CVPR
#22318

MEDIC-AD: Towards Medical Vision-Language Model's Clinical Intelligence

Anonymous CVPR submission

Paper ID 22318

Abstract

Lesion detection, symptom tracking, and visual explainability are central to real-world medical image analysis, yet current medical Vision-Language Models (VLMs) still lack mechanisms that translate their broad knowledge into clinically actionable outputs. To bridge this gap, we present MEDIC-AD*, a clinically oriented VLM that strengthens these three capabilities through a stage-wise framework. First, learnable anomaly-aware tokens (<Ano>) encourage the model to focus on abnormal regions and build more discriminative lesion centered representations. Second, inter-image difference tokens (<Diff>) explicitly encode temporal changes between studies, allowing the model to distinguish worsening, improvement, and stability in disease burden. Finally, a dedicated explainability stage trains the model to generate heatmaps that highlight lesion-related regions, offering clear visual evidence that is consistent with the model's reasoning. Through our staged design,* MEDIC-AD *steadily boosts performance across anomaly detection, symptom tracking, and anomaly segmentation, achieving state-of-the-art results compared with both closed source and medical-specialized baselines. Evaluations on real longitudinal clinical data collected from real hospital workflows further show that* MEDIC-AD *delivers stable predictions and clinically faithful explanations in practical patient-monitoring and decision-support workflows.*

1. Introduction

Vision–Language Models (VLMs) have rapidly evolved [14, 39, 61] from simple image–text tasks such as visual question answering (VQA) [2] and captioning [49] to more advanced capabilities including visual grounding [44, 57] and multi–image reasoning [3, 11, 33]. These advances have inspired the emergence of *Medical Foundation VLMs* [34, 40, 47], which aim to integrate visual and textual medical knowledge for comprehensive diagnostic understanding. Trained on large-scale image-report pairs and multimodal instructions, these models have achieved strong results across tasks such as disease classification, report generation, and Med–VQA, demonstrating the promise of language-driven clinical reasoning.

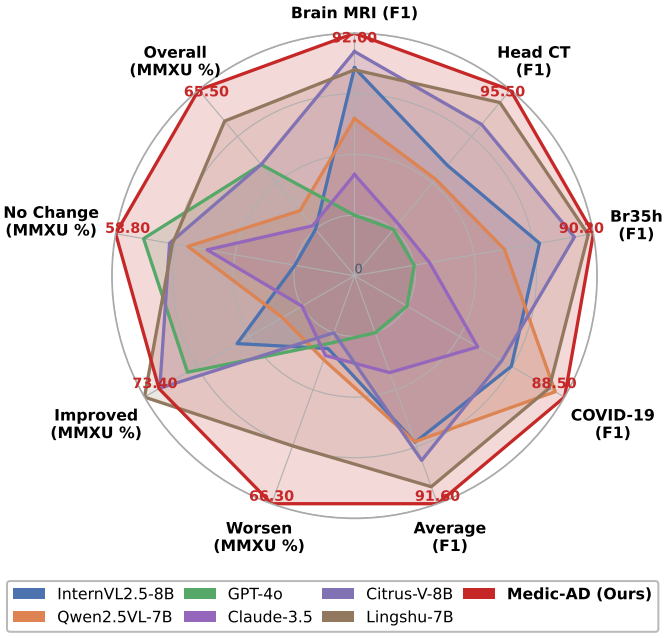


Figure 1. Overall performance of VLMs on Medical Anomaly Detection and Medical Symptom Tracking (MMXU [41]).

However, most Medical Foundation VLMs remain optimized for broad medical knowledge coverage rather than real clinical application [50, 56]. Their training typically relies on long–form captioning, OCR–based instruction tuning, and medical chain–of–thought reasoning that enhance generic reasoning ability, but overlook key properties required for real-world clinical workflows [1]: (1) accurate lesion detection, (2) reliable temporal symptom tracking, and (3) transparent visual explainability of the reasoning process. Addressing these limitations demands a paradigm shift from generalized intelligence toward clinically grounded perception, and understanding.

To that end, we explore three research questions guiding the design of a clinically usable medical VLM.

RQ1: How can lesion and symptom recognition be improved in VLMs for real clinical settings? Even as medical VLMs expand in knowledge, accurate abnormality detection remains essential for safe deployment. We define an *abnormality* as any pathological deviation within an image and propose to enhance this recognition through explicit anomaly–aware representations. By injecting learnable <Ano> tokens into the transformer layers, the model highlights abnormal regions and strengthens its discriminative reasoning. Experiments on brain MRI, head CT, and chest X-ray datasets show that this design achieves strong performance in medical anomaly detection.

RQ2: How can a VLM disentangle temporal medical images to enable more accurate symptom tracking? Existing foundation models that support multi–image inputs typically concatenate visual features, thus failing to capture the temporal progression between scans. To model clinically meaningful changes, we introduce <Diff> tokens that compare anomaly features extracted from multiple images of the same patient. These representations allow the model to reason about whether a condition has worsened, improved, or remained unchanged. On benchmarks such as MMXU [41], which assess longitudinal symptom understanding, our approach achieves superior performance, highlighting its effectiveness as a practical clinical tool for patient monitoring.

RQ3: How can visual explainability be integrated into medical VLM reasoning? Interpretability is indispensable for clinical decision-making. To visually justify model's predictions, we design a heatmap decoder that fuses anomaly features and visual features to generate attention maps highlighting regions responsible for each prediction. These region-level explanations enhance transparency by providing visual evidence for both lesion detection and change assessment, ultimately bridging the gap between black-box reasoning and clinical trust.

These three research directions culminate in **MEDIC-AD**, a stage-wise medical VLM with clinical intelligence, designed to integrate anomaly detection, temporal reasoning, and visual explainability. MEDIC-AD is trained in three stages: *Stage 1: Anomaly Detection.* Learn discriminative abnormality embeddings via injected anomaly tokens, <Ano>, adapting contrastive architecture between normal and abnormal regions for enhancing sensitivity to pathological cues. *Stage 2: Difference Reasoning.* Encode cross-scan variations using <Diff> tokens that disentangle temporal progression of abnormal features and enable fine-grained symptom tracking. *Stage 3: Visual Explainability.* Generate interpretable heatmaps that ground textual outputs on visually abnormal regions, ensuring verifiable reasoning.

