}^{C_{\text{org}} \times H \times W}$. After flattening and a linear projection, we obtain visual tokens

$$\mathbf{X}_{\text{vis}} \in \mathbb{R}^{N \times D}, \quad N = H \cdot W, \tag{1}$$

which are then processed by the cross-modal fusion module.

3.3. Embedding-Level Fusion via Shallow Cross-Attention

We design a single-layer cross-attention module that injects semantic information from $\mathbf{e}_{\text{txt}}$ into visual tokens while keeping the decoder architecture unchanged.

Let $\mathbf{X}_{\text{vis}} \in \mathbb{R}^{N \times D}$ be the visual tokens and $\mathbf{e}_{\text{txt}} \in \mathbb{R}^{D_{\text{text}}}$ the precomputed text embedding. We first project $\mathbf{e}_{\text{txt}}$ into the decoder dimension:

$$\mathbf{T} = \mathbf{e}_{\text{txt}} \mathbf{W}_K = \mathbf{e}_{\text{txt}} \mathbf{W}_V \in \mathbb{R}^D,$$

where $\mathbf{W}_K$ and $\mathbf{W}_V$ are learned linear layers. We then treat $\mathbf{X}_{\text{vis}}$ as queries and $\mathbf{T}$ as keys and values in a multi-head cross-attention layer:

$$\mathbf{Z} = \text{CrossAttn}\big(Q = \mathbf{X}_{\text{vis}},\, K = \mathbf{T},\, V = \mathbf{T}\big). \tag{2}$$

The fused representation is computed using a simple residual formulation:

$$\mathbf{X}_{\text{fused}} = \mathbf{X}_{\text{vis}} + \alpha\, \mathbf{Z},$$

where α is a fixed scalar weight (set to 1.0 in all experiments). This design enables clean ablations: removing the fusion block eliminates all text-derived signals while keeping the remainder of the architecture identical, allowing us to isolate the contribution of semantic information.

3.4. Feature Jittering

To improve robustness to scanner-specific artifacts, acquisition noise, and moderate distribution shift, we apply stochastic perturbations to the visual tokens in the spatial domain during training. This encourages the decoder to learn representations that are stable under local intensity variations rather than memorizing low-level patterns.

Concretely, given the fused tokens $\mathbf{X}_{\text{fused}}$, we add Gaussian noise proportional to the feature norm:

$$\mathbf{X}_{\text{jitter}} = \mathbf{X}_{\text{fused}} + \epsilon, \quad \epsilon \sim \mathcal{N}\big(0, \sigma^2 \mathbf{I}\big), \tag{3}$$

with probability p. The noise scale σ and probability p are fixed across all datasets. When no jittering is applied, we simply set $\mathbf{X}_{\text{jitter}} = \mathbf{X}_{\text{fused}}$. In practice, this simple spatial perturbation is sufficient to regularize MCMIAD and improves generalization without adding architectural complexity.

3.5. Denoising Transformer Decoder

The perturbed tokens $\mathbf{X}_{\text{jitter}}$ are passed to a Wide Transformer decoder (MoEAD head) composed of repeated self-attention, neighborhood-masked attention, and Feed Forward Network blocks.

Given position embeddings $\mathbf{P}$, the decoder outputs reconstructed tokens $\mathbf{Y} \in \mathbb{R}^{N\times D}$, which are mapped back to the original channel dimension via a linear layer and reshaped into $\mathbf{F}_{\text{out}} \in \mathbb{R}^{C_{\text{org}}\times H\times W}$.

3.6. Anomaly Localization and Scoring

We compute anomaly scores by combining pixel-wise reconstruction error and cosine similarity between input and reconstructed features:

$$\mathbf{S}_{\text{MSE}} = \|\mathbf{F}_{\text{in}} - \mathbf{F}_{\text{out}}\|_2^2, \tag{4}$$

$$\mathbf{S}_{\text{cos}} = 1 - \cos\big(\mathbf{F}_{\text{in}}, \mathbf{F}_{\text{out}}\big), \tag{5}$$

where cosine similarity is computed channel-wise at each spatial location. The final anomaly score is obtained by a direct sum of the two terms:

$$\mathbf{S} = \mathbf{S}_{\text{MSE}} + \mathbf{S}_{\text{cos}}. \tag{6}$$

The score map is then upsampled via bilinear interpolation to generate pixel-wise anomaly heatmaps.

