Boundary Guidance for Efficient 3D CT Vision–Language Reasoning

SooYong Kim[1†], **Kyeonghun Kim**[2†], **Taejin Kim**[1], **Yoonkyung Jeon**[1], **Jungmin Shin**[1], **Dayoon Lee**[1]
Hyunjun Kim[1], Won-jae Lee[1], Woo-kyung Jung[1], Hyuk-jae Lee[1], Pa Hong[3*], Namjoon Kim[1*]

[1]Seoul National University
[2]G-New Soft
[3]Samsung Changwon Hospital
[†]Equal contribution (Co-first authors)
[*]Corresponding authors

Abstract

Vision–language models (VLMs) for 3D computed tomography (CT) analysis face the dual challenge of achieving precise visual grounding in high-dimensional data while maintaining computational efficiency. Although state-of-the-art models with multi-billion-parameter decoders have demonstrated strong performance, their attention mechanisms are often distracted by clinically irrelevant but visually similar confounding features, leading to errors in reasoning. To mitigate this, we introduce **Dual-Polarity Bounding Box Prompting**, a novel visual instruction method that provides both positive and negative spatial cues. For each question, we overlay a **green box** on the region of interest (ROI) and a **red box** on a plausible but incorrect distractor region. This contrastive prompting scheme explicitly trains the model to focus its attention on relevant evidence while actively ignoring confounding information. We pair this technique with compact Qwen decoders (0.5B to 3B parameters) and evaluate it on the RadGenome-ChestCT and PMC-VQA benchmarks. Our results show that this dual-prompt strategy substantially improves both closed-ended and open-ended VQA performance. Notably, our 1.5B model, guided by dual-polarity prompts, surpasses the accuracy of a 7B baseline model, demonstrating that explicit negative guidance is a highly effective, parameter-efficient approach to enhancing the reliability and evidence-based reasoning of medical VLMs.

1 Introduction

The interpretive burden on radiologists has significantly increased due to the escalating volume of computed tomography (CT) scans in clinical workflows (1). While existing artificial intelligence tools offer assistance, they are often designed for narrow, single-pathology detection tasks and thus fall short of the holistic reasoning required for complex cases (2). **Vision–Language Models (VLMs)** present a more comprehensive solution, enabling natural language interaction with medical images. However, extending these models to volumetric 3D data introduces formidable computational challenges, primarily due to the high memory demands of large language decoders and the cubic scaling of feature tensors (16).

Recent models such as **Med3DVLM** have demonstrated the feasibility of high-performance 3D medical VQA, but they rely on massive 7-billion-parameter decoders, which limits their practical deployment in clinical settings due to significant hardware requirements (5). This has motivated a shift towards more efficient methodologies, including **Parameter-Efficient Fine-Tuning (PEFT)**

techniques like **Low-Rank Adaptation (LoRA)** (18) and the adoption of smaller yet powerful language models (10; 11).

Beyond computational efficiency, a critical limitation of current VLMs is their vulnerability to visual distractors. These models often struggle to differentiate a true pathological finding from a benign structure or artifact that appears visually similar—a fundamental task in differential diagnosis. Existing visual prompting methods attempt to improve spatial grounding by highlighting the region of interest (ROI) with visual cues like contours or scribbles (8; 9). However, these "positive-only" prompts do not explicitly teach the model what to *ignore*, leaving it susceptible to confounding features elsewhere in the image.

To address this gap, we introduce **Dual-Polarity Bounding Box Prompting**, a novel contrastive visual instruction strategy. Our approach supplements the standard positive prompt (a **green box** on the ROI) with a negative prompt (a **red box** on a distractor region). This mechanism forces the model to develop more robust, evidence-based reasoning. We demonstrate that this technique enables a compact 1.5B parameter model to outperform a 7B baseline, proving that teaching a model what to ignore is as important as teaching it where to look.

2 Related Work and Background

2.1 Volumetric Vision–Language Foundations

The transition from 2D radiographs to 3D medical VLMs was enabled by two key advancements: specialized encoders and large-scale, grounded datasets.

