

OmniV-Med: A Scalable Training Recipe for Unified Medical Vision-Language Understanding

Songtao Jiang^{1†} Yuan Wang^{1†} Chenyi Zhou¹
Tianxiang Hu¹ Fucheng Niu¹ Zuozhu Liu^{1*}

¹ZJU-Angelalign R&D Center for Intelligence Healthcare, Zhejiang University, China

[†]Equal contribution ^{*}Corresponding author

{songtao.22, zuozhuliu}@intl.zju.edu.cn

Abstract

Practical deployment of medical vision-language models requires a single model that interprets heterogeneous clinical visual inputs: 2D images, 3D volumes, and medical videos. Yet how to train such unified models remains underexplored, as these modalities differ in spatial-temporal structure, token length, and data availability. We present OmniV-Med, a scalable training recipe built on four empirical findings. First, broad modality coverage requires balancing modality diversity with task-level supervision, motivating OmniV-Med-Instruct, a 252K corpus spanning 14 medical modalities and 11 clinical tasks. Second, we identify a modality-ordering effect: introducing 3D and video too early destabilizes optimization and weakens 2D grounding, motivating a progressive curriculum that first builds 2D alignment before extending to spatial-temporal inputs. Third, we observe an emergent spatiotemporal transfer effect: once strong 2D patch representations are in place, lightweight rotary position adaptation extrapolates them to 3D slices and video frames, letting a single 2D ViT replace modality-specific encoder stacks. Fourth, the imbalance in sequence length across modalities and the redundancy of long volumes and videos hinder scaling; our Redundancy-Guided Token Pruning (RGTP) exploits inter-slice and inter-frame redundancy to cut visual tokens by 60% while preserving performance. Across nine 2D, 3D, and video medical benchmarks, OmniV-Med achieves robust performance. Extensive component-wise ablations systematically refine the design space, yielding a simple yet efficient and scalable recipe for training high-performance unified Med-VLMs. Code is released at <https://github.com/ZJUI-AI4H/Hulu-Med>.

1 Introduction

Practical medical vision-language systems must interpret heterogeneous clinical visual inputs, including 2D medical images, 3D volumetric scans,

and surgical videos. Recent medical VLMs have advanced 2D biomedical visual understanding (Li et al., 2023a; Moor et al., 2023; Jiang et al., 2024; Zhang et al., 2024), 3D radiology (Wu et al., 2025; Bai et al., 2024), and surgical-video analysis (Seeni-vasan et al., 2022, 2023; Jin and Jeong, 2024). Most, however, remain specialized by input type. This fragmentation limits clinical workflows in which one model may need to process endoscopic video, volumetric CT or MRI, and conventional 2D images under textual instructions (Wang et al., 2023).

A common solution is to assign each modality its own encoder: image encoders for 2D inputs, 3D encoders for volumes, and temporal encoders for videos. However, our empirical exploration suggests that this modality-specific paradigm is not necessarily optimal for medical VLMs. VLM training depends critically on aligning visual representations with a language model; separate encoders fragment the representation space, increase alignment burden, and introduce substantial training and inference overhead. Moreover, specialized 3D and video encoders are usually pretrained on much smaller corpora than CLIP-style 2D ViTs, whose representations are often stronger and easier to align. This raises a key question: *can a single ViT serve as a unified visual backbone for 2D images, 3D volumes, and videos?*

Unified Med-VLM training also faces challenges in learning dynamics and scalability. High-quality 3D/video instruction data is far more limited than 2D medical image data, and we observe that introducing spatial-temporal modalities too early can destabilize optimization and weaken 2D performance. Meanwhile, 3D volumes and videos produce much longer token sequences, causing imbalanced training throughput and potential visual-token-length bias. Our experiments show that much of this cost comes from redundant inter-slice and inter-frame information. These observations mo-

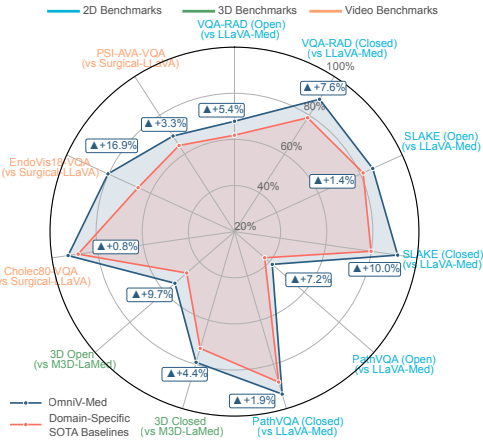


Figure 1: Comparison with state-of-the-art methods on medical 2D/3D image and video benchmarks.