Image-level anomaly scores are aggregated using a weighted sum of the maximum, mean, and Top-K mean pixel scores:

$$S_{\text{image}} = w_1 \cdot \max(\mathbf{S}) + w_2 \cdot \text{mean}(\mathbf{S}) + w_3 \cdot \text{TopKMean}_K(\mathbf{S}), \tag{7}$$

with fixed weights $(w_1, w_2, w_3) = (0.5, 0.3, 0.2)$ and $K = 5$ across all datasets.

3.7. Summary

MCMIAD integrates frozen encoders, a shallow cross-attention fusion block, feature jittering, and a compact reconstruction decoder. All semantic information flows through the precomputed CLIP text embedding and a single cross-attention layer, making the influence of language both interpretable and easy to ablate. We next evaluate MCMIAD on BMAD and compare it with state-of-the-art UAD baselines.

4. Experiments

4.1. Benchmark and Datasets

We evaluate MCMIAD on the BMAD benchmark (Bao et al., 2024), which provides clinically curated, pixel-annotated datasets for: (i) brain tumor MRI (T2-FLAIR), (ii) liver tumor CT, and (iii) retina OCT. Each dataset includes normal and anomalous samples with expert delineations of lesions. We follow the official BMAD splits and report Area Under the Receiver Operating Characteristic (AUROC) at both *image* level (classifying each slice as normal or abnormal) and *pixel* level (localizing anomalous regions). This dual evaluation reflects both detection and localization performance.

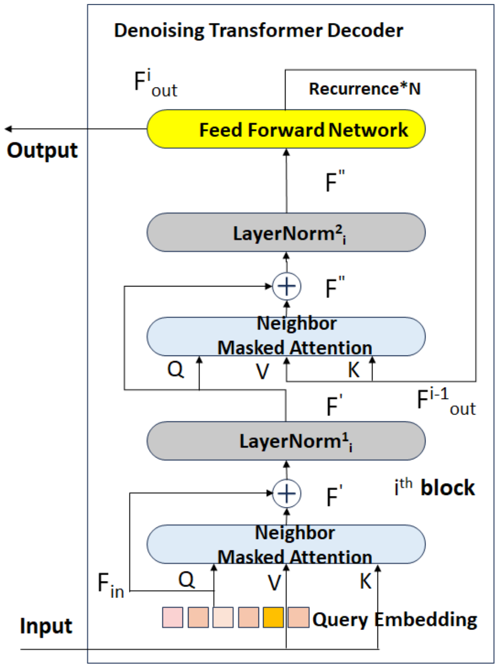

Figure 2: Denoising decoder of MCMIAD. The decoder stacks self-attention, neighborhood-masked attention to reconstruct normal features from jittered, cross-modally fused tokens.

4.2. Implementation Details

All experiments are conducted on a single NVIDIA RTX 4090 GPU. Input images are resized to 224 × 224 and normalized using ImageNet statistics. Unless otherwise stated, we train one MCMIAD model per modality to align with BMAD baselines; in §4.5 we additionally analyze the unified multi-modality setting.

Backbone and offline CLIP text embedding. We use Efficient-B4 (Koonce, 2021) with weights as the visual backbone. After a light in-domain contrastive adaptation, both the image and text encoders remain frozen for all experiments.

To construct the semantic prior, we encode 15 clinically phrased normal prompts (five per modality), L_2-normalize each embedding, and take their average: this yields a single modality-agnostic text representation $\mathbf{e}_{\text{txt}}$ used throughout training and inference.

Ablations that remove CLIP-based guidance operate by substituting $\mathbf{e}_{\text{txt}}$ with a zero vector of identical dimension, ensuring that architectural capacity remains unchanged while isolating the effect of semantic information.

Training protocol. We follow the standard UAD setting and train MCMIAD using only normal images; anomalous images are reserved for validation and test. We use AdamW with learning rate 2×10^{-4}, weight decay 10^{-4}, cosine or step learning-rate decay (selected via validation), gradient clipping with max norm 0.1, and batch size 16. Each model is trained

Table 1: Image-AUROC (%) and Pixel-AUROC (%) on the BMAD benchmark. Best results in **bold**.

