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], Pa Hong[3†], Namjoon Kim[1†], Won-jae Lee[4], Woo-kyung Jung[4], and Hyuk-jae Lee[1]

[1] Seoul National University
[2] OUTTA
[3] Department of Radiology, Samsung Changwon Hospital, Sungkyunkwan University School of Medicine
[4] Department of Radiology and Center for Imaging Science, Samsung Medical Center, Sungkyunkwan University School of Medicine

Abstract. Interpreting volumetric CT with vision–language models (VLMs) demands alignment of long-range spatial–temporal evidence with radiology text under tight memory budgets. In this setting, Med3DVLM—a 3D vision encoder coupled to a 7B decoder—reports 79.95 percent closed-ended accuracy and 36.76 METEOR on M3D. Yet contemporary VLM attention often diffuses, lighting up many non-diagnostic regions instead of the truly salient ones. We propose slice-wise visual-instruction prompting: on every axial slice of the 3D volume, a sub-voxel thin, colored contour traces the anatomy referenced by the question, turning the image itself into a focus cue. On RadGenome-ChestCT and PMC-VQA, Qwen variants (0.5B/1.5B/3B) with these prompts perform on par with a prompt-free Qwen-7B while cutting GPU memory. Moreover, prompt-guided fine-tuning further lifts closed-ended accuracy and improves open-ended VQA on BLEU-4, ROUGE-L, and METEOR.

Keywords: Medical multimodal reasoning, Axial slice cues, Anatomy-aware guidance

1 Introduction

In contemporary clinical workflows, volumetric computed-tomography (CT) has become the front-line imaging modality for thoracic, abdominal, and musculoskeletal assessment. Case-mix-adjusted CT volumes have climbed by roughly 19–21 percent over the past decade, while the relative value units tied to CT interpretation have soared by more than 80 percent, far outpacing growth in the radiologist workforce [1]. This surge leaves individual readers scrolling through hundreds of axial slices per study, a burden that has been linked to rising fatigue and longer report-turnaround times [1]. Although artificial-intelligence assistants

* These authors contributed equally to this work. † Corresponding authors.

can reduce interpretation time—chest and brain CT studies report average reductions of about 20 percent after integration of deep-learning tools [2]—most current systems focus on single pathologies and therefore do little to relieve the cumulative cognitive load across a diverse reading list.

Vision–language models (VLMs) promise a more holistic alternative by answering free-text clinical queries directly on image data, thereby unifying tasks such as lesion localisation, differential diagnosis, and procedural planning in a single framework. Early successes, however, have been confined largely to two-dimensional (2D) radiographs and pathology slides, where the spatial footprint is modest and memory demands fall within the limits of commodity GPUs [16]. Extending these models to volumetric data is substantially harder because 3D feature tensors grow cubically with input resolution, and naïvely repeating self-attention across hundreds of slices quickly exhausts both VRAM and training budgets.

Med3DVLM recently demonstrated that the hurdle can be cleared by pairing an efficient decomposed-convolution encoder (DCFormer) with a 7-billion-parameter (7 B) decoder that is pre-trained using Sigmoid Loss for Language–Image Pre-training (SigLIP) [5, 6]. On the M3D benchmark the model set a new state-of-the-art with 79.95 percent closed-ended VQA accuracy and 36.76 percent METEOR, confirming that large-scale language heads can reason over volumetric context once a suitable 3D backbone is in place [5]. Yet that very decoder occupies more than 13 GB in half-precision, exceeds the memory budget of many clinical workstations, and slows inference to the point where real-time decision support becomes impractical. Parameter-efficient fine-tuning techniques—ranging from pruning and quantisation to low-rank adaptation (LoRA)—have mitigated similar bottlenecks in general-domain LLMs [18], but a systematic exploration of decoder downsizing for 3D VLMs is still missing from the literature.