tivate two further questions: *how can the model extend from 2D to 3D/video inputs without disrupting 2D grounding, and how can long visual inputs be processed under a practical token budget while preserving clinical details?*

We answer these questions with OmniV-Med, a scalable training recipe for unified medical visual understanding across text, 2D images, 3D volumes, and videos. We construct OmniV-Med-Instruct, a 252K-sample multimodal instruction corpus spanning 14 medical modalities and 11 clinical tasks. Architecturally, OmniV-Med adapts a single 2D ViT into a unified visual backbone by representing single images, multi-image inputs, 3D volumes, and videos as variable-length patch sequences encoded with 2D rotary position-adaptive embeddings. This enables inputs with different resolutions, lengths, and modality structures to share one representation space, analogous to variable-length text tokens in language models.

To enable efficient and stable training, we adopt a progressive curriculum: OmniV-Med first builds 2D medical vision-language alignment and instruction-following ability, then extends this foundation to spatiotemporal inputs. Our ablations show a clear modality-ordering effect and provide empirical evidence that strong 2D patch representations transfer to 3D slices and video frames. For scalable long-input processing, we introduce **Redundancy-Guided Token Pruning (RGTP)**, which computes L_1 distances between visual patches from adjacent slices or frames and prunes tokens below a pre-defined threshold. RGTP reduces visual tokens by 60% on average while maintaining robust performance, making large 3D inputs and hour-scale medical videos tractable on consumer GPUs.

We evaluate OmniV-Med on both training-distribution and held-out benchmarks covering 2D medical VQA, 3D volumetric understanding and report generation, surgical video VQA, and long-video captioning. As shown in Figure 1, OmniV-Med consistently improves over strong domain-specific baselines on the representative benchmarks summarized there. The 7B model achieves strong overall performance, while the 1.5B variant is trainable on $8 \times$ RTX 3090 GPUs. Component-wise ablations foreground the recipe’s central design choices, and additional reliability tests expose both encouraging safety results and remaining clinical limitations.

2 Related Work

Medical vision-language models. Medical VLMs have largely developed along modality-specific lines. LLaVA-Med, Med-Flamingo, Med-MoE, HuatuoGPT-Vision, and BiomedGPT study 2D biomedical instruction following or task transfer (Li et al., 2023a; Moor et al., 2023; Jiang et al., 2024; Chen et al., 2024; Zhang et al., 2024). RadFM and M3D-LaMed extend language-conditioned modeling to 3D radiology, but rely on volumetric processing components (Wu et al., 2025; Bai et al., 2024). Surgical-VQA, SurgicalGPT, and Surgical-LLaVA target surgical video understanding (Seenivasan et al., 2022, 2023; Jin and Jeong, 2024). A parallel line of work broadens coverage further, toward general video reasoning, generalist medical models, and other biomedical data modalities (Jiang et al., 2025, 2026; Liu et al., 2026). OmniV-Med complements these systems with a single visual backbone and a controlled study of how to order and balance 2D, 3D, and video training.

Efficient visual-token processing. Generic efficiency methods either merge similar tokens, as in ToMe (Bolya et al., 2023), or restrict token interactions through sparse attention (Child et al., 2019). RGTP instead uses redundancy specifically across adjacent slices or frames. We compare these alternatives under a matched token budget and separately test whether pruning preserves small, diagnostically important lesions.

