MedPruner: Training-Free Hierarchical Token Pruning for Efficient 3D Medical Image Understanding in Vision-Language Models

Shengyuan Liu[1*], Zanting Ye[2,3*], Yunrui Lin[3*], Chen Hu[2,4], Wanting Geng[5], Xu Han[6], Bulat Ibragimov[7], Yefeng Zheng[2†], and Yixuan Yuan[1†]

[1]Chinese University of Hong Kong [2]Westlake University
[3]Southern Medical University [4]Jiangnan University
[5]Dalian University of Technology [6]Shanghai Jiao Tong University
[7]University of Copenhagen
yxyuan@ee.cuhk.edu.hk
zhengyefeng@westlake.edu.cn
[*] Equal contributions.
[†] Corresponding Author

Abstract. While specialized Medical Vision-Language Models (VLMs) have achieved remarkable success in interpreting 2D and 3D medical modalities, their deployment for 3D volumetric data remains constrained by significant computational inefficiencies. Current architectures typically suffer from massive anatomical redundancy due to the direct concatenation of consecutive 2D slices and lack the flexibility to handle heterogeneous information densities across different slices using fixed pruning ratios. To address these challenges, we propose MedPruner, a training-free and model-agnostic hierarchical token pruning framework specifically designed for efficient 3D medical image understanding. MedPruner introduces a two-stage mechanism: an Inter-slice Anchor-based Filtering module to eliminate slice-level temporal redundancy, followed by a Dynamic Information Nucleus Selection strategy that achieves adaptive token-level compression by quantifying cumulative attention weights. Extensive experiments on three 3D medical benchmarks and across three diverse medical VLMs reveal massive token redundancy in existing architectures. Notably, MedPruner enables models such as MedGemma-1.5 to maintain or even exceed their original performance while retaining fewer than 5% of visual tokens, thereby reducing visual-token overhead and validating the necessity of dynamic token selection for practical clinical deployment. Our code is available at here.

Keywords: Medical Vision-Language Models · 3D Medical Imaging · Token Pruning.

1 Introduction

Vision-Language Models (VLMs), ranging from proprietary systems like GPT-4o [9], Gemini [6] to high-performance open-source models such as the LLaVA [17]

and Qwen-VL [3] series, have demonstrated extraordinary universal perceptual and reasoning capabilities. Drawing upon these advancements, specialized medical VLMs [28,15,29,5,23,8,31,24] such as Med-PaLM M [25] and LLaVA-Med [15], have achieved exceptional proficiency in 2D medical image interpretation, providing critical support for accurate diagnosis and clinical decision-making [19,26,22]. Beyond 2D modalities, specialized architectures [2,1,14,13,20] have been developed to navigate 3D volumetric data (e.g., CT and MRI), enabling the analysis of volumetric anatomical structures and temporal dynamics that are critical for complex clinical scenarios. Recently, advanced medical VLMs, such as Hulu [11], MedGemma-1.5 [23], and RadFM [27], have further extended their capabilities to simultaneously process both 2D and 3D medical inputs, aiming to consolidate multimodal medical image understanding within a unified framework. Despite these advancements, extending these models from 2D images to 3D clinical scenarios remains challenging, as high-resolution volumetric medical images can induce a substantial increase in token count. This soaring computational demand necessitates an intelligent token pruning mechanism to maintain reasoning integrity while ensuring clinical-grade inference speeds.

However, current processing pipelines and general-purpose pruning methods exhibit critical limitations when applied to 3D medical inputs. First, their architectures [23,11,3,29] typically rely on feeding 2D slices along a single axis into the model, where the generated tokens are directly concatenated. Since consecutive slices in 3D volumes share extreme spatial similarity, this direct concatenation introduces massive redundancy that exhausts the LLMs context window and hinders the processing of auxiliary clinical information. Second, existing token pruning methods [11,30,18,12] typically employ a static, predefined pruning ratio, which fails to account for the inherent heterogeneity of information density. While certain slices capture the intricate boundaries of a tumor, others may only contain uniform tissue with minimal diagnostic value; a fixed ratio either risks losing fine-grained pathological details or wastes tokens on irrelevant backgrounds. Crucially, these static approaches ignore the variance in attention distributions across different vision backbones. Different models exhibit distinct perceptual biases toward medical features, rendering model-agnostic pruning sub-optimal.

