TGUMIAD: Text-Guided Unified Model for Medical Image Anomaly Detection

Anonymous submission

Abstract

Accurate anomaly detection in medical imaging is critical for clinical decision-making, yet many methods rely on disease-specific models and extensive labels. We present **TGUMIAD**, a unified vision–language framework that combines a frozen CLIP image encoder and CLIP text encoder with explicit cross-modal fusion and a denoising Transformer decoder to deliver robust, interpretable anomaly detection across retina, brain tumor, and liver tumor benchmarks. Our design emphasizes *human-in-the-loop use*, *explainability* (prompt-guided heatmaps), and *clinical usability* (compact model size and fast inference). Experiments show strong image- and pixel-level AUROC, especially in few-shot settings, indicating practical value when annotated data are scarce. We discuss deployment constraints, fairness/robustness under shifts, and how our interface supports clinician oversight in real workflows.

Introduction

Anomaly detection (AD) in medical imaging is critical for early disease diagnosis and trustworthy clinical decision-making. However, widespread deployment of AD models in practice remains challenging due to their reliance on category-specific architectures and the need for large, annotated datasets. These constraints hinder adaptability and scalability, especially in domains where abnormal cases are rare and expert labeling is costly.

Traditional supervised approaches often require dense, pixel-level annotations and may generalize poorly to unseen pathologies or new imaging modalities. In contrast, unsupervised anomaly detection (UAD) methods, such as memory-based or reconstruction-based models, reduce annotation requirements but frequently suffer from limited semantic interpretability, suboptimal localization, and vulnerability to domain shifts (Bao et al. 2024; Roth et al. 2022; Zavrtanik, Kristan, and Skočaj 2021).

Recently, advances in multimodal vision-language models (VLMs) (Radford et al. 2021) have opened new avenues for cross-modal medical AI by bridging image and text semantics. This enables more interpretable, open-vocabulary, and data-efficient learning. However, directly applying vision-language models to clinical anomaly detection is non-trivial due to domain gaps, insufficient localization ability, and the lack of robust integration with clinical language data (Zhang et al. 2024; Jeong et al. 2023). Moreover, many VLM-based systems are not designed to operate across multiple diseases or imaging modalities, limiting their clinical applicability and scalability.

A central and underexplored challenge in medical anomaly detection is achieving unified and multi-class anomaly detection. Unlike traditional single-disease or single-modality approaches, a unified model aims to identify diverse types of abnormalities across various anatomical regions, disease categories, or imaging techniques within a single, generalizable framework. Such a paradigm is essential for modern healthcare systems, where large-scale multimodal data (e.g., MRI, CT, OCT) is routinely acquired, and operational efficiency precludes deploying and maintaining multiple specialized models.

Despite progress, there remains a clear gap: existing solutions rarely combine efficient cross-modal feature alignment, interpretability, parameter efficiency, and real-world deployability in a single model. Furthermore, for clinical translation, it is crucial that models can adapt to limited annotated data, operate under domain shifts, and provide trustworthy, explainable predictions, especially as AI becomes more deeply integrated into diagnostic workflows.

To address these challenges, we propose TGUMIAD, a unified, multimodal vision-language framework for robust and explainable medical anomaly detection. TGUMIAD fuses multi-scale Vision Transformer (ViT) visual features with CLIP-guided text embeddings, enabling adaptive cross-modal reasoning, robust anomaly localization, and generalization across diverse medical domains with minimal reliance on category-specific tuning or large-scale annotations. By explicitly aligning visual and semantic features, TGUMIAD supports scalable, modality-agnostic, and interpretable anomaly detection suitable for dynamic, real-world clinical scenarios.

We extensively validate TGUMIAD on three heterogeneous benchmarks: retina, brain tumor, and liver tumor datasets, demonstrating consistent improvements over state-of-the-art baselines in both standard and few-shot settings. In addition, we show strong performance in resource-limited and real-world clinical environments, with rapid inference, low computational requirements, and explainable outputs that align with expert annotations.