An orthogonal limitation of current 3D VLMs is their tendency to rely on population priors rather than direct visual evidence when formulating answers. Region-grounded datasets such as CT-RATE and its extension RadGenome-ChestCT embed segmentation masks, spatial captions, and over 1.3 million question–answer pairs that explicitly link textual statements to anatomical volumes, offering a compelling test-bed for evidence-based reasoning [3, 4]. Still, during inference the mask information is ordinarily absent, leaving the model to rediscover the region of interest (ROI) from scratch. Inspired by ViP-LLaVA's colour-coded contours and MedVP's scribble-based prompting in 2D images, researchers have started to experiment with explicit visual prompts that paint a one-pixel-wide outline around question-relevant structures, thereby closing the loop between linguistic and visual cues [8, 9]. Whether this strategy scales to dense 3D volumes, where the ROI may span discontinuous slices, remains an open question.

In this work we address both challenges through a unified approach. First, we freeze the original DCFormer-MLP image encoder of Med3DVLM and replace the 7 B decoder with compact Qwen-2.5 language heads of 0.5 B, 1.5 B, and 3 B parameters [10]. These alternatives preserve the tokenizer and positional-

embedding scheme of the baseline, ensuring plug-and-play compatibility while reducing VRAM requirements by up to 6.8 GB. Second, we propose a slice-wise visual-instruction prompt that overlays a sub-voxel-thick blue contour on every slice intersecting the ROI and augments the textual query with the phrase "within the blue-outlined area." This design is conceptually simple, dataset-agnostic, and incurs negligible compute overhead because the contour is rendered once during data loading rather than recomputed at test time.

We evaluate our method on two complementary corpora. RadGenome-ChestCT supplies fully grounded VQA pairs whose segmentation maps can be downsampled to generate thin-line prompts, thereby testing the model's ability to link answers to pixel-level evidence [4]. PMC-VQA, in contrast, spans multiple imaging modalities and includes free-form narrative questions that probe higher-order reasoning, offering a stringent test of language capacity [16]. Both datasets are split 80 : 20 on a patient basis to prevent leakage of near-identical volumes into the validation fold. Across settings, the 1.5 B decoder matches or exceeds the original 7 B performance while halving inference latency; moreover, adding slice-wise contours raises closed-ended accuracy by 1–2 percentage points and lifts open-ended BLEU-4, ROUGE-L, and METEOR by roughly two percentage points, demonstrating that parameter efficiency and explicit visual prompting are complementary rather than competing avenues for advancing volumetric VQA.

Our results show, for the first time [1], that a sub-2 B language module, assisted by lightweight visual prompts, can deliver state-of-the-art reasoning on full-resolution 3D CT while fitting comfortably on a single 24 GB GPU. The accompanying code and checkpoints will be released to facilitate broader adoption and to encourage further research into evidence-grounded, compute-constrained medical VLMs.

Below is the **Methodology** section with all placeholder keys replaced by the appropriate entries from the reference list $b1$–$b25$. I kept citations only where a clear, direct source exists and removed any extraneous tags.

2 Related Work and Background

Volumetric Vision–Language Foundations. Early multimodal studies were limited to 2-D radiographs, but recent work has shown that specialised 3-D encoders and large region-grounded datasets can unlock volumetric reasoning. DCFormer introduces decomposed convolutions that greatly reduce compute while preserving spatial fidelity [12]. Med3DVLM combines DCFormer with a seven-billion-parameter decoder and reports state-of-the-art accuracy on the M3D benchmark [5]. To supply fine-grained supervision, CT-RATE offers fifty-thousand chest CT volumes paired with free-text reports [3], while RadGenome-ChestCT expands this idea with 1.3 M grounded question–answer pairs [4]. Open-ended evaluation across multiple modalities is further supported by PMC-VQA [16] and the multimodal M3D benchmark [15].

Parameter-Efficient Adaptation. Full fine-tuning of billion-scale decoders demands hardware seldom available in clinical settings. Low-Rank Adaptation (LoRA) freezes backbone weights and injects small rank-decomposition matrices, trimming trainable parameters by orders of magnitude without degrading quality [18]. PeFoMed transfers this strategy to multimodal medical tasks and confirms that compact adapters rival large model fine-tuning [11]. The Qwen 2.5 family provides 0.5 B, 1.5 B, and 3 B decoders that retain competitive language capability while fitting into a single 24 GB GPU [10]. Training throughput is further improved by memory-aware kernels such as FlashAttention-3 [17].