To address these challenges, we propose MedPruner, a training-free and model-agnostic hierarchical token pruning approach tailored for 3D medical image understanding in VLMs. Specifically, we first incorporate an inter-slice anchor-based filtering module to effectively manage the high temporal redundancy inherent in 3D medical volumes. Additionally, we develop a dynamic information nucleus selection strategy to achieve adaptive compression across slices with varying information densities by quantifying the cumulative attention weights contributed by visual tokens. To validate the efficacy of our proposed method, we conduct extensive experiments on three 3D medical benchmarks and three VLMs. Our results reveal substantial token redundancy in current medical vision-language models. Notably, MedPruner allows MedGemma-1.5 to maintain or even exceed its original performance while using less than 5% of the origi-

nal visual tokens. This extreme compression highlights a highly skewed attention distribution in medical VLMs, demonstrating that dynamic token selection is essential for effectively filtering background noise and capturing critical diagnostic signals. Our contributions can be summarized as follows:

- We introduce MedPruner, which, to the best of our knowledge, is the first training-free and model-agnostic hierarchical token pruning framework tailored to 3D medical VLMs.
- We employ a training-free, two-stage mechanism to dynamically prune redundant information at both the slice and token levels.
- We conduct comprehensive experiments across 3 datasets and 3 VLMs, consistently demonstrating the effectiveness and robustness of our approach.

2 Methods

2.1 3D Vision-Language Models

Existing VLM architectures generally consist of three components: a visual encoder, a modality projector, and an LLM backbone. In 3D medical imaging, a common approach is to slice a 3D volume $V \in \mathbb{R}^{D\times H\times W}$ along the axial axis, resulting in a sequence of 2D slices $\mathcal{I} = \{I_1, I_2, \ldots, I_D\}$. Each slice I_i is independently processed by a visual encoder $\mathcal{E}_v$ and a modality projector $\mathcal{P}$. The final visual representation H_v is the concatenation of the projected tokens from all slices:

$$H_v = \bigoplus_{i=1}^{D} \mathcal{P}(\mathcal{E}_v(I_i)), \tag{1}$$

where $\bigoplus$ denotes sequence concatenation. If each slice yields M tokens, the total visual token count $D \times M$ leads to a severe sequence length explosion. To mitigate this, we propose MedPruner, a hierarchical framework consisting of two components: Inter-slice Anchor-based Filtering (IAF), which reduces inter-slice redundancy by tracking content evolution, and Dynamic Information Nucleus Selection (DINS), which adaptively prunes uninformative tokens based on attention distribution. The overview of MedPruner is shown in Fig. 1.

2.2 Inter-slice Anchor-based Filtering

To effectively manage the high temporal redundancy inherent in 3D medical volumes, we introduce Inter-slice Anchor-based Filtering (IAF). Rather than employing a static or fixed-interval sampling rate, IAF utilizes a dynamic, content-aware strategy to adaptively identify slices that provide significant anatomical information. The process operates sequentially by maintaining a dynamic anchor slice, denoted as I_{anc}. We initialize this filtering process by setting the first slice of the volume as the starting anchor, such that $I_{anc} = I_1$. As we traverse the remaining sequence, we evaluate the informational divergence of each incoming

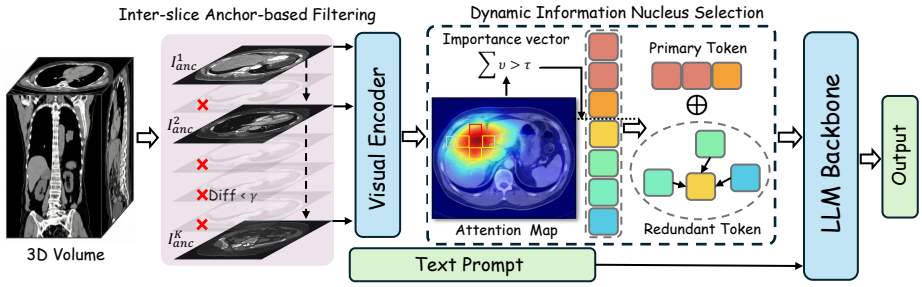

Fig. 1: The overview of our MedPruner.