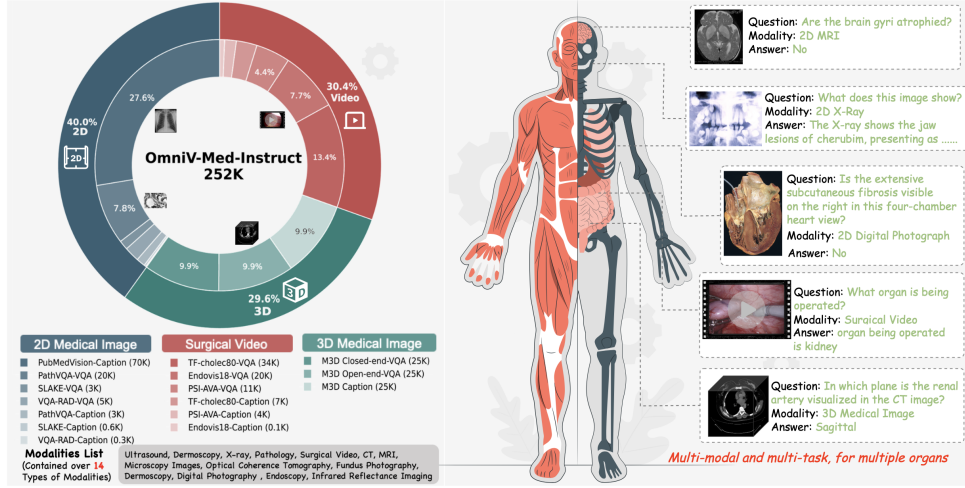


Figure 2: Data distribution, coverage and examples of OmniV-Med-Instruct.

3 OmniV-Med

3.1 OmniV-Med-Instruct Construction

Figure 2 summarizes the modality coverage, sample sizes, tasks, and statistics of OmniV-Med-Instruct. The dataset covers 14 medical imaging modalities across diverse diseases, including common modalities such as CT, MRI, X-ray, and pathology, as well as domain-specific modalities such as dermoscopy and endoscopy. For 2D vision-language instructions, we curate mainstream medical datasets covering Medical VQA (Med-VQA) and Medical Report Generation (MRG), including VQA-RAD (Lau et al., 2018), MIMIC-Ext-MIMIC-CXR-VQA (Bae et al., 2024), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020). We further sample 70K instruction-following examples from PubMedVision (Chen et al., 2024) to strengthen instruction-following ability. For 3D vision-language data, we sample 75K examples from M3D (Bai et al., 2024), covering tasks such as image-text retrieval, MRG, VQA, and positioning.

For medical video-language data, we construct an open-source instruction-following dataset from existing surgical video resources. We build video VQA examples from Cholec80 (Seenivasan et al., 2022), PSI-AVA (Valderrama et al., 2022), and Endovis18 (Seenivasan et al., 2023) by pairing consecutive-frame video segments with their original questions and answers, enabling the model to learn from continuous surgical sequences. For video captioning, since open-source medical-specific video VLMs are unavailable, we use Qwen2.5-VL-72B (Bai et al., 2025) to generate

captions by combining frame-level QA annotations with corresponding video segments. We apply rejection sampling (Touvron et al., 2023) to filter high-quality video-text pairs, with details provided in Appendix A.2 and Figure 8.

3.2 Benchmark Decontamination

To rule out test-set leakage and ensure fair evaluation, we apply a multi-stage de-duplication pipeline to OmniV-Med-Instruct against every downstream benchmark. For each test split we fingerprint every image by perceptual hash and every text by normalized n -gram MinHash, and discard any training sample whose image ≥ 0.95 similarity or whose QA ≥ 0.85 Jaccard overlap with any test entry. For 3D volumes and surgical videos we additionally drop any training clip sharing source patient or procedure IDs with a test split.

3.3 Training Pipeline

Our training process consists of three stages: 1) 2D Med-VL alignment, enhancing the model’s understanding of 2D medical images with VL caption pairs; 2) medical instruction-following tuning, enabling the model to follow complex instructions based on medical image understanding; and 3) medical mixed-modal tuning, improving the model’s capacity to interpret diverse spatial and temporal VL inputs. OmniV-Med-7B uses Qwen2.5-7B-Instruct, and OmniV-Med-Tiny-1.5B uses Qwen2.5-1.5B-Instruct, as their language backbones. Hyperparameters and stage-wise trainability are in Appendix A.1.