Through extensive evaluation, MEDIC-AD consistently demonstrates state-of-the-art (SOTA) performance across diverse medical modalities and tasks as shown in Fig. 1. It outperforms medical foundation models [34, 50, 56], anomaly–specialized models [17, 55], and closed-source counterparts [4, 24] in both lesion detection, and temporal symptom reasoning. Moreover, MEDIC-AD delivers superior visual interpretability generating spatially grounded explanations that align model decisions with clinical evidence. Beyond numerical gains, its stage–wise design encodes the clinical diagnostic workflow—detect, compare, explain—into the model's learning curriculum, transforming general-purpose vision–language understanding into clinically actionable intelligence.

Our main contributions are as follows:

- We present a unified, stage-wise framework (MEDIC-AD) that integrates anomaly detection, longitudinal reasoning, and visual grounding to enable explainable medical inference.
- We introduce anomaly- and difference-token mechanisms that endow medical VLMs with explicit lesion sensitivity and temporal reasoning capability.
- We conduct comprehensive evaluations on multiple medical tasks, as well as on real longitudinal datasets from hospital sites, demonstrating superior reliability and usability compared to both open- and closed-source foundation models, and showcasing deployment readiness for real-world clinical practice workflows.

2. Related Works

2.1. Vision-Language Models for Medical Imaging

Vision Language Models (VLMs) [3, 11, 14, 33, 39, 61] have unified visual and textual reasoning across domains through large-scale contrastive or instruction-tuned learning. Building upon these, medical VLMs have adapted multimodal alignment to clinical imaging and reporting tasks. Early medical VLMs focused on contrastive alignment and report-level representation learning [6, 51], while later instruction-tuned architectures expanded multimodal reasoning through larg scale medical–text pretraining [34, 46, 48, 54]. More recently, Lingshu and Citrus-V [50, 56], built on Qwen-VL 2.5 [3], introduced multi-stage training with shallow/deep alignment, medical instruction tuning, and reinforcement learning with verifiable rewards, achieving state-of-the-art results on single-image medical VQA benchmarks such as SLAKE, PathVQA, VQA-RAD, and OmniMedVQA [19, 22, 32, 48]. These datasets collectively evaluate anatomical localization, factual consistency, and report generation, forming the empirical foundation for single-image medical VLMs. Beyond single-image understanding, medical VLMs have advanced toward difference-aware reasoning, modeling longitudinal changes and dis-

ease progression between paired studies. Generic difference captioning frameworks describe visual changes across image pairs [43, 58], whereas medical-specific longitudinal reasoning integrates anatomical or report-based temporal modeling [12, 21]. A recent unified framework [15] proposes a *Report Generator–Answer Generator (RG–AG)* architecture. While these studies have achieved meaningful progress in understanding longitudinal medical images, they remain largely task-specific, focusing on objectives such as VQA and report generation, without fully leveraging the extensive medical knowledge and generative reasoning capabilities offered by Medical Foundation VLMs.

2.2. Zero-Shot Anomaly Detection

Traditional anomaly detection (AD) methods primarily focused on low-level visual irregularities in industrial datasets such as MVTec-AD [8] and VisA [62], later extending toward zero-shot recognition through text–image alignment. CLIP-based approaches introduced adapter- or prompt-based mechanisms for open-vocabulary detection [10, 23, 60], while Q-Former-based [36] architectures further connected anomaly detection and instruction tuning [17, 55] in general AD tasks. In the medical context, unified benchmarks such as BMAD [7] integrate diverse medical anomaly detection datasets—covering Brain MRI, Chest X-Ray, Liver CT, Retinal OCT, and Pathology—into a single evaluation framework. BMAD consolidates various modality-specific datasets to enable consistent cross-dataset and cross-organ evaluation under a unified protocol. In parallel, chest-specific datasets such as ChestX-Det [38] further provide detailed pixel-level annotations for thoracic disease localization and anomaly segmentation. Collectively, these works mark a shift from handcrafted features to medically grounded anomaly understanding.

2.3. Explainability in Vision–Language Models

Explainability has become a key criterion for assessing the reliability of VLMs, especially in safety-critical domains such as medicine. Recent VLMs utilize cross-attention heatmaps and token-conditioned activations to visualize how linguistic tokens attend to visual regions during reasoning, thereby revealing the internal correspondence between textual semantics and spatial evidence [16, 35, 39, 42]. Such mechanisms improve transparency and enable systematic model auditing and error analysis by linking visual attention patterns with generated textual outputs.

In medical imaging, explainability has been advanced through explicit grounding and concept-level alignment. Grounded VLMs explicitly associate textual rationales with anatomical regions [6, 15, 25], while concept-disentanglement approaches align clinical entities with visual concepts to enhance interpretability and trustworthiness [37]. Building upon these developments, we extend

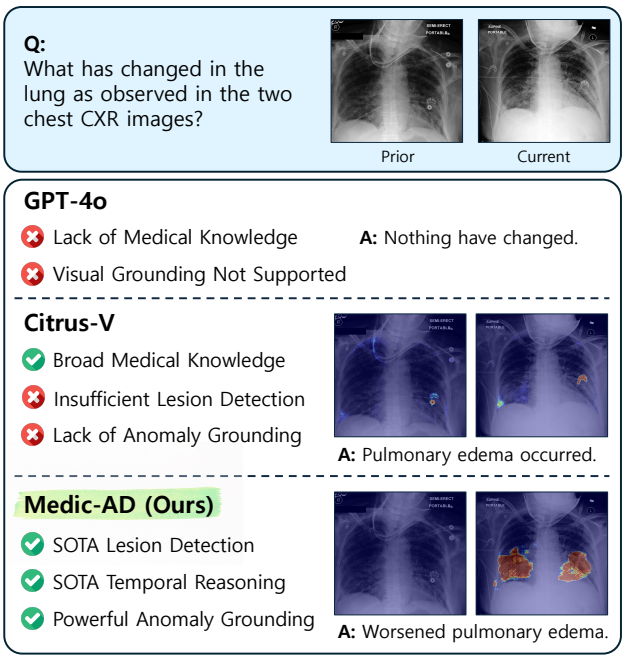


Figure 2. Comparison of VLMs on clinical applications. Medic-AD provides stronger lesion detection, temporal reasoning, and visual grounding than GPT-4o [24] and Citrus-V [50].