slice I_i ($i > 1$) relative to the current active anchor. This divergence is quantified using the pixel-wise mean L_1 distance:

$$\Delta(I_i, I_{anc}) = \frac{1}{N} \sum_{j=1}^{N} |I_{i,j} - I_{anc,j}|, \tag{2}$$

where N denotes the total number of pixels in a slice. This distance $\Delta(I_i, I_{anc})$ serves as a proxy for morphological change. A small distance indicates that the structural information in I_i is already well-represented by I_{anc}, rendering the current slice redundant. The core of the IAF mechanism lies in its threshold-driven update logic. We continuously compare the distance $\Delta(I_i, I_{anc})$ against a predefined sensitivity threshold γ. If the distance exceeds γ, the slice I_i is deemed to contain significant novel anatomical features. Consequently, I_i is preserved and immediately takes over as the new active anchor ($I_{anc} \leftarrow I_i$) for evaluating the remaining sequence. Conversely, if the distance falls below the threshold, the slice lacks sufficient new information and is entirely filtered out. Through this continuous traversal and updating process, the original dense volume of length D is distilled down to a sparse, informative subsequence comprising only the dynamically preserved anchor frames $\mathcal{I}_{filtered} = \{I^1_{anc}, I^2_{anc}, \ldots, I^k_{anc}\}$, k is the final number of retained slices ($k < D$). By discarding the non-anchor intermediate frames, IAF compresses the sequence length. This ensures that the model concentrates its computational budget solely on regions with high structural variance, including the boundaries of organs or the appearance of lesions. Consequently, it prepares a highly condensed and representative sequence for the subsequent token-level optimization.

2.3 Dynamic Information Nucleus Selection

Following the inter-slice filtering, we further optimize the token density within each preserved slice by deriving token importance directly from the self-attention layers of the vision encoder. First, we calculate the attention score for each head

as follows:

$$S_h = \text{Softmax}\left(\frac{Q_h K_h^\top}{\sqrt{D_h}}\right), \tag{3}$$

where D_h is the head dimension, and Q_h and K_h represent the query and key matrices, respectively. By averaging these scores across all heads, we obtain an aggregated attention matrix S_{avg}. To evaluate the raw significance of each visual token, we compute the average of S_{avg} along the sequence dimension, yielding an initial importance vector $\hat{v} \in \mathbb{R}^M$, where M is the number of tokens in the slice. To transform these raw scores into a comparable probability distribution and allow for adjustable selection sensitivity, we apply a temperature-scaled softmax normalization such that:

$$v_i = \frac{\exp(\hat{v}_i/T)}{\sum_{j=1}^{M} \exp(\hat{v}_j/T)}, \tag{4}$$

where the temperature coefficient T serves as a smoothing factor; a lower T sharpens the distribution to emphasize high-scoring tokens, while a higher T retains broader contextual information.

The inherent heterogeneity of information density across medical slices suggests that a fixed pruning ratio is suboptimal, as it fails to distinguish between salient anatomical features and uninformative backgrounds. To address this, we utilize a strategy inspired by nucleus filtering to adaptively capture the essential core of each slice by sorting the normalized weights in v in descending order to obtain v_{sorted}. We then dynamically select the minimal set of top-ranked tokens, designated as *primary tokens* $\mathcal{K}$, whose cumulative attention mass reaches a predefined information threshold τ:

$$\mathcal{K} = \{\text{Top-}k \text{ tokens} \mid \min k \quad \text{s.t.} \quad \sum_{j=1}^{k} v_{\text{sorted},j} \geq \tau\}. \tag{5}$$

By anchoring the selection boundary to the cumulative probability mass, this mechanism ensures that slices with concentrated attention are significantly compressed, while those with dispersed, critical details retain a larger token set to maintain diagnostic integrity. Token dimensions remain unchanged.