2D Med-VL Alignment. In stage 1, we align medical vision and language in-

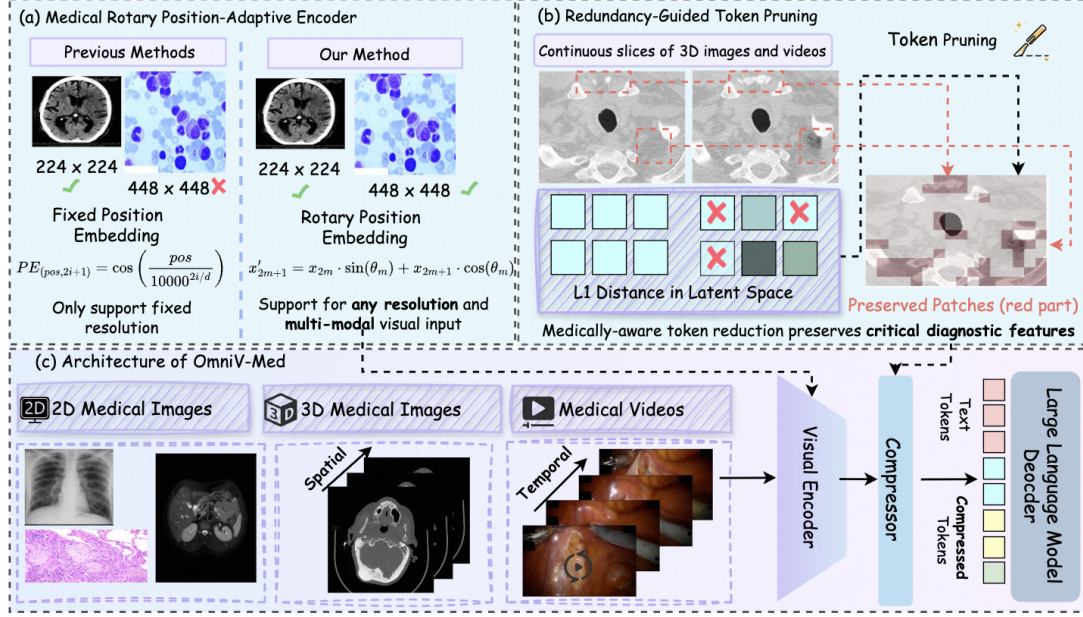


Figure 3: OmniV-Med comprises (a) a rotary position-adaptive encoder for modalities and resolutions of different sizes, (b) RGTP for redundant frames and slices, and (c) a shared vision-language architecture.

puts using 647K image-caption pairs from PubMedVision_Alignment (Chen et al., 2024). We freeze the LLM backbone and fine-tune the pre-trained SigLIP (Zhai et al., 2023) along with the MLP projector as the vision encoder. To better handle different image resolutions, we replace the original fixed position embedding in SigLIP with a rotary position embedding, creating the *rotary position-adaptive encoder*. The rotary position embedding is defined as:

$$\mathbf{R}_{\theta,d} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}, \quad (1)$$

where θ represents the rotation angle and d denotes the embedding dimension. For a given position m , the rotary embedding is applied to the query $\mathbf{q}_m$ and key $\mathbf{k}_n$ vectors as:

$$\mathbf{q}'_m = \mathbf{R}_{\theta,d}(m)\mathbf{q}_m, \quad \mathbf{k}'_n = \mathbf{R}_{\theta,d}(n)\mathbf{k}_n. \quad (2)$$

This approach enhances the model’s ability to capture relative positional relationships, improving its adaptability to diverse medical visual inputs.

Medical Instruction-Following Tuning. In stage 2, we make all model parameters trainable to improve the model’s capability in following complex textual instructions based on visual inputs. We train the model using 647K instruction-answer pairs from PubMedVision_SFT (Chen et al., 2024). Both stage 1 and stage 2 focus on 2D images, as our experiments indicate that a strong foundational

understanding of 2D images is crucial to extend Med-VLMs to complex 3D images and videos. By the end of this stage, the model can effectively handle various visual tasks related to 2D images.

Medical Mixed-Modal Tuning. In stage 3, the goal is to expand the model’s capabilities from 2D images to 3D images (spatial) and videos (temporal). However, processing 3D images and videos is computationally expensive. In response, we introduce a token pruning mechanism to remove redundant information while preserving critical content. We found that consecutive slices in 3D medical images and adjacent frames in medical videos often contain repetitive or similar information. Hence, we compress visual tokens by calculating the L_1 distance between patches of consecutive views or frames. Tokens with an L_1 distance below a threshold of 0.1 are identified as redundant and removed. This pruning method effectively reduces unnecessary visual data without impacting model performance. The model is trained with all 252K samples in OmniV-Med-Instruct with a standard autoregressive loss. All parameters are trainable to fully unleash the model’s potential, enabling it to extrapolate visual understanding to spatiotemporal dimensions and enhance downstream performance across diverse tasks. Our approach achieves SoTA accuracy with improved efficiency, enhancing its practicality for real-world medical tasks.