Beyond technical gains, TGUMIAD addresses key as-

pects valued in clinical AI:

- **Multimodal and multiclass scalability:** One model can be seamlessly applied to multiple diseases and modalities, supporting hospital-wide deployments and reducing maintenance overhead.
- **Data efficiency and practicality:** TGUMIAD performs strongly with limited or few-shot labeled data, enhancing utility in rare disease or emerging epidemic scenarios.
- **Explainability and trustworthiness:** Vision–language alignment and interpretable feature fusion enable clinicians to understand, trust, and validate AI predictions.
- **Ethical and robust design:** We consider model robustness under distribution shift, fairness, and bias, supporting safe, transparent, and ethical deployment in real healthcare systems.

Our main contributions are summarized as follows:

- We propose TGUMIAD, a unified, multimodal vision–language framework that integrates multi-scale Vision Transformer features and CLIP-guided semantic fusion, enabling robust, interpretable, and modality-agnostic anomaly detection and localization across multiple disease categories and imaging modalities.
- We introduce a cross-modal feature alignment module and a frequency–spatial feature perturbation strategy, significantly improving model robustness, fine-grained localization, and generalization under domain shifts and limited supervision.
- We achieve new state-of-the-art performance on three challenging medical datasets (retina, brain tumor, liver tumor), with superior image-level and pixel-level AUROC and a highly compact model size of 3.5M parameters, which is substantially smaller than prior transformer and diffusion-based competitors.
- We provide extensive quantitative and qualitative analyses, including ablation and efficiency studies, validating the benefit of vision–language fusion and the adaptability of TGUMIAD to few-shot and one-shot detection scenarios.
- We discuss clinical and societal impact, ethical considerations, and the path toward integrating TGUMIAD with large foundation models and digital twin systems, supporting future explainable, scalable, and human-centric AI in healthcare.

In summary, TGUMIAD bridges key gaps in medical anomaly detection, offering a unified, explainable, and efficient solution with immediate clinical relevance and a foundation for next-generation, trustworthy AI systems in medicine.

Related Work

Unsupervised Anomaly Detection in Medical Imaging. Unsupervised anomaly detection (UAD) in medical imaging has evolved from traditional reconstruction- and embedding-based paradigms to recent unified and cross-modal frameworks. Early UAD methods leveraged pre-trained CNNs to extract feature embeddings and identified anomalies by measuring deviations from the distribution of normal samples (Roth et al. 2022; Rudolph, Wandt, and Rosenhahn 2021). Memory-based methods such as PatchCore (Roth et al. 2022) use nearest-neighbor matching in latent space, while statistical approaches fit explicit distributions to detect outliers. However, these designs often struggle with domain adaptation and fine-grained localization, particularly in the heterogeneous context of medical imaging.

Synthesis- and Reconstruction-Based Approaches. Synthesis-based methods, including DRAEM (Zavrtanik, Kristan, and Skočaj 2021), improve robustness by generating pseudo-anomalies via noise or texture augmentation. Reconstruction-based models aim to restore input images, flagging anomalies through high reconstruction errors (Liu et al. 2022; Deng and Li 2022). Multi-scale and edge-aware architectures enhance spatial sensitivity, but scalability is limited by category-specific designs and the need for dense annotations.

Vision–Language Models and Cross-Modal AD. Recent advances in vision–language models (VLMs), notably CLIP (Radford et al. 2021), have introduced cross-modal reasoning to anomaly detection, enabling zero-shot and open-vocabulary capabilities. By aligning image and text embeddings, models such as WinCLIP (Jeong et al. 2023) and CLIP Surgery (Li et al. 2023) facilitate semantically-aware anomaly localization. Nevertheless, directly applying VLMs to medical domains often faces challenges due to modality-specific gaps and insufficiently tailored integration strategies (Zhang et al. 2024).

Unified and Transformer-Based Anomaly Detection. To improve scalability and generalization, unified frameworks such as UniAD (You et al. 2022), HVQ-Trans (Lu et al. 2023), and DiAD (He et al. 2024) adopt transformer decoders and quantization-based reconstruction for multi-class and cross-domain anomaly detection. While these approaches advance accuracy and flexibility, many require complex inference or are computationally intensive.