Image-AUROC (%)					
Category	DRAEM	UniAD	SimpleNet	DiAD	MCMIAD (Ours)
Liver Tumor	59.1	61.0	55.8	59.2	**75.0**
Brain Tumor	69.2	89.9	82.3	**93.7**	89.1
Retina	51.7	84.6	**88.8**	88.3	82.8
Average	60.0	78.5	75.6	80.4	**82.3**

Pixel-AUROC (%)					
Category	DRAEM	UniAD	SimpleNet	DiAD	MCMIAD (Ours)
Liver Tumor	52.9	97.1	**97.4**	97.1	95.8
Brain Tumor	52.0	**97.4**	94.8	95.4	96.6
Retina	57.4	94.8	95.5	95.3	**95.5**
Average	54.1	**96.4**	95.9	95.9	96.1

for 100 epochs, and results are averaged over three random seeds; standard deviations are reported in the appendix.

Reconstruction and scoring. We set $\lambda = 0.7$ in Eq. (5) to balance intensity- and structure-based differences. Image-level scores use the fixed weights $(w_1, w_2, w_3) = (0.5, 0.3, 0.2)$. Hyperparameters are selected on a validation split of liver CT and kept unchanged for brain MRI and retina OCT.

4.3. CLIP Text Prompt Details

We construct the offline text embedding from clinically phrased descriptions of normal anatomy (e.g., healthy brain MRI, retina OCT, and liver CT slices). Representative examples are shown here, and the complete prompt list is provided in the Appendix.

4.4. Quantitative Comparison with Baselines

Table 1 compares MCMIAD with representative UAD baselines, including DRAEM (Zavrtanik et al., 2021), UniAD (You et al., 2022), SimpleNet (Liu et al., 2023), and DiAD (He et al., 2024). Across liver CT, brain MRI, and retina OCT, MCMIAD achieves the best *average* image-level AUROC (82.3%) and competitive pixel-level AUROC (96.1%), slightly below UniAD (96.4%). The largest image-level gain appears on liver CT, where MCMIAD surpasses the strongest baseline by +14 AUROC while maintaining accurate localization. On brain and retina, it remains competitive, illustrating the trade-off between extreme per-dataset performance and parameter efficiency.

4.5. Model Size vs. Accuracy

To quantify efficiency, Table 2 compares the number of trainable parameters in the task-specific head and the average AUROC across BMAD. vision and text backbones are frozen

Table 2: Model size vs. AUROC. "Avg." denotes the mean of image- and pixel-level AUROC (in %). Trainable parameters refer to the task-specific head; CLIP backbone parameters are frozen.

Model	Trainable Parameters (Millions)	Avg. Image-/Pixel-AUROC (%)
DiAD (He et al., 2024)	1300	88.2
UniAD (You et al., 2022)	7.7	87.5
MCMIAD (Ours)	**2.97**	**89.2**

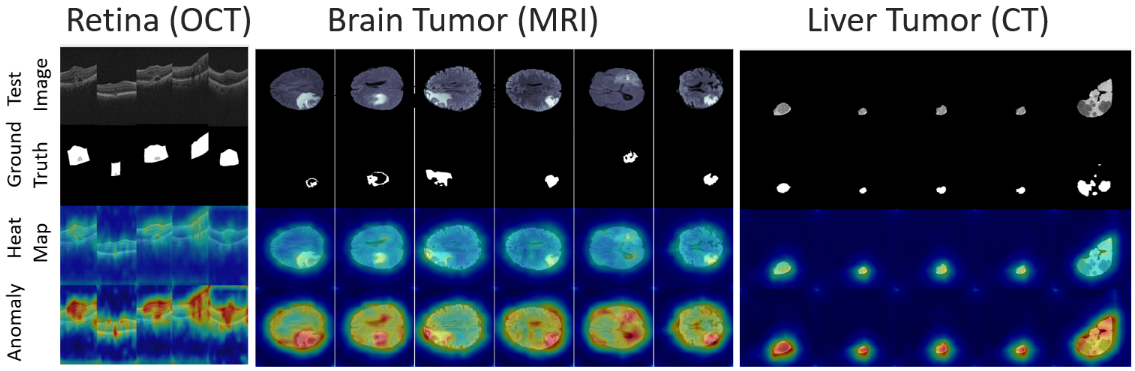

Figure 3: Qualitative visualization across retina OCT, brain tumor MRI, and liver tumor CT. MCMIAD accurately localizes anomalies while keeping false positives low, especially for liver lesions, and produces prompt-guided heatmaps that align with expected clinical regions of interest.

and shared wherever used. MCMIAD achieves the competitive average AUROC while using less than a fifth of the parameters of UniAD and orders of magnitude fewer parameters than diffusion-based models.