Finally, the remaining unselected tokens are treated as *redundant tokens*. To retain the global structural context without increasing the sequence length, we apply a bipartite matching and clustering operation following [30]. These clustered redundant tokens are subsequently concatenated with the primary tokens and passed to the modality projector for the final VLM inference.

3 Experiments

3.1 Experiment Setting

Datasets. In our experiments, we evaluate our approach on three 3D medical benchmarks: M3D [2], 3D-RAD [7], and AMOS-MM [10]. Both M3D and 3D-RAD are comprehensive 3D Medical Visual Question Answering (VQA) datasets

Table 1: Quantitative resluts on the 3DRad [7] and M3D [2] VQA datasets. **R-Rate** denotes the token retention rate. **Bold** and underlined values represent the best and second-best performance, respectively.

Method	M3D						3DRad					
	Acc	Rouge-1	Rouge-L	BLEU-1	BLEU-4	R-Rate	Acc	Rouge-1	Rouge-L	BLEU-1	BLEU-4	R-Rate
Hulu-Med-7B [11]												
Original	75.317	34.188	34.047	31.820	9.541	100.00%	78.767	23.280	22.917	18.760	5.695	100.00%
Hulu-L1 [11]	**77.790**	47.334	47.167	45.234	12.491	66.03%	78.082	23.765	23.487	20.389	6.280	45.06%
VisionZip [30]	77.623	**47.596**	**47.418**	**45.451**	12.549	69.96%	79.232	25.618	25.336	22.391	6.949	69.74%
HiPrune [18]	77.616	47.072	46.199	45.417	12.574	**22.30%**	79.061	25.721	25.559	21.339	6.759	**22.30%**
MedPruner	77.452	47.434	47.262	45.304	**12.580**	52.10%	**79.280**	**26.247**	**25.982**	**22.865**	**7.123**	51.88%
MedGemma1.5-4B [23]												
Original	32.717	6.797	5.888	3.741	0.835	100.00%	59.155	4.865	4.094	2.629	0.516	100.00%
Hulu-L1 [11]	44.638	7.810	4.473	2.779	0.651	65.98%	57.868	5.009	1.596	1.034	0.217	44.89%
VisionZip [30]	**45.428**	7.925	4.364	2.707	0.639	69.98%	56.537	5.091	1.155	0.754	0.165	69.82%
HiPrune [18]	44.420	8.201	4.447	2.792	0.664	22.30%	55.959	**6.109**	2.045	**1.365**	**0.304**	22.30%
MedPruner	43.718	**8.983**	**5.845**	**3.923**	**1.005**	**4.87%**	**60.843**	5.168	**2.057**	1.317	0.277	**4.62%**

that support both open-ended and closed-ended evaluations. Specifically, M3D queries detailed anatomical and pathological attributes, such as imaging planes, contrast phases, specific organs, and abnormalities. 3D-RAD focuses specifically on radiology CT scans and introduces complex reasoning challenges across six diverse VQA tasks, including anomaly detection, medical computation, and multi-stage temporal diagnosis. Additionally, we utilize the AMOS-MM benchmark, which comprises CT and MRI of abdominal organs, primarily to assess the model's performance in 3D medical report generation.

Implementation Details. We evaluate MedPruner across three VLMs which support 3D medical imaging input, including a general-purpose model, Qwen3-VL-8B [3], and two medical domain-specific models, Hulu-Med-7B [11] and MedGemma-1.5-4B [23]. For the evaluation metrics, we utilize Accuracy for closed-set VQA tasks to rigorously measure classification performance. For open-set reasoning and 3D medical report generation, we employ standard Natural Language Generation metrics including BLEU [21] (BLEU-1, BLEU-4), Rouge [16] (Rouge-1, Rouge-L), and METEOR [4]. All experiments are conducted on $8 \times$ NVIDIA H20 GPUs. Hyperparameters are set as $\gamma = 0.05$ and $\tau = 0.9$.