Summary. Despite substantial progress, prior art is still constrained by category-specific tuning, lack of cross-modal flexibility, or insufficient scalability. Our TGUMIAD framework addresses these gaps by unifying Vision Transformer and CLIP-guided embeddings, enabling robust, adaptive, and generalizable anomaly detection across diverse medical imaging domains.

Proposed Framework: TGUMIAD

Architecture Overview

TGUMIAD (Multi-Class Model for Medical Image Anomaly Detection) is a unified vision–language framework for robust and generalizable anomaly detection in medical imaging. As shown in Fig. 1, TGUMIAD integrates a **CLIP image encoder (ViT-B/16, frozen)**, a **CLIP text encoder (frozen)**, an explicit cross-modal alignment module, and a denoising Transformer decoder. This unified pipeline enables adaptive anomaly detection and precise localization across heterogeneous modalities, while minimizing reliance on large labeled datasets or category-specific model designs.

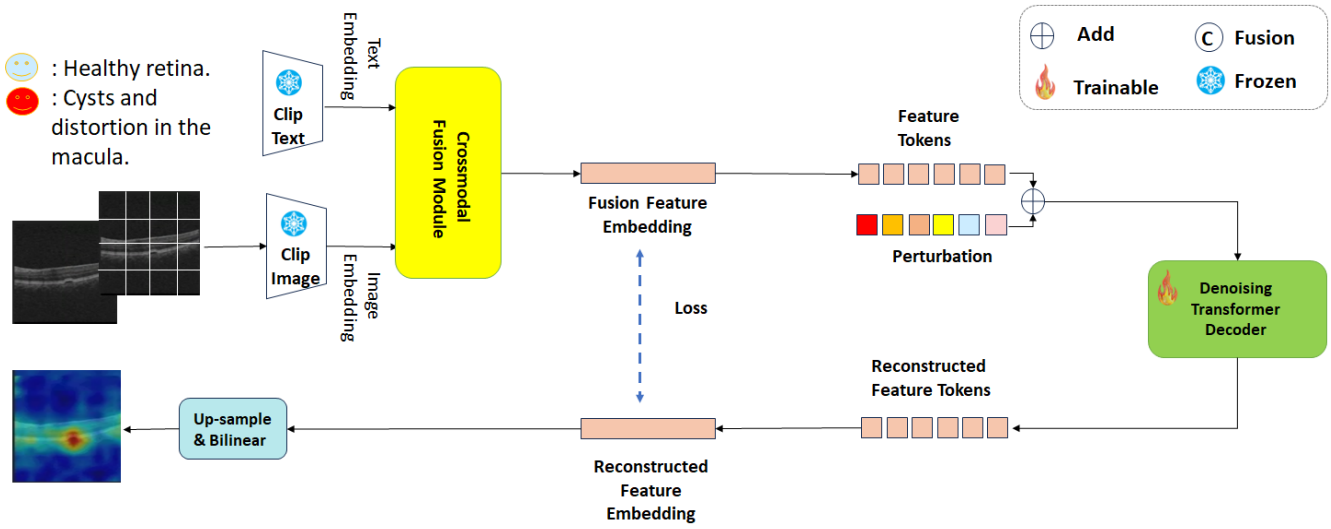

Figure 1: Overall architecture of the proposed TGUMIAD framework.

1. **CLIP Image Encoder (frozen):** extracts patch-level visual tokens from the medical image.
2. **CLIP Text Encoder (frozen):** embeds clinical prompts/descriptions into semantic text tokens.
3. **Cross-Modal Fusion Module:** explicitly aligns image/-text tokens and outputs a fused feature embedding.
4. **Feature Jittering Module:** applies spatial/frequency perturbations to encourage robust, denoised representations.
5. **Denoising Transformer Decoder (trainable):** reconstructs normal feature tokens; a feature-space reconstruction loss is computed between the fused and reconstructed embeddings.
6. **Upsample to Heatmap:** bilinear up-sampling converts reconstruction residuals into a pixel-wise heatmap.
7. **Anomaly Localization & Scoring:** produces pixel-level anomaly maps and stable image-level scores (e.g., max/mean/Top-K), enabling modality-agnostic AD with minimal labels.