4.6. Qualitative Visualization

As illustrated in Fig. 3, MCMIAD yields compact, well-localized heatmaps across all three BMAD modalities. On liver CT, the model highlights focal lesions while suppressing background parenchyma, whereas on brain MRI and retina OCT it emphasizes tumor cores and pathological retinal layers with minimal spurious responses. These examples illustrate how CLIP-guided fusion and reconstruction-based scoring jointly produce interpretable, clinically plausible anomaly maps that are suitable for human-in-the-loop review.

4.7. One-Shot Performance

We evaluate the one-shot setting, where only a single normal example per dataset is available during adaptation. Table 3 reports image- and pixel-level AUROC. MCMIAD yields substantial gains in liver CT (image-level AUROC +42.4 over UniAD) and improves or matches UniAD in most pixel-level metrics, demonstrating strong sample efficiency.

Table 3: One-shot anomaly detection on BMAD (AUROC %). Best per column in **bold**.

Dataset	Image-AUROC (UniAD)	Image-AUROC (MCMIAD)	Pixel-AUROC (UniAD)	Pixel-AUROC (MCMIAD)
Liver Tumor	35.0	**77.4**	88.5	**96.0**
Brain Tumor	50.0	**59.4**	**93.6**	88.8
Retina OCT	53.5	**58.7**	80.7	**84.0**
Average	46.2	**65.2**	87.6	**89.6**

Table 4: Ablation on the **cross-modal fusion module** (AUROC %).

Method	Image-AUROC	Pixel-AUROC
w/o CLIP fusion	80.9	94.8
w/ CLIP fusion	**82.3**	**96.1**

5. Ablation Study: Effectiveness of Cross-Modal Fusion

We assess the effect of CLIP-guided fusion by enabling or disabling `use_text_fusion`, with all other components kept fixed. Table 4 summarizes the results.

The improvements over the non-fusion variant show that the benefit is not due merely to added parameters or architectural depth. Instead, the CLIP-derived semantic prior provides meaningful guidance, yielding consistent gains in both image- and pixel-level AUROC. This confirms that prompt-based semantics play a central role in MCMIAD's effectiveness.

6. Ethics, Data, Reproducibility, and Limitations

All experiments use public, de-identified datasets. Code, configurations, prompts, and precomputed CLIP embeddings will be released to support reproducibility. We report training protocols, evaluation metrics, and random seeds, and we examine potential biases such as scanner artifacts and class imbalance. Although MCMIAD improves robustness on BMAD (Bao et al., 2024), broader multi-site validation is required before clinical use. Domain-adapted CLIP encoders may still encode spurious correlations, underscoring the need for human oversight. MCMIAD currently operates on 2D slices, which may limit performance for modalities where 3D context is important. The unified text embedding is derived from normal-only descriptions; richer prompt sets may enhance localization.

7. Conclusion and Future Work

We introduced **MCMIAD**, a unified and efficient vision–language framework for medical anomaly detection. MCMIAD integrates a vision encoder, offline normal-text prompts, a shallow cross-modal fusion block, and a compact denoising Transformer decoder. With only **2.97M** trainable parameters and **3.26G** FLOPs in the task-specific head, the model is suitable for standard clinical GPUs while achieving strong image- and pixel-level AUROC across retina OCT, brain MRI, and liver CT.

Looking forward, extending MCMIAD to 3D volumes, exploring unified multi-modality checkpoints, and incorporating retrieval-augmented reasoning or large language models may further enhance scalability and explainability in real-world clinical workflows.

References

Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4053, 2024.

Hanqiu Deng and Xingyu Li. Anomaly detection via reverse distillation from one-class embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9737–9746, June 2022.

Haoyang He, Jiangning Zhang, Hongxu Chen, Xuhai Chen, Zhishan Li, Xu Chen, Yabiao Wang, Chengjie Wang, and Lei Xie. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 8472–8480, 2024.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar 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.

Brett Koonce. Efficientnet. In *Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization*, pages 109–123. Springer, 2021.

Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, pages arXiv–2304, 2023.

Tongkun Liu, Bing Li, Zhuo Zhao, Xiao Du, Bingke Jiang, and Leqi Geng. Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection. *arXiv preprint arXiv:2210.14485*, 2022.