3.2 Comparison Results

In this section, we conduct a comprehensive evaluation of MedPruner against three existing training-free token reduction methods: the L1-compression method proposed in Hulu-Med [11] (Hulu-L1), VisionZip [30], and HiPrune [18].

Table 1 presents quantitative results on the 3DRad and M3D VQA datasets using Hulu-Med and MedGemma-1.5. In cases where datasets involve massive slice counts, such as M3D with an average of 87 and a maximum of over 600 slices, MedPruner plays a critical role in mitigating information overload. Notably, on the M3D dataset, MedPruner frequently outperforms the uncompressed baseline.

Table 2: Quantitative results on the AMOS-MM dataset [10]. **R-Rate** denotes the token retention rate. **Average** represents the mean percentage of performance across all metrics relative to the original model. **Speed** indicates the average processing time per sample in seconds (s). **Bold** and underlined values represent the best and second-best performance, respectively.

Method	Rouge-1	Rouge-L	BLEU-1	BLEU-4	METEOR	Average↑	R-Rate↓	Speed↓
Hulu-Med-7B [11]								
Original	39.518	28.763	37.317	13.375	31.729	100.00%	100.00%	9.212
Hulu-L1 [11]	33.063	24.826	30.754	9.799	26.212	81.65%	**16.35%**	8.039
VisionZip [30]	38.519	28.216	36.876	12.705	30.738	97.25%	49.69%	8.435
HiPrune [18]	39.093	27.956	35.764	12.513	31.086	96.70%	22.30%	9.111
MedPruner	**39.255**	**28.860**	**37.339**	**13.120**	**31.136**	**99.19%**	54.20%	**7.931**
MedGemma1.5-4B [23]								
Original	13.791	9.906	5.845	0.253	9.721	100.00%	100.00%	38.001
Hulu-L1 [11]	**13.999**	9.787	5.843	0.241	9.782	99.25%	16.07%	36.193
VisionZip [30]	13.830	10.043	**6.050**	**0.260**	10.129	**102.42%**	49.61%	36.682
HiPrune [18]	13.369	9.782	5.828	0.236	9.939	98.21%	21.85%	37.619
MedPruner	13.706	**10.234**	5.985	0.235	**10.252**	100.65%	**2.46%**	**35.889**
Qwen3-VL-8B [3]								
Original	18.717	18.317	26.994	1.077	18.821	100.00%	100.00%	11.179
Hulu-L1 [11]	18.201	17.808	25.839	0.870	17.788	93.10%	**16.50%**	9.569
VisionZip [30]	18.590	**18.047**	25.604	0.925	18.116	94.98%	49.61%	10.654
HiPrune [18]	18.217	17.255	25.886	0.919	**18.600**	94.33%	22.30%	10.052
MedPruner	**19.469**	17.908	**26.119**	**0.937**	17.648	**95.86%**	43.28%	**9.044**

This occurs because directly concatenating hundreds of slices introduces significant background noise and redundant structures that can overwhelm the LLM's context window. By filtering these uninformative tokens, MedPruner achieves the highest BLEU-4 scores on M3D (12.580) and 3DRad (7.123) with Hulu-Med while maintaining competitive accuracy and reducing the token retention rate (R-Rate) to approximately 52%.

Table 2 reports the performance on the AMOS-MM dataset. Across all tested architectures, MedPruner achieves the optimal balance between accuracy and efficiency, delivering the fastest inference speeds while maintaining exceptional performance. In several instances, it even surpasses the original baselines, such as reaching a 100.65% average score on MedGemma-1.5. These results highlight the model-agnostic robustness of MedPruner in optimizing diverse VLMs under real-world computational constraints.