Frozen CLIP Encoders and Feature Extraction

During training, the CLIP image and text encoders are kept frozen. Given an input image $\mathbf{I}$, the **CLIP image encoder (ViT-B/16, frozen)** outputs a patch-level feature map $\mathbf{F}_{in} \in R^{C_{org} \times H \times W}$. For each case, a clinically informative text prompt is processed by the CLIP text encoder to yield a semantic embedding. This embedding is fused with the visual feature map to guide the downstream reconstruction. After spatial flattening and linear projection, we obtain the final token sequence:

$$\mathbf{X} \in R^{H \cdot W \times D}. \tag{1}$$

Embedding-Level Fusion

Given visual embeddings $\mathbf{X}_{vis} \in R^{N \times D}$ and textual embeddings $\mathbf{X}_{text} \in R^{C \times D}$, we adopt a cross-attention fusion strategy. Specifically, the visual tokens serve as *queries*, while the text embeddings act as *keys* and *values*:

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

where $\mathbf{Z}$ denotes the cross-modally attended visual representation. The fused features are then refined through layer normalization, self-attention, and an MLP to enhance semantic consistency across modalities. This embedding-level fusion enables adaptive alignment between visual and textual information, supporting robust and generalizable cross-modal anomaly reasoning.

Feature Jittering Module

To enhance robustness against distribution shifts, scanner variability, and acquisition noise, we introduce a **dual-domain feature perturbation module** that randomly applies perturbations in either the spatial or frequency domain during training. In the spatial domain, additive Gaussian noise is injected into the fused tokens. In the frequency domain, the top $q\%$ high-frequency Fourier coefficients (empirically $q = 30\%$) are selectively perturbed using a binary mask:

$$\mathbf{X}_{\text{jitter}} = \begin{cases} \mathbf{X}_{\text{fused}} + \epsilon, & (spatial) \\ \mathcal{F}^{-1}\big(\mathcal{F}(\mathbf{X}_{\text{fused}}) + \epsilon\, I_{\text{high}}\big), & (frequency) \end{cases} \tag{3}$$

where ϵ is zero-mean Gaussian noise and I_{high} denotes the high-frequency mask in the Fourier spectrum. This dual augmentation encourages the model to learn invariance to both spatial-level artifacts and frequency-specific textural shifts.

Denoising Transformer Decoder

The perturbed tokens $\mathbf{X}_{\text{jitter}}$ are passed through a multi-layer denoising Transformer decoder (with self-attention and feed-forward networks (Vaswani et al. 2017)) to reconstruct the expected normal feature distribution:

$$\mathbf{F}_{out} \in R^{C_{org} \times H \times W}. \tag{4}$$

Each decoder block comprises Multi-Head Self-Attention (MHA) and FFN modules, supporting both global context and fine-grained local structure. We also employ local neighborhood-masked attention (following UniAD (You et al. 2022)) to balance fine localization with tractable compute.

Anomaly Localization and Scoring

Anomaly scores are computed by combining pixel-wise reconstruction error and perceptual similarity:

$$\mathbf{S} = \|\mathbf{F}_{in} - \mathbf{F}_{out}\|_2^2 + \lambda \cdot (1 - SSIM(\mathbf{F}_{in}, \mathbf{F}_{out})), \tag{5}$$

where λ weights structural and intensity differences. Image-level anomaly scores are then aggregated (define $K = 5$ unless specified) by:

$$S_{image} = w_1 \cdot \max(\mathbf{S}) + w_2 \cdot mean(\mathbf{S}) + w_3 \cdot TopKMean_K(\mathbf{S}), \tag{6}$$

with w_1, w_2, w_3 as fixed or learnable coefficients.