- **Architectures:** The **DCFormer** model introduced an efficient architecture for processing 3D data without prohibitive computational costs (12). Building upon this, **Med3DVLM** achieved state-of-the-art performance on the M3D benchmark, establishing the efficacy of large language modules for reasoning over volumetric data (5; 15).
- **Datasets:** Progress has been significantly fueled by richly annotated datasets. **RadGenome-ChestCT** provides over a million question-answer pairs grounded in segmentation masks (4), while datasets like **PMC-VQA** offer diverse questions across various medical imaging modalities (16).

2.2 Parameter-Efficient Adaptation

Fine-tuning billion-scale models is often infeasible in resource-constrained environments.

- **LoRA: Low-Rank Adaptation (LoRA)** is a pivotal technique that freezes a model's pre-trained weights and fine-tunes only a small set of injected adapter matrices, dramatically reducing the number of trainable parameters (18). This strategy has been successfully applied to medical multimodal tasks in models such as **PeFoMed** (11).
- **Efficient Models & Kernels:** The development of smaller, yet highly capable language models like the **Qwen 2.5** family provides efficient alternatives to massive decoders (10). Furthermore, optimized software kernels like **FlashAttention** accelerate training by enhancing the speed and memory efficiency of the core attention mechanism (17; 25).

2.3 Explicit Spatial Prompting

To ensure models base their reasoning on visual evidence rather than dataset biases, visual prompting methods explicitly guide model attention.

- **Positive Prompts:** Initial works like **ViP-LLaVA** and **MedVP** demonstrated that overlaying visual cues such as colored outlines on 2D images can effectively ground a model's reasoning (8; 9).
- **Our Contribution (Negative Prompts):** These existing approaches, however, exclusively provide **positive guidance**. Our work introduces the concept of **negative visual prompts** to train the model to actively suppress attention to known confounders, thereby fostering a more robust and advanced reasoning capability.

3 Methodology

3.1 Dual-Polarity Visual Prompt Generation

Our core contribution is a visual prompting scheme that provides both positive and negative spatial guidance. For each question-answer pair grounded in a region of interest (ROI), we generate two distinct visual cues on every relevant 2D slice of the 3D CT volume.

1. **Positive Prompt (Green Box):** From the ground-truth segmentation mask M of the target anatomy, we compute the minimal bounding box that encloses the ROI. This box is rendered as a 3-pixel-thick **green** outline on each intersecting 2D slice. This serves as the primary instruction, directing the model's attention to the relevant evidence.
2. **Negative Prompt (Red Box):** To train the model to ignore distractors, we generate a negative bounding box. The distractor region is selected by sampling a different anatomical structure, often one that is nearby or visually similar to the ROI (e.g., a healthy blood vessel near a lung nodule, or a rib artifact). We use masks from a pre-trained segmentation model like VISTA-3D (14) to identify these candidate distractor regions. A bounding box is then drawn around the selected distractor and rendered as a 3-pixel-thick **red** outline.

This process transforms a standard CT volume into a visually prompted volume where each slice contains clear "go" and "no-go" signals for the model's attention mechanism.

3.2 Prompt-Conditioned Model and Instruction Tuning

Our model architecture builds upon Med3DVLM, retaining the frozen DCFormer encoder (12). We replace the decoder with more compact Qwen-2.5 language models (10) and fine-tune them using LoRA adapters (18). The integration of our dual-polarity prompts occurs at the input level.

The visually prompted 3D volume (with green and red boxes) is fed to the vision encoder. Crucially, we rewrite the corresponding textual prompt to make the model aware of the visual cues. A given question q is programmatically transformed into an instruction-following format:

```
"Referring to the CT images, answer the following question based on the
anatomy inside the green-outlined area, while ignoring anything inside the
red-outlined area. Question: [original question q]"
```

This explicit textual instruction, combined with the visual boxes, creates a powerful, multi-modal signal that conditions the model to perform contrastive reasoning. The combined visual and textual embeddings are fused and processed by the Qwen decoder to generate an answer.