visual grounding beyond a mere auxiliary visualization tool. In our medical VLM, interpretability is achieved by grounding the reasoning process on anomalous features such as lesions or symptoms, thereby providing explicit visual evidence that supports the model's clinical conclusions and ensuring clinically verifiable visual explainability.

3. Methodology

3.1. Overview

Standard VLMs encode an input image $\mathbf{I}$ and textual instruction $\mathbf{T}$ into a joint multimodal sequence to generate a response $\mathbf{R}$:

$$\mathbf{R} = f_l([\, f_p(\mathbf{V})\,;\, \mathrm{Emb}(\mathbf{T})\,]), \tag{1}$$

where $f_p(\cdot)$ denotes a visual projection layer that maps visual features $\mathbf{V}$, extracted from $\mathbf{I}$ via a vision encoder, into the text embedding space, and $f_l(\cdot)$ represents the large language model (LLM). While this general formulation supports broad multimodal reasoning, it lacks the inductive biases required for clinically meaningful perception including reliable lesion localization, temporal tracking, and visual justification, all of which are essential for trustworthy decision support.

MEDIC-AD extends this paradigm into a clinically grounded reasoning framework through a stage-wise optimization pipeline, composed of three progressive stages where each stage incrementally enhances the model's reasoning capability. Stage 1 learns anomaly-aware represen-

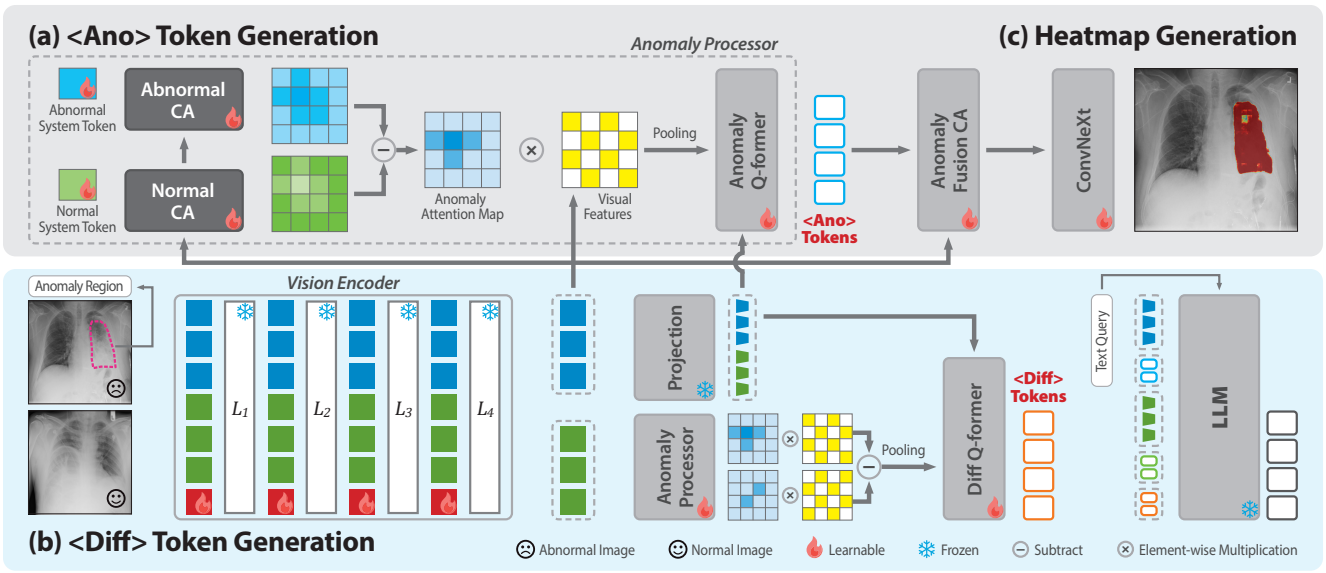


Figure 3. Architecture of MEDIC-AD. (a) Stage 1: <Ano> Token Generation, (b) Stage 2: <Diff> Token Generation, and (c) Stage 3: Heatmap Generation illustrate each stage of the proposed framework. Note that CA denotes Cross-Attention.

tations that encode lesion-specific semantics. Stage 2 builds on these representations to disentangle temporal variations between prior and current studies, yielding difference tokens. Stage 3 introduces grounding supervision that aligns the learned anomaly features with spatial heatmaps, enabling visually verifiable predictions.

We first introduce anomaly-aware tokens, <Ano>, derived from a cross-attention mechanism applied to the visually enhanced feature representation $\mathbf{V}^*$. Here, $\mathbf{V}^*$ denotes the anomaly-augmented visual features obtained by incorporating *visual soft prompts* [26] into the original visual embeddings $\mathbf{V}$. This process encourages the model to emphasize lesion-relevant regions and discriminative cues. The resulting tokens are concatenated with the visual and textual embeddings as

$$\mathbf{R} = f_l([\, f_p(\mathbf{V}^*)\,;\, \texttt{<Ano>}\,;\, \mathrm{Emb}(\mathbf{T})\,]). \tag{2}$$