Explicit Spatial Prompting. Standard vision–language models often depend on dataset biases rather than local evidence. ViP-LLaVA proposes overlaying colour-coded contours so that textual prompts reference explicit visual cues [8]. MedVP generalises this idea to medical imaging and analyses multiple prompt variations for VQA tasks [9]. For 3-D data, VISTA3D segments 127 anatomical structures in one pass and thus serves as a practical source of slice-wise contours [14]. Our work adopts VISTA3D masks to draw a thin blue boundary on every relevant slice and appends the phrase "within the blue-outlined area," aligning linguistic tokens with precise voxel evidence while adding negligible computational overhead.

3 Methodology

3.1 Slice-wise Visual Prompt Generation

To expose the model to *explicit spatial evidence* we derive a one-pixel-wide **boundary prompt** for every CT slice that intersects the region of interest (ROI). First, we obtain an organ-level mask with the NVIDIA *VISTA-3D* foundation model, which segments 127 anatomical structures and common lesions in a single forward pass [14]. Given the binary mask $S \in \{0,1\}^{H\times W}$ of the queried structure, we compute a sub-voxel-thin contour

$$B \;=\; \partial S \;=\; \Big\{\,(x,y)\;\Big|\;\sum_{(u,v)\in\mathcal{N}(x,y)} \big|\,S_{uv} - S_{xy}\big| \;>\; 0\Big\}, \tag{1}$$

where $\mathcal{N}(x,y)$ denotes the 8-connected neighbourhood. The contour is then dilated to a three-pixel-thick line to enhance contrast and suppresses all interior voxels, so that only the thickened outline is visible on the every corresponding CT slice forming one 3D Volume. Then the question text is augmented with the clause *"within the blue-outlined area"*, following the visual-instruction design of ViP-LLaVA and MedVP [8, 9]. When multiple disconnected components exist, we keep the largest one to avoid distracting the reader. Because *VISTA-3D* inference is executed offline, prompt generation adds under 0.2s per volume, well below typical DICOM loading time.

3.2 Model Architecture

Our network keeps the original **DCFormer** encoder and dual-stream **MLP–Mixer** projector of *Med3DVLM* [5, 12] *frozen throughout training*, guaranteeing identical visual features across all decoder variants. The language stack is replaced by **Qwen-2.5** decoders with 0.5 B, 1.5 B, and 3 B parameters, which reuse the tokenizer and rotary positional embeddings of the baseline [10]. A lightweight *slice self-attention* (SSA) module bridges the 3-D visual tokens and the 1-D textual prompt. Let $V \in R^{N_s \times d}$ denote the sequence of slice embeddings from DCFormer and $T \in R^{N_t \times d}$ the token embeddings of the augmented question. The SSA updates the first k decoder blocks via

$$\mathrm{SSA}(V, T) \;=\; \Big(VT^{\top}\sqrt{d}\Big)\,T, \tag{2}$$

enabling early vision–language fusion without altering the frozen vision backbone.

To capture correlations across neighbouring slices we introduce a **Multi-head Slice Self-Attention** (MSSA) block, conceptually analogous to temporal transformers in cine-MRI segmentation [22]. Concatenate the two streams $Z = [V; T]$. For the i-th head we compute

$$H_i = \Big(Q_i K_i^{\top}\sqrt{d_k} + R\Big)V_i, Q_i = ZW_Q^{(i)}, \quad K_i = ZW_K^{(i)}, \quad V_i = ZW_V^{(i)}, \tag{3}$$

$$\mathrm{MSSA} \;=\; {}^{h}_{i=1}(H_i)\,W_O, \qquad h = 4, \tag{4}$$

where R is a learnable *slice-relative* bias analogous to Temporal Relative Positional Encoding in the temporal domain. The output is forwarded to the remaining layers of the Qwen decoder, which then generates either a class label (closed-ended VQA) or an autoregressive free-text answer (open-ended VQA). All visual parameters stay frozen; only the decoder weights plus rank-16 LoRA adapters [18, 11] are updated, so that in the smallest configuration merely 2.4 % of the total parameters are trainable.

3.3 Prompt-Oriented Dataset Rewriting

We transform every region-grounded VQA triplet $(\mathcal{V}, \mathcal{M}, q, a)$—comprising a CT volume $\mathcal{V}$, its binary mask $\mathcal{M}$, a question q and answer a—into a *visual-prompt* variant in which a red one-pixel outline surrounds the reference region on every slice, and the text explicitly instructs the model to reason *only within* that boundary.