3.3 Training Objective

We train the model end-to-end on the VQA task. Since our datasets contain both closed-ended (e.g., yes/no, multiple choice) and open-ended (free-text) questions, we use a composite loss function.

$$\mathcal{L}_{total} = \lambda_{cls}\mathcal{L}_{focal} + \lambda_{gen}\mathcal{L}_{nll}$$

where:

- $\mathcal{L}_{focal}$ is the **Focal Loss** (19) for the closed-ended VQA task. This loss function effectively handles class imbalance by down-weighting the loss attributed to well-classified examples, allowing the model to focus on harder, less frequent cases.
- $\mathcal{L}_{nll}$ is the standard token-level **Negative Log-Likelihood** loss for the auto-regressive open-ended answer generation task.
- λ_{cls} and λ_{gen} are hyper-parameters to balance the two tasks, which we set to 1.0.

During training, only the LoRA adapter weights and the language model's normalization layers are updated, making the process highly parameter-efficient.

4 Experiments

4.1 Datasets and Preprocessing

We evaluate our approach on two large-scale medical VQA benchmarks. **RadGenome-ChestCT** (4) is a comprehensive 3D dataset featuring 25,692 chest CT volumes linked to approximately 1.3 million grounded question-answer pairs. It provides organ-level segmentation masks, which we use to generate the positive bounding boxes. We adhere to the official patient-level 70/10/20 split to prevent data leakage. **PMC-VQA** (16) is a diverse 2D dataset containing over 227,000 QA pairs from 149,000 images sourced from biomedical literature. We filter this dataset for CT images, resulting in 91,000 QA pairs. An article-level 80/10/10 split is used to ensure train and test sets are distinct. For both datasets, we generate our dual-polarity prompts offline. Negative prompts are created by selecting distractor regions from non-target anatomical masks provided by the VISTA-3D model (14).

4.2 Implementation Details

All models were trained on a single NVIDIA A6000 GPU with 48 GB of VRAM, using PyTorch 2.2 and Flash-Attention 3 (17) for memory and speed optimization. We employed mixed-precision (`bf16`) training. The AdamW optimizer (20) was used with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 0.01 applied only to the LoRA adapter weights. We used a cosine learning rate schedule with 500 warm-up steps, with a peak learning rate of 2×10^{-5} for the 0.5B and 1.5B models, and 1×10^{-5} for the 3B model. All vision encoder weights remained frozen. We trained rank-16 LoRA adapters (18) on the attention and feed-forward layers of the Qwen decoders for 25 epochs.

4.3 Baselines and Evaluation

We compare three experimental conditions across the 0.5B, 1.5B, and 3B Qwen decoder variants.

1. **No Prompt**: The standard baseline where models are trained on original images and text without any visual cues.
2. **Positive Prompt Only**: An ablation study where models are guided only by the positive (green) bounding box around the ROI.
3. **Dual-Polarity Prompt (Ours)**: The proposed method using both positive (green) and negative (red) bounding boxes.

For reference, we also include the reported performance of the original 7B parameter Med3DVLM as a high-end baseline. We evaluate closed-ended VQA using **Accuracy (ACC)** and open-ended VQA using **BLEU-4**, **ROUGE-L**, and **METEOR**.

4.4 Results and Discussion

Table 1 presents the main results of our VQA experiments. Several key trends are evident. First, across all model sizes, providing a **Positive Prompt** consistently improves performance over the **No Prompt** baseline, confirming that explicit spatial guidance is beneficial.

Second, and most importantly, our **Dual-Polarity Prompt** method yields the highest performance in all settings. The addition of a negative prompt provides a significant boost over the positive-only prompt, particularly for open-ended metrics like METEOR on PMC-VQA, which require more nuanced reasoning. This demonstrates that teaching the model what to ignore is a powerful mechanism for reducing confusion and improving focus.