Next, to model temporal changes between prior and current images, our model learns anomaly-aware representations across time and produces <Diff> tokens. Given two input images, their corresponding anomaly-augmented visual features, $\mathbf{V}_1^*$ and $\mathbf{V}_2^*$, are extracted by the vision encoder and formulated as

$$\mathbf{R} = f_l([\, f_p(\mathbf{V}_1^*)\,;\, \texttt{<Ano>}\,;\, f_p(\mathbf{V}_2^*)\,;\, \texttt{<Ano>}\,;\, \mathrm{Emb}(\mathbf{T})\,;\, \texttt{<Diff>}\,]), \tag{3}$$

following the modified chat template as illustrated in Appendix Sec. A. Finally, the anomaly-aware tokens, <Ano>, are fed into a heatmap decoder $f_h(\cdot)$ together with the corresponding visual features $\mathbf{V}^*$ to generate grounding maps $\mathbf{M}$. These maps provide region-level visual evidence that supports and justifies textual predictions as

$$\mathbf{M} = f_h([\, f_p(\mathbf{V}^*)\,;\, \texttt{<Ano>}\,]). \tag{4}$$

3.2. Architecture and Stage-wise Training

In this section, we present the detailed architecture of MEDIC-AD and the corresponding training pipelines designed to implement the framework illustrated in Sec. 3.1. Each stage progressively enhances the capability of the baseline Medical Foundation VLM, Lingshu [56], which is a strong backbone model pretrained on large-scale medical data, by introducing specialized modules for anomaly reasoning, temporal difference analysis, and visual explainability.

Stage 1: Anomaly Detection. The first stage focuses on training an *Anomaly Processor* that produces <Ano> tokens, compact latent representations capturing lesion-related semantics, as illustrated in Fig. 3 (a). These tokens are constructed through two learnable system tokens, the *Abnormal System Token* and *Normal System Token*. They interact with multi-scale visual features extracted from four intermediate layers of the vision encoder via a cross-attention mechanism, producing *Abnormal* and *Normal Attention Scores* for each visual patch.

To preserve the pretrained vision encoder's representational stability while adapting it for anomaly detection, we adopt *Visual Soft Prompt Tuning* [26] to the selected four layers instead of fully updating their parameters. Unlike conventional attention head using Softmax normalization, our design applies Sigmoid activation to obtain patch-wise anomaly probabilities. The difference between abnormal

and normal attention weights yields an *Anomaly Attention Map*, which reflects the likelihood of each patch being abnormal. This map modulates the original visual features through element-wise multiplication, adjusting their magnitudes according to anomaly salience.

Subsequently, 2D global pooling is performed over the modulated visual features to derive *Anomaly Queries*. These queries are passed through an *Anomaly Q-Former*, where the LLM-projected visual tokens act as keys and values. The Anomaly Q-Former outputs are then fed into a 2-layer MLP to yield the final Anomaly-aware tokens, <Ano>, which are used in both downstream LLM inference and later heatmap generation in Stage 3.

Training for this stage utilizes a diverse collection of medical anomaly datasets spanning MRI, X-ray, and CT modalities, including **BMAD**, **ChestX-Det** [7, 38], as well as multimodal VQA datasets such as **SLAKE**, **PathVQA**, and **VQA-RAD** [19, 32, 48], ensuring both robust visual grounding and generalizable medical reasoning capability.

Stage 2: Difference Reasoning. The second stage focuses on modeling inter-image differences to analyze disease progression over time. Here, the goal is to learn difference tokens, <Diff>, that *disentangle* the variations in abnormal regions across time or paired studies (e.g., follow-up vs. baseline scans), effectively separating genuine pathological progression from visual or acquisition-related noise.

As shown in Fig. 3 (b), the modulated visual features derived in Stage 1 from two images are contrasted and disentangled through a *Diff Q-Former*, which isolates lesion-specific change patterns. Then, each image's projected visual tokens, $f_p(\mathbf{V_1^*})$ and $f_p(\mathbf{V_2^*})$, serves as keys and values to encode structured inter-image relationships. Passing these through the Diff Q-Former and a subsequent 2-layer MLP yields the difference tokens, <Diff>, which are appended to the multimodal input sequence. By explicitly isolating temporal anomalies from static visual context, this stage enables the LLM to reason over fine-grained temporal variations in lesion appearance or intensity, thereby enhancing its ability for longitudinal disease reasoning.

Stage 2 training requires temporally paired or longitudinal datasets. We use **MIMIC-Diff-VQA**[21], a dataset built for multi-image reasoning in clinical follow-up scenarios, allowing the model to learn spatial correspondence and temporal progression patterns in real patient studies.

Stage 3: Visual Explainability. The final stage introduces a heatmap generation module designed to achieve visual grounding and enhance interpretability. While prior Medical Foundation VLMs often rely on pretrained vision decoders (e.g., SAM2 [45] used in Citrus-V [50]), our approach leverages the learned <Ano> tokens and ConvNeXt-based [52] segmentation head to directly link visual reasoning with interpretable evidence.

As illustrated in Fig. 3 (c), we fuse <Ano> tokens with the vision encoder's intermediate feature maps via a fusion block, reinforcing the model's focus on lesion-relevant regions driving the LLM's prediction. The fused features are then processed by a compact ConvNeXt-based segmentation head to generate a heatmap $\mathbf{M}$ spatially aligned with the input image. This heatmap is overlaid on the original image to provide region-level visual evidence that supports the textual output, thereby connecting model reasoning with clinically observable cues.

Stage 3 is trained on datasets with pixel-level segmentation masks, such as selected subsets of **BMAD** and **ChestX-Det** [7, 38]. Leveraging anomaly-token-guided fusion, our model achieves substantially improved anomaly-localization accuracy compared with recent grounding-based medical VLMs, as demonstrated in Sec. 4.3. For the more detailed training configurations of each stages, please see Appendix Sec. B.

4. Experiments

To validate the effectiveness of MEDIC-AD, we conduct a series of experiments corresponding to each stage of the proposed framework introduced in Sec. 3. Each stage is validated on a task specifically aligned with its objective. Stage 1 evaluates the discriminative capability of the learned <Ano> tokens via **Medical Zero-shot Anomaly Detection**. Stage 2 assesses temporal reasoning through the **MMXU Benchmark** for medical symptom tracking. Finally, Stage 3 examines **Medical Visual Explainability**, evaluating region-level grounding and the consistency between visual evidence and textual reasoning using segmentation-based metrics.