Zhikang Liu, Yiming Zhou, Yuansheng Xu, and Zilei Wang. Simplenet: A simple network for image anomaly detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20402–20411, 2023.

Ruiying Lu, YuJie Wu, Long Tian, Dongsheng Wang, Bo Chen, Xiyang Liu, and Ruimin Hu. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36:8487–8500, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

Karsten Roth, Latha Pemula, Joaquin Zepeda, Bernhard Schölkopf, Thomas Brox, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.

Marco Rudolph, Bastian Wandt, and Bodo Rosenhahn. Same same but differnet: Semi-supervised defect detection with normalizing flows. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1907–1916, 2021.

Zhiyuan You, Lei Cui, Yujun Shen, Kai Yang, Xin Lu, Yu Zheng, and Xinyi Le. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35:4571–4584, 2022.

Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8330–8339, 2021.

Xiaoran Zhang, Meng Xu, Dong Qiu, Ruibin Yan, Nan Lang, and Xiang Zhou. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 458–468, 2024.

Appendix A. Normal Text Prompts and CLIP Adaptation Procedure

This appendix describes (i) the light domain adaptation applied to CLIP and (ii) the offline construction of a unified text embedding for MCMIAD.

A.1. Frozen CLIP Encoders and Offline Prompt Embeddings

We adopt an OpenCLIP ViT-B/16 encoder as the visual backbone and its paired text encoder for prompt processing. Before integrating CLIP into MCMIAD, we apply two lightweight steps to improve alignment with medical imaging while keeping most CLIP parameters frozen.

(1) Light Domain Adaptation of CLIP. We collect a small set of clinically normal text descriptions across the BMAD modalities—brain MRI (FLAIR), retina OCT, and liver CT. These prompts describe diagnostically unremarkable anatomy and are used as *positive anchors* in a short contrastive fine-tuning stage on in-domain normal images. Only the projection layers are updated, while the majority of CLIP weights remain frozen. This reduces domain mismatch without sacrificing generalization. The full list of prompts is provided in Section A.2.

(2) Offline Text Embedding Construction. After adaptation, each prompt is encoded by the updated CLIP text encoder and L2-normalized. We average these vectors to obtain a single semantic prior:

$$\mathbf{e}_{\text{normal}} = \frac{1}{15} \sum_{i=1}^{15} \frac{\mathbf{t}_i}{\|\mathbf{t}_i\|_2},$$

which serves as a modality-agnostic reference for normal anatomy. This embedding remains fixed during training and inference. All MCMIAD variants use the same averaged embedding unless otherwise stated.

Motivation. These steps preserve CLIP's broad semantic knowledge while adding medically relevant priors in a parameter-efficient manner. The resulting embedding provides a stable normal reference that improves cross-modal anomaly localization when combined with our denoising decoder.

A.2. Full List of Normal Text Prompts

Below are the exact clinically oriented prompts used for both CLIP adaptation and offline embedding construction.

Brain MRI (T2-FLAIR)

- A FLAIR brain MRI slice of a healthy adult with no tumor, mass effect, or abnormal hyperintense regions.
- A normal brain MRI FLAIR image showing symmetric ventricles and clean white matter without lesions.
- A T2-FLAIR axial brain MRI with intact gray–white matter boundaries and no signs of edema or glioma.
- A neurologically normal brain MRI without hemorrhage, infarction, or abnormal contrast in the parenchyma.
- A diagnostically unremarkable brain MRI scan used as a reference for healthy anatomy.

Retina OCT

- A healthy retinal OCT B-scan with smooth, continuous retinal layers and a normal foveal depression.
- A normal macular OCT image without intraretinal or subretinal fluid, cysts, or hyperreflective foci.
- A high-quality retinal OCT slice showing intact outer retinal layers and a regular retinal pigment epithelium band.
- A clinically normal OCT scan of the macula used as a reference for healthy retinal structure.
- A retinal OCT image from an eye without diabetic macular edema, neovascularization, or drusen.

Liver CT

- A normal contrast-enhanced liver CT slice with homogeneous liver parenchyma and no focal lesions.
- An abdominal CT image where the liver appears smooth, with clear margins and no visible tumors or nodules.

- A healthy liver CT slice showing normal vessel enhancement and no hypodense or hyperdense abnormal areas.
- A diagnostic liver CT scan from a patient without hepatic tumors, cysts, or metastases.
- A standard liver CT image used as a normal reference in medical imaging studies.