Finally, our results highlight a remarkable improvement in parameter efficiency. The 1.5B model equipped with dual-polarity prompts not only surpasses the 3B model without prompts but also achieves performance that is highly competitive with—and in some cases exceeds—the 7B baseline. For instance, on RadGenome-ChestCT, our guided 1.5B model achieves 75.2% accuracy, outperforming the unguided 3B model (72.7%) and approaching the 7B baseline (79.9%). This shows that our contrastive prompting scheme can compensate for a smaller parameter count by fostering more efficient and accurate reasoning.

Table 1: VQA performance on RadGenome-ChestCT (Closed-Ended) and PMC-VQA (Open-Ended). Our Dual-Polarity Prompt method enables smaller models to match or exceed larger baselines. All scores are percentages (%). Best performance for each model size is in **bold**.

2***Size**	2***Prompt Type**	**RadGenome-ChestCT**				**PMC-VQA**			
		ACC	B-4	R-L	MET	ACC	B-4	R-L	MET
7B	Med3DVLM	79.9	48.1	51.2	36.8	75.4	34.5	37.1	32.5
3*3B	No Prompt	72.7	46.6	49.5	33.5	71.8	32.1	35.4	31.7
	Positive Only	75.1	47.9	50.8	35.2	73.5	33.5	36.9	32.9
	Dual-Pol. (Ours)	**76.5**	**48.8**	**51.9**	**36.4**	**74.6**	**34.8**	**37.9**	**33.8**
3*1.5B	No Prompt	70.1	44.3	47.6	31.4	69.5	29.5	33.2	28.8
	Positive Only	73.0	46.1	49.2	33.8	71.8	31.6	35.1	30.9
	Dual-Pol. (Ours)	**75.2**	**47.5**	**50.6**	**35.5**	**73.1**	**32.9**	**36.2**	**32.1**
3*0.5B	No Prompt	67.8	41.1	44.9	29.8	66.2	25.1	28.3	23.9
	Positive Only	69.5	43.2	46.5	31.5	68.0	27.2	30.4	26.1
	Dual-Pol. (Ours)	**71.3**	**44.8**	**47.9**	**33.1**	**69.4**	**28.9**	**31.8**	**27.5**

5 Conclusion

In this work, we addressed the critical challenge of visual grounding and distractibility in 3D medical VLMs. We introduced **Dual-Polarity Bounding Box Prompting**, a novel contrastive prompting technique that guides model attention using both a positive (green) cue for the region of interest and a negative (red) cue for a confounding distractor. By coupling this method with parameter-efficient Qwen decoders, we demonstrated a powerful and practical approach to building robust medical VQA systems.

Our experiments on RadGenome-ChestCT and PMC-VQA confirmed that this dual-prompt strategy significantly enhances model performance across all metrics. By explicitly teaching the model to focus on relevant evidence while inhibiting attention to irrelevant features, our method fosters a more nuanced and reliable reasoning process. The most compelling finding is that our guided 1.5B parameter model achieves performance competitive with a 7B baseline, showcasing a superior trade-off between accuracy and computational cost.

This work has limitations that open avenues for future research. Our method currently relies on pre-computed segmentation masks to generate prompts. Future work could explore end-to-end models that learn to generate these spatial prompts dynamically. Furthermore, the selection of negative prompts is currently heuristic; developing a more advanced strategy, perhaps using an adversarial approach to identify the most challenging distractors, could yield further improvements. Finally, extending and validating this contrastive prompting approach on other medical imaging modalities, such as MRI and PET, is a promising direction. Ultimately, by enabling finer-grained control over model attention, we believe our work represents a key step towards developing safer, more interpretable, and clinically trustworthy AI assistants.

References

[1] T. C. Kwee and R. M. Kwee, "Workload of diagnostic radiologists in the foreseeable future based on recent (2024) scientific advances: Updated growth expectations," *European Journal of Radiology*, vol. 187, Art. no. 112103, 2025.

[2] A. Kurmukov, V. Chernina, R. Gareeva *et al.*, "The impact of deep-learning aid on the workload and interpretation accuracy of radiologists on chest computed tomography: a cross-over reader study," arXiv preprint arXiv:2406.08137, 2024.