As this section focuses on the stage-specific tasks central to our framework, evaluations on conventional medical VQA benchmarks (**VQA RAD** [32], **SLAKE** [48], **PathVQA** [19], and **MMMU Med** [59]) are provided in Appendix Sec. C, demonstrating that MEDIC-AD successfully preserves general medical knowledge of backbone while simultaneously reinforcing its stage-wise clinical applicability.

4.1. Medical Zero-shot Anomaly Detection

Experimental Settings. In the zero-shot anomaly detection setting, the model is tested on datasets that are entirely unseen during training to ensure a fair comparison with competing models. We evaluate across four heterogeneous modalities—**Brain MRI**, **Head CT**, **Br35h**, and **COVID-19 X-ray** [13, 18, 28, 30]—to assess generalization capability. Each test sample is queried with a consistent instruction prompt: *"Is there any abnormality in this image?"* The

Table 1. Results on Zero-shot Medical Anomaly Detection (Brain MRI [28], Head CT [30], Br35h [18], and COVID-19 [13]). For each dataset, we report Precision, Recall, and F1. The rightmost column shows the average F1 over all tasks.

Category	Model	Size	Brain MRI			Head CT			Br35h			COVID-19			Avg. F1
			Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	
General	Qwen2.5-VL	7B	76.1	92.9	83.6	61.6	100.0	76.3	66.4	94.5	78.0	88.6	83.4	85.9	81.0
	InternVL2.5	8B	81.2	97.4	88.6	67.6	96.0	79.3	71.2	99.0	82.8	94.8	60.2	73.6	81.1
Closed	GPT-4o	–	59.3	98.6	74.1	48.7	100.0	65.5	48.9	99.9	65.6	29.1	92.9	44.4	62.4
	Claude-3.5	–	64.0	100.0	78.1	50.0	100.0	66.7	51.2	100.0	67.7	47.2	100.0	64.2	69.2
Anomaly	AnomalyGPT	13B	61.0	96.8	74.8	50.8	97.0	66.7	47.2	88.6	61.6	33.9	57.2	42.5	61.4
	Anomaly-OV	7B	53.1	71.6	61.0	39.8	66.0	49.6	42.2	73.0	53.5	29.6	47.0	36.3	50.1
Medical	LLaVA-MED	7B	69.0	95.5	80.1	54.1	77.9	63.8	60.8	92.7	73.4	46.1	86.3	60.1	69.4
	Citrus-V	8B	85.5	95.5	90.2	78.7	100.0	88.1	79.5	97.5	87.6	58.4	90.4	70.9	84.2
	Lingshu	7B	83.1	94.3	88.4	91.4	94.1	92.8	86.0	93.3	89.5	84.0	84.4	84.2	88.7
	MEDIC-AD	7B	91.1	92.9	**92.0**	95.1	96.0	**95.5**	90.8	89.5	**90.2**	96.6	81.6	**88.5**	**91.6**

Table 2. Results on MMXU [41]. Models are categorized into general-purpose, closed-source, and medical-domain VLMs.

Category	Model	Size	Worsen	Improved	No Change	Overall (↑)
General	InternVL2.5	8B	0.486	0.607	0.402	0.498
	Qwen2.5-VL	7B	0.499	0.545	0.513	0.519
Closed	Claude-3.5	–	0.494	0.518	0.493	0.502
	GPT-4o	–	0.480	0.675	0.559	0.571
Medical	Citrus-V	8B	0.468	0.713	0.532	0.571
	Lingshu	7B	0.597	**0.734**	0.529	0.620
	MEDIC-AD (Ours)	7B	**0.663**	0.714	**0.588**	**0.655**

model's binary response is evaluated using the F1 score to capture both precision and recall performance.

We compare MEDIC-AD against a range of baselines spanning three categories: (1) **General-purpose open-source VLMs:** Qwen2.5-VL-7B and InternVL2.5-8B [3, 11], (2) **Closed-source models:** GPT-4o and Claude-3.5 [4, 24], and (3) **Anomaly Detection Specialized VLMs:** AnomalyGPT and Anomaly-OV [17, 55], as well as (4) **Medical Foundation VLMs:** LLaVA-Med-7B, Citrus-V-8B, and Lingshu-7B [34, 50, 56].

Results. As summarized in Table 1, MEDIC-AD achieves **SOTA performance** across all four datasets, demonstrating superior generalization to unseen medical imaging modalities and conditions. The results indicate that the learned anomaly-aware tokens, `<Ano>`, effectively capture lesion-relevant features, enabling reliable and robust zero-shot abnormality discrimination even in unseen datasets. Notably, MEDIC-AD surpasses all medical-specialized baselines and even outperforms closed-source models, validating the efficacy of its anomaly representation learning. While some closed-source models (e.g., GPT-4o) occasionally produce conservative outputs for normal cases due to safety-alignment bias [29, 53] (e.g., responding with *"I'm unable to analyze medical images ..."*), MEDIC-AD consistently produces well-calibrated predictions across both normal and abnormal cases, indicating greater clinical reliability and decision consistency. Overall, Stage 1 demonstrates that anomaly representation learning offers a reliable foundation for zero-shot medical abnormality detection.