Method	VQA-RAD		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
<i>Representative & SoTA methods with numbers reported in the literature (Non-VLM Based Methods)</i>						
VL Encoder-Decoder (Bazi et al., 2023)	-	82.47	-	-	-	85.61
Q2ATransformer (Liu et al., 2023)	-	81.20	-	-	54.85	88.85
Prefix T. Medical LM (van Sonsbeek et al., 2023)	-	-	-	82.01	-	87.00
PubMedCLIP (Eslami et al., 2023)	-	80.00	-	82.50	-	-
BiomedCLIP (Zhang et al., 2023a)	-	79.80	-	89.70	-	-
M2I2 (Li et al., 2023b)	-	83.50	-	91.10	-	88.00
BiomedGPT-S (Zhang et al., 2024)	13.40	57.80	66.50	73.30	10.70	84.20
BiomedGPT-M (Zhang et al., 2024)	53.60	65.07	78.30	86.80	12.5	85.70
CLIP-ViT w/ GPT2-XL	-	-	84.30	82.10	40.0	87.00
<i>Results (VLM Based Methods)</i>						
HuatuoGPT-Vision-34B (Chen et al., 2024)	-	63.80	-	74.50	-	59.90
HealthGPT-14B (Lin et al., 2025)	-	73.70	-	76.40	-	85.90
Lingshu-7B (LASA Team et al., 2025)	35.44	77.29	78.54	77.67	31.72	77.00
LLaVA-Med (LLama7B) (Li et al., 2023a)	61.52	84.19	83.08	85.34	37.95	91.21
LLaVA-Med (Vicuna7B) (Li et al., 2023a)	<u>64.39</u>	81.98	84.71	83.17	38.87	91.65
Med-MoE-3.6B (Jiang et al., 2024)	58.55	82.72	<u>85.06</u>	85.58	34.74	91.98
Med-MoE-2B (Jiang et al., 2024)	50.08	80.07	83.16	83.41	33.79	91.30
OmniV-Med-Tiny-1.5B	59.22	<u>84.93</u>	84.59	<u>87.98</u>	<u>38.87</u>	<u>92.66</u>
OmniV-Med-7B	67.87	88.24	85.92	91.51	41.68	93.42

Table 1: Performance on Med-VQA tasks. Open-ended columns report recall; closed-ended columns report accuracy, following the LLaVA-Med protocol. **Bold** denotes the best result and underlining the second best.

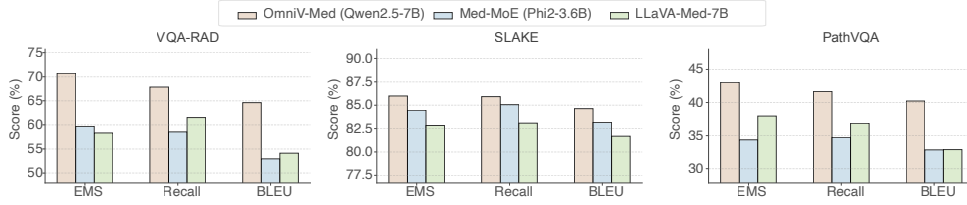


Figure 4: Detailed comparison with additional metrics in open settings. **EMS** denotes *Exact Match Score*. More baseline comparisons are provided in Table 12.

Model	RAD	SLAKE	Path	PMC
GPT-4.1 [†]	65.0	72.2	55.5	55.5
Claude-Sonnet-4 [†]	67.6	70.6	54.2	56.8
Gemini-2.5-Flash [†]	68.5	75.8	55.4	55.4
GPT-5.2	66.7	77.4	53.2	57.6
Claude-Sonnet-4.5	70.5	72.7	55.8	58.7
Gemini-2.5-Pro	69.2	76.0	58.1	59.1
OmniV-Med-7B	78.0	86.8	65.6	66.8

Table 2: Proprietary-model accuracy (%). [†]Values are cited from Lingshu (LASA Team et al., 2025); the remaining values are measured under the same GPT-4.1 judging protocol.

4 Experiment

4.1 Settings

We distinguish evaluations drawn from training distributions from held-out, out-of-distribution (OOD) tests. The in-domain group comprises VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), PathVQA (He et al., 2020), PMC-VQA (Zhang et al., 2023b), M3D-VQA (Bai et al., 2024), Cholec80-VQA (Seenivasan et al., 2022), PSI-AVA-VQA (Valderrama et al., 2022), and EndoVis18-VQA (Seenivasan et al., 2023); test instances are decontaminated, but their source distri-

butions occur during training. The OOD group contains the held-out OmniMedVQA split across eight modalities (Hu et al., 2024), the MMMU Health & Medicine validation track (Yue et al., 2024), and MedLVBench. For multiple-choice evaluation, the model returns the option letter; for open-ended evaluation, it returns a single word or phrase. We report closed-ended accuracy, open-ended recall, and Rouge-L for report or caption generation. Tables 3 and 4 report the held-out multiple-choice results, and Appendix A.4 describes MedLVBench.