[3] I. E. Hamamci, S. Er, C. Wang *et al.*, "Developing generalist foundation models from a multimodal dataset for 3D computed tomography," arXiv preprint arXiv:2403.17834, 2025.

[4] X. Zhang, C. Wu, Z. Zhao, J. Lei, Y. Zhang, Y. Wang, and W. Xie, "RadGenome-Chest CT: A grounded vision-language dataset for chest CT analysis," arXiv preprint arXiv:2404.16754, 2024.

[5] Y. Xin, G. C. Ates, K. Gong, and W. Shao, "Med3DVLM: An efficient vision-language model for 3D medical image analysis," arXiv preprint arXiv:2503.20047, 2025.

[6] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language-image pre-training," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 11975–11985.

[7] G. Shinde, A. Ravi, E. Dey, S. Sakib, M. Rampure, and N. Roy, "A survey on efficient vision-language models," arXiv preprint arXiv:2504.09724, 2025.

[8] M. Cai, H. Liu, D. Park *et al.*, "ViP-LLaVA: Making large multimodal models understand arbitrary visual prompts," arXiv preprint arXiv:2312.00784, 2023.

[9] K. Zhu, Z. Qin, H. Yi *et al.*, "Guiding medical vision-language models with explicit visual prompts: Framework design and comprehensive exploration of prompt variations," arXiv preprint arXiv:2501.02385, 2025.

[10] A. Yang, B. Yang, B. Zhang *et al.*, "Qwen 2.5 technical report," arXiv preprint arXiv:2412.15115, 2025.

[11] J. He, P. Li, G. Liu *et al.*, "PeFoMed: Parameter-efficient fine-tuning of multimodal large language models for medical imaging," arXiv preprint arXiv:2401.02797, 2024.

[12] G. C. Ates, Y. Xin, K. Gong, and W. Shao, "DCFormer: Efficient 3D vision–language modeling with decomposed convolutions," arXiv preprint arXiv:2502.05091, 2025.

[13] I. E. Hamamci, S. Er, C. Wang *et al.*, "Developing generalist foundation models from a multimodal dataset for 3D computed tomography (CT-RATE)," arXiv preprint arXiv:2403.17834, 2024.

[14] Y. He, P. Guo, Y. Tang *et al.*, "VISTA3D: A unified segmentation foundation model for 3D medical imaging," arXiv preprint arXiv:2406.05285, 2024.

[15] F. Bai, Y. Du, T. Huang, M. Q.-H. Meng, and B. Zhao, "M3D: Advancing 3D medical image analysis with multi-modal large language models," arXiv preprint arXiv:2404.00578, 2024.

[16] X. Zhang, C. Wu, Z. Zhao, W. Lin, Y. Zhang, Y. Wang, and W. Xie, "PMC-VQA: Visual instruction tuning for medical visual question answering," arXiv preprint arXiv:2305.10415, 2023.

[17] J. Shah, G. Bikshandi, Y. Zhang, V. Thakkar, P. Ramani, and T. Dao, "FlashAttention-3: Fast and accurate attention with asynchrony and low-precision," arXiv preprint arXiv:2407.08608, 2024.

[18] E. J. Hu, Y. Shen, P. Wallis *et al.*, "LoRA: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[19] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[20] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2019.

[21] J. Su, Y. Lu, S. Pan *et al.*, "RoFormer: Enhanced transformer with rotary position embedding," arXiv preprint arXiv:2104.09864, 2021.

[22] G. Bertasius, H. Wang, and L. Torresani, "Is space–time attention all you need for video understanding?" in *Proc. Intl Conf. Machine Learning (ICML)*, 2021.

[23] J. Wasserthal, H.-C. Breit, M. T. Meyer *et al.*, "TotalSegmentator: Robust segmentation of 104 anatomical structures in CT images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, Art. e230024, 2023.

[24] C. P. Langlotz, "RadLex: A new method for indexing online educational materials," *Radiographics*, vol. 26, no. 6, pp. 1595–1597, 2006.

[25] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.