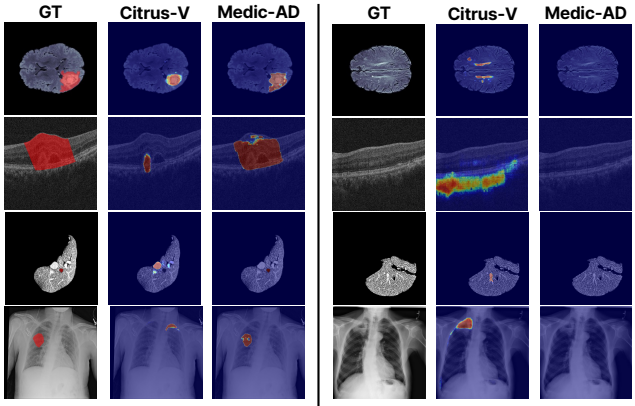
4.2. Medical Symptom Tracking

Experimental Settings. We evaluate temporal reasoning ability using the **MMXU** benchmark [41], which involves paired chest X-ray studies from the same patient. For each instance, the model must classify disease progression as *worsened*, *improved*, or *unchanged* based on two images and a multiple-choice question such as: *"What is the condition of the left lower lung zone across the two chest CXR images? A: No significant change, B: Improved, C: Worsened."* Only models supporting multi-image reasoning are included in this comparison, and the evaluation metric follows the accuracy of categorical predictions derived from the model's textual responses.

Results. As shown in Tab. 2, MEDIC-AD demonstrates clear advantages over existing multimodal models on the MMXU benchmark. By leveraging the `<Diff>` tokens representations disentangled through the Diff Q-Former, the model effectively captures localized lesion changes while being robust to irrelevant appearance variations such as illumination or positional shifts. Unlike general VLMs that

Table 3. Visual grounding performance on BMAD [7] (BraTS2021 [5], RESC [20], BTCV + LiTs [9, 31]) and ChestX-Det [38] datasets. Each dataset reports AUC and mIoU metrics (higher is better).

Model	BraTS2021		RESC		BTCV + LiTs		ChestX-Det	
	AUC (↑)	mIoU (↑)	AUC (↑)	mIoU (↑)	AUC (↑)	mIoU (↑)	AUC (↑)	mIoU (↑)
Citrus-V	98.8	32.6	87.6	2.1	82.1	1.7	98.0	12.4
Medic-AD (Ours)	**99.8**	**87.6**	**100**	**97.2**	**97.2**	**83.6**	**99.8**	**79.8**



(a) Abnormal samples (b) Normal samples

Figure 4. Visual Grounding comparison between MEDIC-AD and Citrus-V [50] on diverse abnormal and normal samples.

simply concatenate multiple images, MEDIC-AD explicitly encodes inter-image relationships, yielding consistent reasoning about temporal dynamics. In particular, qualitative inspection shows that MEDIC-AD highlights pathological regions with true clinical changes (e.g., consolidation growth or opacity reduction), whereas other models often mistake global contrast shifts for disease progression (Fig.2). These findings indicate that Stage 2 effectively separates clinically relevant temporal changes.

4.3. Medical Visual Explainability

Experimental Settings. To assess visual grounding and interpretability, we evaluate MEDIC-AD on medical datasets that include pixel-level anomaly masks, specifically a subset of **BMAD** [7] (BraTS2021 [5], RESC [20], and BTCV + LiTs [9, 31]) and the **ChestX-Det** [38] dataset. For each image, the model generates a heatmap using the same anomaly-detection query as in Sec. 4.1. Predicted heatmaps are compared with ground-truth masks using AUC and mIoU. We compare against **Citrus-V** [50], the most recent medical VLMs supporting visual grounding.

Results. MEDIC-AD shows consistently strong performance across datasets, outperforming Citrus-V on both AUC and mIoU (Tab. 3). By integrating <Ano> tokens with intermediate visual features, MEDIC-AD produces heatmaps that more accurately localize pathological regions aligned with the model's textual rationale. In contrast, Citrus-V, which use a SAM2 decoder [45], tends to produce less precise masks, sometimes highlighting non-lesion areas or yielding diffuse activations in normal images. Representative examples in Fig. 4 highlight these differences. These results show that Stage 3 improves the alignment between visual evidence and model reasoning, yielding more interpretable and clinically coherent responses.

5. Analysis

In this section, we present a comprehensive analysis of the proposed MEDIC-AD framework through ablation studies, hyperparameter sensitivity experiments, and real-world evaluations. Across all studies, the results consistently support the central claim of this work: temporal reasoning in medical image pairs fundamentally benefits from coherent integration of anomaly-aware spatial cues and temporally grounded difference representations.

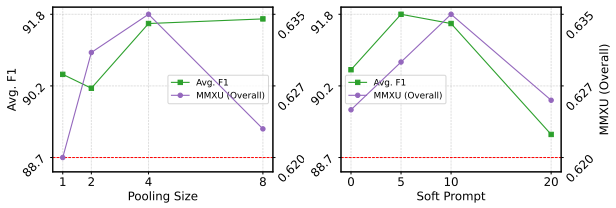
5.1. Effect of <Ano> Tokens on Temporal Reasoning

The first analysis focuses on how <Ano> tokens contributes to constructing reliable temporal representations. In MEDIC-AD, the construction of <Diff> tokens is grounded in the anomaly-aware visual features generated during the <Ano> tokens estimation process, where patch-wise anomaly likelihood modulates the magnitude of visual representations (Sec. 3.2). To isolate the role of <Ano> tokens, we generate <Diff> tokens using the unmodified, original visual features without salience adjustment.

As reported in Tab. 4 (*Effect of <Ano> Tokens*), removing <Ano> tokens consistently degrades performance across tasks. The drop reveals two important observations. First, anomaly-aware magnitude modulation enhances the expressiveness of the visual embeddings used for temporal differencing, allowing <Diff> tokens to capture clinically relevant cues rather than global appearance changes. Second, incorporating <Ano> tokens during inference adds complementary contextual information that stabilizes the interpretation of new findings. Together, these results show that <Ano> and <Diff> tokens form a mutually reinforcing pair: one grounds spatial anomaly cues, while the other captures temporal evolution, and both are required for accurate modeling of medical image progression.

Table 4. Effect of utilizing <Ano> tokens and comparison of visual feature extraction strategies.