Summary

TGUMIAD integrates CLIP-based visual and textual encoders with explicit cross-modal alignment, lightweight feature perturbations, and a Transformer-based denoising decoder to reconstruct normal patterns and reveal anomalies. This unified framework delivers robust, interpretable, and generalizable anomaly detection across diverse medical imaging modalities, while reducing dependence on category-specific labels or hand-crafted architectures. We additionally report trainable parameters and FLOPs to reflect deployability in clinical environments, and the design naturally supports user-friendly clinician interaction (e.g., simple editable text prompts and adjustable anomaly masks).

Experiments

Benchmark and Dataset

To comprehensively evaluate the proposed TGUMIAD framework, we adopt the BMAD benchmark (Bao et al. 2024), which provides clinically curated, pixel-annotated datasets for Brain MRI, Liver CT, and Retina OCT. We focus on these three modalities because they include pixel-level ground truth for precise anomaly localization. Each dataset is divided into normal and anomalous subsets following the BMAD protocol. See Table 1 for the main comparison.

We report AUROC as the primary evaluation metric at both the *image* level (detecting whether an image is anomalous) and the *pixel* level (localizing anomalous regions). This dual-level evaluation enables a comprehensive assessment of both overall detection and spatial localization accuracy.

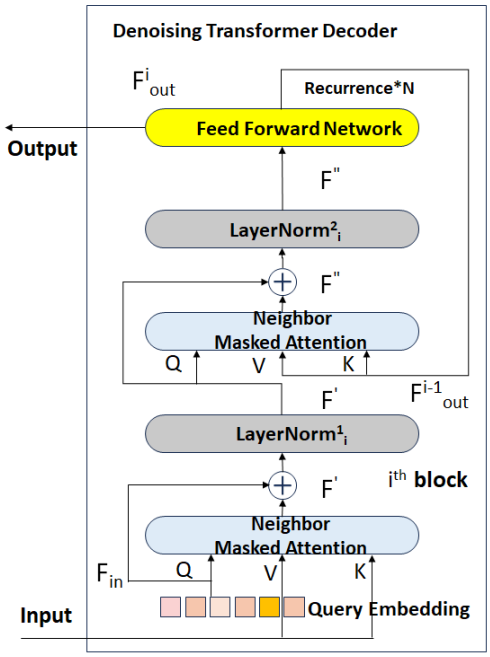

Figure 2: Denoising decoder of the proposed TGUMIAD framework.

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. We adopt the CLIP image encoder (`ViT-B/16`) as the frozen visual backbone and extract patch-level embeddings with the classification head removed. The CLIP text encoder (Transformer-based) is also kept frozen to provide semantic guidance through text prompts.

Cross-Modal Feature Fusion. Visual embeddings from the CLIP image encoder are fused with the corresponding CLIP text embeddings via a cross-attention module, where image tokens serve as queries and text tokens act as keys and values, yielding a unified representation $\mathbf{F}_{fused}$.

Feature Jittering. During training, a dual-domain feature perturbation module randomly applies spatial-domain Gaussian noise or frequency-domain perturbation to the fused tokens. For frequency perturbation, the top $q = 30\%$ high-frequency Fourier coefficients are modified and inverse-transformed, promoting robustness against modality-specific artifacts and scanner variations.

Denoising Transformer Decoder. The fused tokens are reconstructed using a Transformer-based denoising decoder consisting of four layers, eight attention heads, and hidden dimension 128. Each block includes multi-head self-attention and local neighborhood-masked attention, following UniAD (You et al. 2022), with a fixed 7×7 spatial window for efficient local reasoning.

Reconstruction and Scoring. The decoder reconstructs the expected normal feature distribution $\mathbf{F}_{out}$, and pixel-level anomaly maps are computed as a weighted combina-

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

Image-AUROC (%)					
Category	DRAEM	UniAD	SimpleNet	DiAD	TGUMIAD (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	TGUMIAD (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**

tion of MSE and SSIM:

$$\mathbf{S} = \lambda \cdot \|\mathbf{F}_{in} - \mathbf{F}_{out}\|_2^2 + (1-\lambda)\left[1 - SSIM(\mathbf{F}_{in}, \mathbf{F}_{out})\right]. \tag{7}$$

Image-level anomaly scores are aggregated using a weighted combination of maximum, mean, and Top-K mean pixel values (default $K = 5$). AdamW is used with learning rate 2×10^{-4}, weight decay 1×10^{-4}, StepLR scheduler, gradient clipping (max norm 0.1), and we report mean$\pm$std over 3 random seeds.

