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

Supplementary Material

This supplementary document provides additional details supporting the main paper.

A. Chat Template

To incorporate the new <Ano> and <Diff> tokens, we update the original chat template to provide more explicit cues for abnormality detection and change description (Fig. 6). While the original template simply presented two images and asked what had changed, MEDIC-AD template attaches anomaly tokens (<Ano>) to each image and adds diff tokens (<Diff>) to the change question. This enriches the prompt with clinically aligned signals, enabling the model to produce more accurate abnormality identification and localized change explanations.

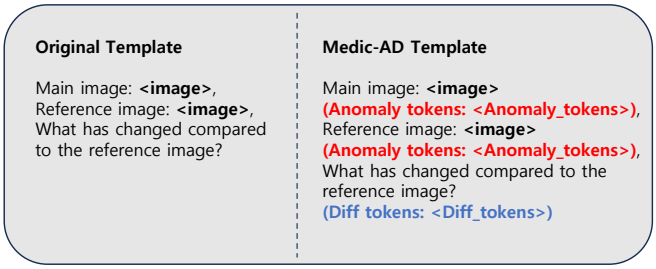

Figure 6. Comparison of chat template between original and MEDIC-AD, utilizing <Ano> and <Diff> tokens.

B. Training Details

All training stages employ the AdamW optimizer together with a cosine learning rate scheduler. **Stage 1** is trained on 3 H200 GPUs for one epoch (3 hours) using an initial learning rate of $1e-4$ and a batch size of 16. **Stage 2** uses the same hardware and optimization settings, trained for a single epoch (6 hours). Both Stage 1 and Stage 2 are optimized using the standard LLM cross-entropy loss. **Stage 3** is trained for 100 epochs (15 hours) with an initial learning rate of $1e-3$ and a batch size of 32, optimizing the DiceCE loss. The parameter size of additional modules introduced in MEDIC-AD follows the specifications summarized in Tab. 6.

Table 6. Parameter size of MEDIC-AD's additional modules.

Category	Param. Size
Visual soft prompts	51K
<Ano> modules	40M
<Diff> modules	40M
Heatmap modules	48M

C. Additional experiments

C.1. Generic-Medical VQA

In this section, we assess the general medical reasoning capability of MEDIC-AD by benchmarking it on a diverse set of established medical VQA and QA datasets, including VQA-RAD [33], SLAKE [51], PathVQA [19], MMMU Med [62], PMC-VQA [63] for visual reasoning, and MedMCQA [44], PubMedQA [28], MedQA [27], MedXpertQA [68] for text-only medical QA. As shown in Tab. 7 and Tab. 8, MEDIC-AD preserves broad medical knowledge acquired during pretraining while maintaining strong question–answering competence across both visual and non-visual domains. Notably, when comparing against a fully fine-tuned Lingshu baseline—trained end-to-end on the Stage 1 and Stage 2 datasets—we observe a pronounced degradation in general medical knowledge, leading to substantial drops across benchmarks. This contrast underscores that MEDIC-AD's stage-wise design, which updates only targeted components such as <Ano> and <Diff>, effectively avoids catastrophic forgetting and retains general-domain medical reasoning capabilities.

C.2. Results on hyperparameter experiments

The experimental results on hyperparameters summarized in Fig. 5 are provided in full detail in Tab. 9, offering a comprehensive breakdown of each experimental configuration.

C.3. Stage-wise vs. Joint Training

Tab. 10 shows that stage-wise training consistently outperforms joint optimization under identical experimental settings. While joint training still yields performance improvements over the baseline, stage-wise optimization achieves superior results by providing a more stable learning curriculum. This result suggests that the proposed curriculum enables each stage-specific objective to be effectively internalized before introducing additional supervision signals.

Table 7. Results on medical VQAs (VQA RAD [33], SLAKE [51], PathVQA [19], MMMU Med [62]), and PMC VQA [63]).

Model	Size	VQA RAD	SLAKE	PathVQA	MMMU Med	PMC VQA
LLaVA-Med	7B	53.7	48.0	32.5	29.3	30.5
Citrus-V	8B	64.3	**84.9**	**62.0**	46.4	55.6
Lingshu	7B	**67.9**	83.1	61.9	54.0	56.0
Lingshu (full finetuning)	7B	40.8	38.4	45.3	24.8	6.4
MEDIC-AD	7B	64.3	78.5	56.5	**54.2**	**56.1**

Table 8. Results on medical QAs (MedMCQA [44], PubMedQA [28], MedQA [27], and MedXpertQA [68]).

Model	Size	MedMCQA	PubMedQA	MedQA	MedXpertQA
LLaVA-Med	7B	39.4	26.4	42.0	9.9
Citrus-V	8B	55.1	74.8	**64.9**	**16.9**
Lingshu	7B	55.9	75.4	63.3	16.5
Lingshu (full finetuning)	7B	1.8	35.4	24.4	10.9
MEDIC-AD	7B	**56.7**	**75.6**	63.6	16.5

Table 9. Ablation on pooling size and soft prompt counts. We report average F1 and MMXU (Overall) for each configuration.

Category	Config. Value	Avg. F1	MMXU
Pooling Size	1	90.5	0.620
	2	90.2	0.631
	4	91.6	0.635
	8	91.7	0.623
Soft Prompt Counts	0	90.6	0.625
	5	91.8	0.630
	10	91.6	0.635
	20	89.2	0.626

Table 10. Comparison between Stage-wise and Joint training.

Model	AVG. F1	MMXU
Lingshu-7B	88.7	0.620
Medic-AD (Joint)	<u>89.7</u>	<u>0.630</u>
Medic-AD (Stage-wise)	**91.6**	**0.655**