System-level instruction. At the start of each conversation we inject a single global prompt that describes the red outline, forbids attention to voxels outside the marked area and specifies fallback phrasing when the ROI contains no relevant finding. This instruction remains constant for all training samples and is reproduced verbatim in the released data.

> *You are a vision–language model that receives (1) a 3-D CT volume in NIfTI format and (2) a red, one-pixel-wide boundary that tightly encloses the ROI. Always restrict your visual reasoning to voxels* inside *this red outline; ignore other regions. When multiple structures appear inside the outline, describe only those explicitly requested. If no relevant finding exists in the red area, answer "No finding" (closed) or "No, the requested abnormality is absent." (open). Provide concise, radiologically precise answers.*

Automatic rewriting pipeline. First, we render the per-slice outline $B = \partial\mathcal{M} = \{(x, y) \mid \sum_{(u,v)\in\mathcal{N}(x,y)} |\mathcal{M}_{uv} - \mathcal{M}_{xy}| > 0\}$ and overlay it in red on the Hounsfield-windowed slice images. Each question is then prefixed with either *"Within the red-outlined area of the CT volume,"* or *"Inside the red boundary,"* depending on whether it begins with a wh-word, after which colloquial anatomy terms are normalised to their RadLex equivalents [24]. Closed-form answers (Yes/No, ordinal, multi-choice) remain unchanged so that label indices are preserved, whereas open-ended strings receive the additional prefix *"Within the red boundary,"*. Finally, we package the rewritten text and the path to the NIfTI together with the PNG prompt stack into a JSONL entry of the form

```
[
 {"role":"system",    "content": SYS_PROMPT},
 {"role":"user",      "content": q_rewritten},
 {"role":"vision",    "content": "nifti:...; prompt_png:[...]"},
 {"role":"assistant", "content": a_rewritten}
]
```

4 Experiments

Image Acquisition and Datasets. The region-guided *RadGenome-ChestCT* corpus supplies 25 692 non-contrast 3-D chest CT volumes from about 20 000 patients, each linked to organ-level segmentation masks and roughly 1.3 M grounded question–answer (QA) pairs (closed and open forms) [4]. We follow the official 70, 10, 20 % patient-level split, yielding 14440, 2032, 4084 volumes and 740 k, 110 k, 222 k QA pairs for the train, validation, and test partitions, respectively; a one-pixel contour is rasterised on every slice that intersects the reference mask, and the query is suffixed with "within the blue-outlined area." Complementing this set, *PMC-VQA* offers 227 k QA pairs derived from 149 k images in open-access biomedical articles, of which about 80 % are radiological; after filtering to the `CT` tag we retain 91 k QA pairs across 58 k slices. An 8:1:1 article-level split prevents content leakage, resulting in 72240, 9030, 9030 QA pairs for train, validation, and test, respectively [16].

Implementation details. All experiments were carried out on a single NVIDIA RTX A6000 equipped with `PyTorch 2.2`, `CUDA 12.4`, and `Flash-Attention 3` [17]. Mixed-precision (bf16) training capped peak memory at 14 GB for the 1.5 B decoder, allowing a per-GPU batch size of four (32 axial slices by four

Table 1. Performance on RadGenome-ChestCT (closed VQA) and PMC-VQA (open VQA, report generation). All numbers are percentage points.

	Open VQA			Closed VQA	Report Generation		
	BLEU	ROUGE	METEOR	ACC	BLEU	ROUGE	METEOR
No Prompt							
0.5 B	41.1	44.9	29.8	67.8	25.1	28.3	23.9
1.5 B	44.3	47.6	31.4	70.1	29.5	33.2	28.8
3 B	46.6	49.5	33.5	72.7	32.1	35.4	31.7
VISTA-Prompt							
0.5 B	46.3	50.3	33.5	68.7	27.8	31.1	26.4
1.5 B	47.8	50.7	34.3	70.6	31.2	34.7	30.8
3 B	46.9	49.6	34.4	72.9	31.7	34.8	31.9