CLIP Text Prompt Details

Brain MRI

Normal: This is a normal brain MRI scan. No mass, lesion, or abnormal enhancement is seen.

Abnormal: This MRI shows a hyperintense lesion in the right frontal lobe with surrounding edema and midline shift.

Retina OCT

Normal: Healthy retina with continuous and well-organized layers.

Abnormal: Intraretinal cysts and fluid accumulation in the macular region, with disrupted retinal architecture.

Liver CT

Normal: Homogeneous hepatic parenchyma with smooth contours and no visible lesion.

Abnormal: Hypodense mass in segment VI showing irregular borders and capsular retraction.

Experimental Results

Quantitative Comparison with State of the Art

Table 1 presents a comprehensive comparison between **TGUMIAD** and representative UAD baselines, including DRAEM (Zavrtanik, Kristan, and Skočaj 2021), UniAD (You et al. 2022), SimpleNet (Liu et al. 2023), and DiAD (He et al. 2024). Across Retina OCT, Brain Tumor MRI, and Liver Tumor CT, TGUMIAD achieves the highest average Image-AUROC (82.3%) and Pixel-AUROC (96.1%). The largest image-level gain appears on Liver CT, where TGUMIAD surpasses the strongest baseline by +14% Image-AUROC while maintaining precise pixel-level localization.

Qualitative Visualization

Few-Shot and One-Shot Performance

We evaluate the one-shot setting in Table 3. TGUMIAD achieves notable gains for both image- and pixel-level AUROC, especially in Liver Tumor with a +42.4% Image-AUROC improvement over UniAD, highlighting strong sample efficiency.

Ablation Study: Effectiveness of Cross-Modal Fusion

We ablate the CLIP-guided fusion by removing the CLIP text encoder and fusion module. Results in Table 4 show consistent improvements with CLIP, confirming the importance of semantic alignment.

Ethics, Data, and Reproducibility

We use only public, de-identified datasets; no protected health information is accessed. Code, prompts, and configuration files will be released to support reproducibility. We report seeds, one-/few-shot protocols, and metrics, and analyze typical failure cases and potential biases (e.g., modality/site shifts).

Conclusion and Future Work

We introduced **TGUMIAD**, a unified and efficient vision–language framework for medical anomaly detection. TGUMIAD integrates a frozen CLIP image encoder, a CLIP text encoder, explicit cross-modal alignment, and a denoising Transformer decoder to deliver robust and interpretable detection across Retina OCT, Brain MRI, and Liver CT. With only **3.5M** trainable parameters and low FLOPs, the model is practical for clinical deployment; we also report parameters and FLOPs to quantify deployability. The design supports clinician-friendly interaction (simple prompts, editable masks) and shows strong one-/few-shot generalization with state-of-the-art Image-/Pixel-level AUROC. Future work will extend TGUMIAD to temporal inputs and integrate retrieval-augmented reasoning and federated training to address privacy and cross-center domain shifts.

Table 2: Model size vs. AUROC. "Avg." denotes the mean of Image- and Pixel-level AUROC (in %).

Model	Parameters (Millions)	Avg. Image-/Pixel AUROC (%)
TGUMIAD (Ours)	3.5	**89.2**
DiAD (He et al. 2024)	1300	88.2
UniAD (You et al. 2022)	7.7	87.5