	<Ano>	Avg. F1 (↑)	MMXU (↑)
Effect of <Ano> Tokens			
<Diff> tokens only	×	–	0.635
MEDIC-AD	✓	–	**0.655**
Feature Selection Strategy			
Last layer	✓	90.9	0.619
Intermediate 4-layers	✓	**91.6**	**0.635**



(a) (b)

Figure 5. Hyperparameter sensitivity analysis on (a) query token pooling size and (b) visual soft prompt counts. The red line denotes the baseline performance of Lingshu [56].

5.2. Layer Selection in Visual Feature Extraction

In generating both <Ano> and <Diff> tokens, MEDIC-AD relies on visual features extracted from the vision encoder of the backbone VLM. To understand how visual representations influence anomaly and temporal token construction, we compare two configurations: using only the last-layer visual features, and aggregating intermediate-layer features together with the final representation.

As shown in Tab. 4 (*Feature Selection Strategy*), using intermediate-layer features consistently outperforms relying solely on the final hidden state. This advantage echoes prior findings [10, 23, 55, 60] that multi-level feature aggregation provides a broader range of semantic cues and improves downstream performance. In our setting, incorporating intermediate features not only enriches the anomaly representations but also produces feature embeddings that are more stable and informative for computing inter-image differences in temporal reasoning.

5.3. Impact of Hyperparameters

MEDIC-AD involves two key hyperparameters: (1) the number of generated <Ano> tokens and <Diff> tokens, and (2) the number of soft prompts injected into the vision encoder. The number of <Ano> and <Diff> tokens is implicitly determined by the 2D pooling size applied to the magnitude-adjusted visual features used for token generation. Therefore, we investigate the effect of varying the pooling size. In parallel, the number of soft prompts influences how the extracted visual features are adapted to MEDIC-AD while also affecting the overall performance of the backbone VLM, motivating a sensitivity analysis.

As illustrated in Fig. 5 (a), the model achieves consistently strong performance on both Anomaly Detection and MMXU benchmarks when the query-token pooling size is set to 4×4, indicating that this granularity provides a favorable balance between spatial abstraction and anomaly localization. Similarly, the soft prompt sensitivity study in Fig. 5 (b) demonstrates that using 10 visual soft prompts yields the most stable and competitive results. We therefore adopt a pooling size of 4×4 and 10 soft prompts as the default configuration for MEDIC-AD.

Table 5. Temporal image captioning performance on a real-world clinical dataset of 300 patients. Scores (0–1) indicate similarity to ground-truth descriptions, evaluated by GPT-4o [24].

Model	GPT-4o eval
Lingshu-7B	0.177
Medic-AD (Ours)	0.291

5.4. Real-world Application

To validate the applicability of MEDIC-AD beyond public benchmarks, we conduct a real-world clinical study using chest X-ray pairs collected from 300 patients who visited a hospital for follow-up examinations. For each image pair, radiologists provide structured annotations describing the presence or absence of specific clinical findings, and the degree of change compared to the previous examination. Using this dataset, we formulate a temporal-difference captioning task in which the model must generate clinically consistent descriptions of symptom progression.

As shown in Tab. 5, MEDIC-AD outperforms Lingshu [56]—the strongest baseline in temporal reasoning (Tab. 2)—under GPT evaluation. These results indicate that MEDIC-AD remains effective not only in controlled benchmarks but also demonstrates robustness and reliability on real-world clinical data. Moreover, the model produces descriptions that align closely with expert assessments, highlighting its potential for integration into clinical workflows that demand interpretable and accurate temporal reasoning.

6. Conclusion

We introduced **MEDIC-AD**, a stage-wise medical VLM that strengthens clinical intelligence: lesion detection, temporal reasoning, and visual explainability through anomaly-aware and difference-token mechanisms. Our unified design enables the model to focus on abnormality cues, capture clinically meaningful changes between images, and provide accurate grounded visual evidence. Furthermore, evaluations on real-world hospital cases show robust alignment with expert assessments, indicating that MEDIC-AD offers a practical and reliable application for clinically usable medical VLMs.

References

[1] Uwa O Aideyan, Kevin Berbaum, and Wilbur L Smith. Influence of prior radiologic information on the interpretation of radiographic examinations. *Academic Radiology*, 2(3): 205–208, 1995. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *CVPR*, pages 2425–2433, 2015. 1

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 2, 6

[4] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022. 2, 6

[5] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, et al. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv preprint arXiv:2107.02314*, 2021. 7

[6] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *CVPR*, pages 15016–15027, 2023. 2, 3

[7] Jinan Bao, Hanshi Sun, Hanqiu Deng, Yinsheng He, Zhaoxiang Zhang, and Xingyu Li. Bmad: Benchmarks for medical anomaly detection. *arXiv preprint arXiv:2306.11876*, 2023. 3, 5, 7

[8] Paul Bergmann, Michael Fauser, David Sattlegger, and Carsten Steger. The mvtec anomaly detection (mvtec ad) dataset: A comprehensive real-world dataset for unsupervised anomaly detection. Technical report, MVTec Software GmbH, 2021. 3

[9] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *MIA*, 84:102680, 2023. 7

[10] Yunkang Cao, Jiangning Zhang, Luca Frittoli, Yuqi Cheng, Weiming Shen, and Giacomo Boracchi. Adaclip: Adapting clip with hybrid learnable prompts for zero-shot anomaly detection. *arXiv preprint arXiv:2407.15795*, 2024. 3, 8

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 2, 6

[12] Yeongjae Cho, Taehee Kim, Heejun Shin, Sungzoon Cho, and Dongmyung Shin. Pretraining vision-language model for difference visual question answering in longitudinal chest x-rays. *arXiv preprint arXiv:2402.08966*, 2024. 3

[13] Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, Mamun Bin Ibne Reaz, and Mohammad Tariqul Islam. Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8:132665–132676, 2020. 5, 6

[14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36: 49250–49267, 2023. 1, 2

[15] Francesco Dalla Serra, Patrick Schrempf, Chaoyang Wang, Zaiqiao Meng, Fani Deligianni, and Alison Q. O'Neil. Grounding chest x-ray visual question answering with generated radiology reports. *arXiv preprint arXiv:2505.16624*, 2025. 3