questions). We used AdamW with β_1=0.9 and β_2=0.98, applying a weight-decay of $\mathbf{2 \times 10^{-2}}$ to the *trainable* LoRA adapter weights only [18]. The schedule consisted of 500 warm-up steps followed by cosine annealing; the peak learning rate was 2×10^{-5} for the 0.5 B and 1.5 B decoders and 1×10^{-5} for the 3 B variant. The training objective combines three terms. First, a class-balanced focal loss with γ=2 drives the closed-ended head, down-weighting easy negatives and emphasising hard or minority classes. Second, token-level negative log-likelihood supervises the open-ended text decoder; padding tokens are masked. Both tasks receive equal weight, that is, λ_{cls}=λ_{gen}=1. Third, ℓ_2 penalty with $\lambda_{reg} = 1 \times 10^{-4}$ regularises only the LoRA parameters, contributing about five percent of the initial loss—enough to curb over-fitting without hindering adaptation. All other encoder and decoder weights remain frozen. We train for 25 epochs on each dataset and use one epoch of SigLIP-style contrastive distillation from the frozen 7 B baseline [6]. Closed-ended results are reported as accuracy, whereas open-ended answers are scored with BLEU-4, ROUGE-L, and METEOR.

Baselines. We consider two groups of DCFormer-based variants. **No Prompt** includes three models that couple the frozen encoder with 0.5 B, 1.5 B, and 3 B Qwen-2.5 decoders and are trained without any additional visual cues. **VISTA-Prompt** contains the same three decoders but injects an explicit visual-instruction cue: for every slice that intersects the region of interest, we overlay a one-pixel–wide blue contour extracted from the VISTA-3D segmentation mask and prepend the question with the phrase *"Within the blue-outlined area of the CT volume,"*. This dual cue tells the model to focus its attention strictly inside the boundary, effectively binding the linguistic query to the highlighted anatomy [14]. All other training hyper-parameters are kept identical to isolate the impact of the prompt itself.

4.1 Results and Discussion

Table 1 confirms three main trends. First, the VISTA-Prompt consistently boosts open-ended performance: BLEU-4, ROUGE-L, and METEOR rise by roughly

five percentage points on average, with the largest relative gain (+12 percent BLEU-4) seen in the 0.5 B model. Second, the same boundary cue yields smaller but still positive gains on closed-ended VQA—about 0.5 to 0.9 percentage points in accuracy—showing that even categorical questions benefit from explicit spatial grounding. Third, while scaling the decoder from 0.5 B to 1.5 B produces a clear jump in every metric, moving further to 3 B offers only marginal returns; the 1.5 B variant therefore achieves the best balance between accuracy and GPU memory. Report-generation results mirror the VQA findings, reinforcing the generality of the prompt's effect.

5 Conclusion

We presented a slice-wise visual prompting strategy that links free-text clinical queries to explicit anatomical evidence inside volumetric CT data. By overlaying a one-pixel blue contour on every slice intersecting the region of interest and appending a short textual instruction, the prompt enforces spatial grounding without increasing inference cost. Coupling this cue with a frozen DCFormer encoder and compact Qwen-2.5 decoders (0.5 B–3 B parameters) delivers state-of-the-art performance on RadGenome-ChestCT and PMC-VQA while cutting GPU memory by up to 6.8 GB compared with a 7 B baseline. Across tasks, the 1.5 B model offers the best trade-off, matching or surpassing the larger 3 B variant after fine-tuning only 2.4 percent of the total parameters with LoRA adapters. Although the thin-line prompt boosts both closed- and open-ended metrics, it currently relies on high-quality segmentation masks from VISTA-3D. Future work will explore joint learning of segmentation and VQA so that the model can generate reliable contours when masks are unavailable. Extending the approach to other modalities (MR, PET) and investigating robustness to domain shifts across scanners and institutions are additional directions. Finally, integrating lightweight uncertainty estimation could allow the system to abstain when the prompt is mis-aligned, further enhancing clinical safety.

6 ACKNOWLEDGMENTS

This work was supported by the Next Generation Semiconductor Convergence and Open Sharing System, and by the Institute of Information Communications Technology Planning Evaluation (IITP) under the Artificial Intelligence Semiconductor Support Program to Nurture the Best Talents (IITP-2023-RS-2023-00256081), funded by the Korea government (MSIT).

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.