A notable observation is the extreme token compression achieved by MedPruner on the MedGemma-1.5 model. As shown in the tables, MedPruner maintains high performance while requiring fewer than 5% of the visual tokens across all three datasets, reaching an incredibly low R-Rate of 2.46% on AMOS-MM. Upon analyzing the model's behavior, we discovered that MedGemma-1.5's attention weights are highly concentrated on a single or a very small subset of tokens. By dynamically anchoring the selection threshold to the cumulative attention mass, our module naturally adapts to this highly skewed distribution. This phenomenon strongly validates the necessity of our dynamic selection strat-

egy. Unlike fixed-ratio methods that arbitrarily discard useful tokens or retain unnecessary ones (e.g., HiPrune fixed at 22.30%), MedPruner intelligently scales the token retention rate based on the intrinsic attention distribution, ensuring optimal efficiency tailored to each specific slice and model architecture.

3.3 Ablation Study

Table 3: Ablation study on the AMOS-MM dataset [10] using the Hulu-Med-7B model [11].

IAF	Primary	Redundant	Average↑	R-Rate↓	Speed↓
✗	✗	✗	100.00%	100.00%	9.212
✓	✗	✗	92.13%	60.33%	**7.751**
✗	✓	✗	98.73%	83.11%	8.894
✗	✓	✓	**100.07%**	88.14%	9.586
✓	✓	✓	99.19%	**54.20%**	7.931

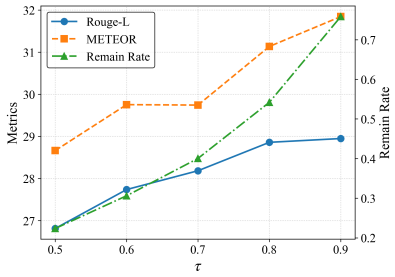

Fig. 2: Ablation study of τ.

Component Analysis. We first evaluate the contribution of each MedPruner component on the AMOS-MM dataset, with results summarized in Table 3. As shown in Table 3, IAF lowers average inference time from 9.2s to 7.7s. While this slice-level filtering initially leads to a performance drop, the introduction of Primary token selection and Redundant token clustering effectively restores diagnostic accuracy. Specifically, incorporating primary tokens and redundant clustering recovers the Average score to over 100% of the baseline. The full MedPruner configuration achieves an optimal balance, maintaining 99.19% of the original performance while reaching the strongest compression with a 54.20% R-Rate.

Sensitivity of Information Threshold τ. Fig. 2 illustrates the impact of the information threshold τ on the Hulu-Med-7B model. As τ increases, the token retention rate rises, leading to improved Rouge-L and METEOR scores. However, these performance gains gradually plateau, indicating that the primary tokens identified by our nucleus selection strategy have already captured the most critical diagnostic features. This confirms that further increasing the token count yields diminishing returns, validating the efficiency of our adaptive selection mechanism.

4 Conclusion

In this paper, we propose MedPruner, a training-free hierarchical pruning framework that bridges the gap between high-performance 3D medical VLMs and the constraints of real-time clinical deployment. By adaptively managing the

non-uniform information density across volumetric data, our approach ensures that critical diagnostic details are prioritized over redundant anatomical backgrounds. Extensive evaluations show that MedPruner improves the efficiency-performance trade-off without compromising diagnostic integrity, offering a scalable and model-agnostic solution for the practical integration of VLMs into complex medical workflows.

Acknowledgments. This work was supported by National Natural Science Foundation of China (No. 62522124), Hong Kong Research Grants Council General Research Fund 14206725 and Hong Kong Innovation and Technology Commission Innovation and Technology Fund PRP/082/24FX.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L.V., Fintelmann, F., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv preprint arXiv:2511.17803 (2025)
2. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3D: Advancing 3D medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of ACL Workshop. pp. 65–72 (2005)
5. Chen, J., Gui, C., Ouyang, R., Gao, A., Chen, S., Chen, G.H., Wang, X., Cai, Z., Ji, K., Wan, X., et al.: Towards injecting medical visual knowledge into multimodal llms at scale. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 7346–7370 (2024)
6. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
7. Gai, X., Liu, J., Li, Y., Meng, Z., Wu, J., Liu, Z.: 3D-RAD: A comprehensive 3D radiology Med-VQA dataset with multi-temporal analysis and diverse diagnostic tasks. arXiv preprint arXiv:2506.11147 (2025)
8. He, S., Nie, Y., Chen, Z., Cai, Z., Wang, H., Yang, S., Chen, H.: MedDR: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. CoRR (2024)
9. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
10. Ji, Y., Bai, H., Ge, C., Yang, J., Zhu, Y., Zhang, R., Li, Z., Zhanng, L., Ma, W., Wan, X., et al.: AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. Advances in NeurIPS **35**, 36722–36732 (2022)