[16] Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting clip's image representation via text-based decomposition. *arXiv preprint arXiv:2310.05916*, 2023. 3

[17] Zhaopeng Gu, Bingke Zhu, Guibo Zhu, Yingying Chen, Ming Tang, and Jinqiao Wang. Anomalygpt: Detecting industrial anomalies using large vision-language models. In *AAAI*, pages 1932–1940, 2024. 2, 3, 6

[18] Ahmed Hamada. Br35h: Brain tumor detection 2020, 2020. 5, 6

[19] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 2, 5, 1

[20] Junjie Hu, Yuanyuan Chen, and Zhang Yi. Automated segmentation of macular edema in oct using deep neural networks. *MIA*, 55:216–227, 2019. 7

[21] Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*, pages 1–16. ACM, 2023. 3, 5

[22] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *CVPR*, pages 22170–22183, 2024. 2

[23] Chaoqin Huang, Aofan Jiang, Jinghao Feng, Ya Zhang, Xinchao Wang, and Yanfeng Wang. Adapting visual-language models for generalizable anomaly detection in medical images. In *CVPR*, 2024. 3, 8

[24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3, 6, 8

[25] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q.H. Truong, Du Du Nguyen Duong, Tan Bui, Pierre Chambon,

Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021. 3

[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *ECCV*, pages 709–727. Springer, 2022. 4

[27] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577, 2019. 1

[28] Pranita Balaji Kanade and PP Gumaste. Brain tumor detection using mri images. *Brain*, 3(2):146–150, 2015. 5, 6

[29] Jaeik Kim, Woojin Kim, Woohyeon Park, and Jaeyoung Do. Mmpb: It's time for multi-modal personalization. *arXiv preprint arXiv:2509.22820*, 2025. 6

[30] Felipe Campos Kitamura. Head ct - hemorrhage, 2018. 5, 6

[31] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, page 12, 2015. 7

[32] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2, 5, 1

[33] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 2

[34] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *NeurIPS*, 36:28541–28564, 2023. 1, 2, 6

[35] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh D. Gotmare, Shafiq Joty, Caiming Xiong, and Steven C.H. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *arXiv preprint arXiv:2107.07651*, 2021. 3

[36] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, pages 19730–19742. PMLR, 2023. 3

[37] Tang Li, Mengmeng Ma, and Xi Peng. Deal: Disentangle and localize concept-level explanations for vlms. *arXiv preprint arXiv:2407.14412*, 2024. 3

[38] Jie Lian, Jingyu Liu, Shu Zhang, Kai Gao, Xiaoqing Liu, Dingwen Zhang, and Yizhou Yu. A structure-aware relation network for thoracic diseases detection and segmentation. *IEEE Transactions on Medical Imaging*, 40(8):2042–2052, 2021. 3, 5, 7

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 1, 2, 3

[40] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 1

[41] Linjie Mu, Zhongzhen Huang, Shengqian Qin, Yakun Zhu, Shaoting Zhang, and Xiaofan Zhang. Mmxu: A multi-modal and multi-x-ray understanding dataset for disease progression. *arXiv preprint arXiv:2502.11651*, 2025. 1, 2, 6

[42] Xichen Pan, Li Dong, Shaohan Huang, Zhiliang Peng, Wenhu Chen, and Furu Wei. Kosmos-g: Generating images in context with multimodal large language models. *arXiv preprint arXiv:2310.02992*, 2023. 3

[43] Yue Qiu, Shintaro Yamamoto, Kodai Nakashima, Ryota Suzuki, Kenji Iwata, Hirokatsu Kataoka, and Yutaka Satoh. Describing and localizing multiple changes with transformers. In *ICCV*, pages 1971–1980, 2021. 3

[44] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *CVPR*, pages 13009–13018, 2024. 1

[45] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5, 7

[46] Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025. 2

[47] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. 1

[48] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024. 2, 5, 1

[49] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015. 1

[50] Guoxin Wang, Jun Zhao, Xinyi Liu, Yanbo Liu, Xuyang Cao, Chao Li, Zhuoyun Liu, Qintian Sun, Fangru Zhou, Haoqiang Xing, et al. Citrus-v: Advancing medical foundation models with unified medical image grounding for clinical reasoning. *arXiv preprint arXiv:2509.19090*, 2025. 1, 2, 3, 5, 6, 7

[51] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *EMNLP*, page 3876, 2022. 2

[52] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *CVPR*, pages 16133–16142, 2023. 5

[53] Tinghao Xie, Xiangyu Qi, Yi Zeng, Yangsibo Huang, Udari Madhushani Sehwag, Kaixuan Huang, Luxi He, Boyi Wei, Dacheng Li, Ying Sheng, et al. Sorry-bench: Systematically evaluating large language model safety refusal. *arXiv preprint arXiv:2406.14598*, 2024. 6

[54] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, et al. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. *arXiv preprint arXiv:2408.02900*, 2024. 2

[55] Jiacong Xu, Shao-Yuan Lo, Bardia Safaei, Vishal M. Patel, and Isht Dwivedi. Towards zero-shot anomaly detection and reasoning with multimodal large language models. *arXiv preprint arXiv:2502.07601*, 2025. 2, 3, 6, 8

[56] Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025. 1, 2, 4, 6, 8

[57] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *CVPR*, pages 9499–9508, 2022. 1

[58] Linli Yao, Weiying Wang, and Qin Jin. Image difference captioning with pre-training and contrastive learning. In *AAAI*, pages 3108–3116, 2022. 3

[59] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, pages 9556–9567, 2024. 5, 1

[60] Q Zhou and et al. Object-agnostic prompt learning for zero-shot anomaly detection. *arXiv preprint arXiv:2310.18961*, 2023. 3, 8

[61] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1, 2

[62] Yang Zou, Jongheon Jeong, Latha Pemula, Dongqing Zhang, and Onkar Dabeer. Spot-the-difference self-supervised pre-training for anomaly detection and segmentation. In *ECCV*, pages 392–408, 2022. 3