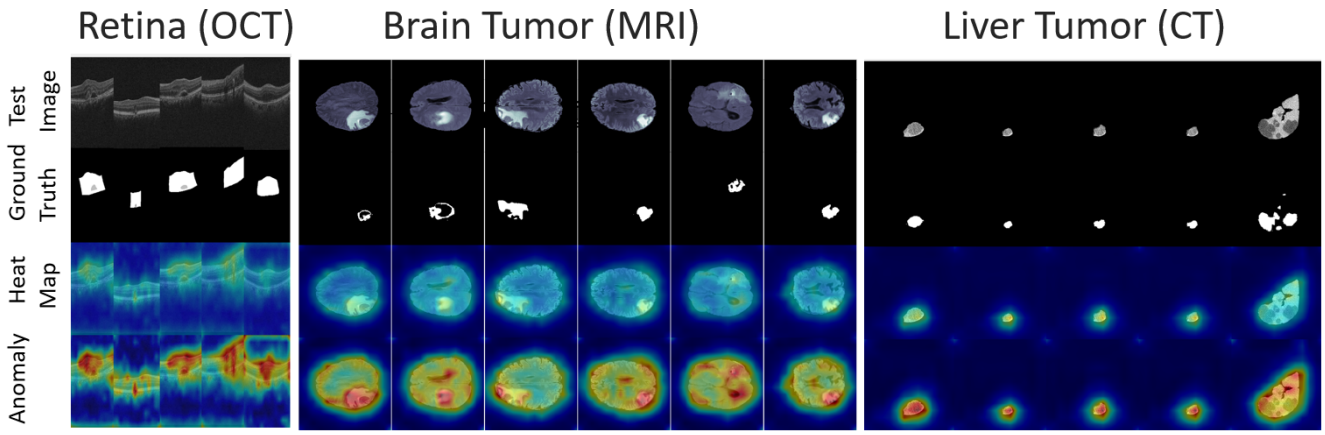

Figure 3: Qualitative visualization across Retina (OCT), Brain Tumor (MRI), and Liver Tumor (CT). TGUMIAD accurately localizes anomalies while minimizing false positives (see also Fig. 2).

References

Bao, J.; Sun, H.; Deng, H.; He, Y.; Zhang, Z.; and Li, X. 2024. Bmad: Benchmarks for medical anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4042–4053.

Deng, H.; and Li, X. 2022. Anomaly Detection via Reverse Distillation From One-Class Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 9737–9746.

He, H.; Zhang, J.; Chen, H.; Chen, X.; Li, Z.; Chen, X.; Wang, Y.; Wang, C.; and Xie, L. 2024. A diffusion-based framework for multi-class anomaly detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 8472–8480.

Jeong, J.; Zou, Y.; Kim, T.; Zhang, D.; Ravichandran, A.; and Dabeer, O. 2023. Winclip: Zero-/few-shot anomaly classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19606–19616.

Li, Y.; Wang, H.; Duan, Y.; and Li, X. 2023. Clip surgery for better explainability with enhancement in open-vocabulary tasks. *arXiv e-prints*, arXiv–2304.

Liu, T.; Li, B.; Zhao, Z.; Du, X.; Jiang, B.; and Geng, L. 2022. Reconstruction from edge image combined with color and gradient difference for industrial surface anomaly detection. *arXiv preprint arXiv:2210.14485*.

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

Lu, R.; Wu, Y.; Tian, L.; Wang, D.; Chen, B.; Liu, X.; and Hu, R. 2023. Hierarchical vector quantized transformer for multi-class unsupervised anomaly detection. *Advances in Neural Information Processing Systems*, 36: 8487–8500.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Roth, K.; Pemula, L.; Zepeda, J.; Schölkopf, B.; Brox, T.; and Gehler, P. 2022. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14318–14328.

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

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

You, Z.; Cui, L.; Shen, Y.; Yang, K.; Lu, X.; Zheng, Y.; and Le, X. 2022. A unified model for multi-class anomaly detection. *Advances in Neural Information Processing Systems*, 35: 4571–4584.

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

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

Dataset	Image-AUROC (UniAD)	Image-AUROC (TGUMIAD)	Pixel-AUROC (UniAD)	Pixel-AUROC (TGUMIAD)
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	78.2	94.8
w/ CLIP	**82.3**	**96.1**

Zhang, X.; Xu, M.; Qiu, D.; Yan, R.; Lang, N.; and Zhou, X. 2024. Mediclip: Adapting CLIP for Few-Shot Medical Image Anomaly Detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 458–468.