11. Jiang, S., Wang, Y., Song, S., Hu, T., Zhou, C., Pu, B., Zhang, Y., Yang, Z., Feng, Y., Zhou, J.T., et al.: Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. arXiv preprint arXiv:2510.08668 (2025)
12. Khaki, S., Guo, J., Tang, J., Yang, S., Chen, Y., Plataniotis, K.N., Lu, Y., Han, S., Liu, Z.: SparseVILA: Decoupling visual sparsity for efficient vlm inference. In: Proceedings of the ICCV. pp. 23784–23794 (2025)
13. Lai, H., Jiang, Z., Zhang, K., Yao, Q., Wang, R., He, Z., Tao, X., Wei, W., Zhou, S.K.: Med3D-R1: Incentivizing clinical reasoning in 3D medical vision-language models for abnormality diagnosis. arXiv preprint arXiv:2602.01200 (2026)
14. Li, C.Y., Chang, K.J., Yang, C.F., Wu, H.Y., Chen, W., Bansal, H., Chen, L., Yang, Y.P., Chen, Y.C., Chen, S.P., et al.: Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nature Communications **16**(1), 2258 (2025)
15. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
16. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text summarization branches out. pp. 74–81 (2004)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in NeurIPS **36**, 34892–34916 (2023)
18. Liu, J., Du, F., Zhu, G., Lian, N., Li, J., Chen, B.: HiPrune: Training-free visual token pruning via hierarchical attention in vision-language models. arXiv preprint arXiv:2508.00553 (2025)
19. Liu, S., Zheng, B., Chen, W., Peng, Z., Yin, Z., Shao, J., Hu, J., Yuan, Y.: EndoBench: A comprehensive evaluation of multi-modal large language models for endoscopy analysis (2025)
20. Murari Vepa, A., Yu, Y., Gan, J., Cuturrufo, A., Li, W., Wang, W., Scalzo, F., Sun, Y.: A multimodal LLM approach for visual question answering on multiparametric 3D brain MRI. arXiv e-prints pp. arXiv–2509 (2025)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the ACL. pp. 311–318 (2002)
22. Peng, Z., Wang, C., Liu, S., Liang, Z., Ye, Z., Ju, M.J., Woo, P.Y., Yuan, Y.: Omnibrainbench: A comprehensive multimodal benchmark for brain imaging analysis across multi-stage clinical tasks. In: Proceedings of the CVPR. pp. 42732–42743 (2026)
23. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: MedGemma technical report. arXiv preprint arXiv:2507.05201 (2025)
24. Tanno, R., Barrett, D.G., Sellergren, A., Ghaisas, S., Dathathri, S., See, A., Welbl, J., Lau, C., Tu, T., Azizi, S., et al.: Collaboration between clinicians and vision–language models in radiology report generation. Nature Medicine **31**(2), 599–608 (2025)
25. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical AI. Nejm Ai **1**(3), AIoa2300138 (2024)
26. Wang, W., Ma, Z., Ding, M., Zheng, S., Liu, S., Liu, J., Ji, J., Chen, W., Li, X., Shen, L., et al.: Medical reasoning in the era of llms: a systematic review of enhancement techniques and applications. arXiv preprint arXiv:2508.00669 (2025)
27. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. Nature Communications **16**(1), 7866 (2025)

28. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (2024)
29. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)
30. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: VisionZip: Longer is better but not necessary in vision language models. In: Proceedings of the CVPR. pp. 19792–19802 (2025)
31. Zambrano Chaves, J.M., Huang, S.C., Xu, Y., Xu, H., Usuyama, N., Zhang, S., Wang, F., Xie, Y., Khademi, M., Yang, Z., et al.: A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. Nature Communications **16**(1), 3108 (2025)