We compare OmniV-Med against both domain-specific SoTA baselines and leading proprietary commercial VLMs. For 2D images, we select Med-MoE (Jiang et al., 2024), LLaVA-Med (Li et al., 2023a), BiomedGPT (Zhang et al., 2024), HealthGPT (Lin et al., 2025), and Lingshu (LASA Team et al., 2025) as open-source baselines. For 3D images, we compare with RadFM (Wu et al., 2025) and M3D-LaMed (Bai et al., 2024). In the video domain, we compare against Surgical-VQA (Seenivasan et al., 2022), LV-GPT (Seenivasan et al., 2023) and Surgical-LLaVA (Jin and Jeong, 2024). As an additional reference, we also benchmark

against four proprietary commercial VLMs—GPT-5.2, GPT-4o, Claude-Sonnet-4.5, and Gemini-2.5-Pro—on the four 2D Med-VQA benchmarks.

4.2 Main Results

Broad supporting evidence. OmniV-Med improves over strong modality-specific baselines on in-domain 2D Med-VQA (Table 1), 3D volumetric VQA (Table 6), and surgical VQA (Table 7). The held-out evaluations in Tables 3 and 4, together with MedLVBench, test generalization beyond those training distributions. RGTP cuts visual tokens by 60% on average for 3D and video inputs and enables long-video inference (Figure 11). Table 2 reports a corrected comparison with proprietary VLMs: the three results imported from Lingshu are explicitly attributed, while newer variants are evaluated separately under the standardized protocol in Appendix A.3.

Performance on long video evaluation. To address the absence of long-video evaluation in the medical domain, we construct **MedLVBench** from the PitVis-2023 Challenge (Das et al., 2025), leveraging its dense step- and instrument-level annotations of the endoscopic transsphenoidal approach to construct captioning clips ranging from 10 min to over 1 h (Appendix A.4). Following rubric-based model judging in HealthBench (Arora et al., 2025) and duration-sensitive evaluation in Video-MME (Fu et al., 2025), we use Rouge-L and GPT-4.1-Judge as complementary lexical and semantic metrics; prompts and the evaluation protocol are in Appendix A.3. As shown in Table 5, OmniV-Med yields higher average Rouge-L (+5.62) and GPT-Judge (+0.49) than the strongest evaluated baseline, Gemini-2.5-Pro. Input budgets differ: Gemini-2.5-Pro alone accepts the full one-hour, 1-fps stream, while the other proprietary baselines use at most 512 sampled frames. Figure 5 further shows less degradation with duration, although this benchmark-based result does not establish clinical deployment readiness.

4.3 Ablation Study and Analysis

Ablation of Arbitrary Resolutions. To validate the effectiveness of 2D rotary position embeddings (RoPE) in learning relationships between image patches, we conducted experiments by replacing fixed position embeddings with RoPE while keeping the visual encoder unchanged. As shown in Table 8, using the same data in Stage 1 and Stage 2, and training only on 2D images in Stage 3,

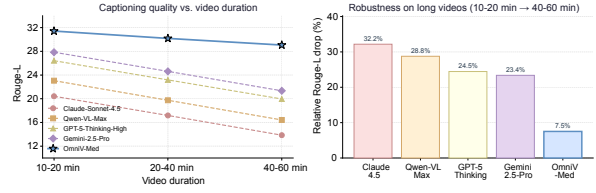


Figure 5: Performance decay with video duration on **MedLVBench**. Left: Rouge-L across duration bins; right: relative Rouge-L drop from 10–20 min to 40–60 min.

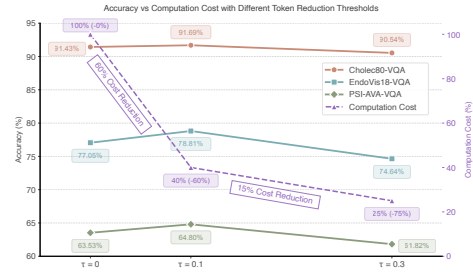


Figure 6: Ablation study on RGTP threshold and cost comparison (Further results in Table 13).

we found that RoPE better captures relative positional relationships between patches, significantly improving performance on 2D tasks.

Ablation of Redundancy-Guided Token Pruning (RGTP). We conduct experiments to evaluate the impact of RGTP on performance and computational cost. As shown in Figure 6, RGTP reduces visual tokens for 3D images and videos by an average of 60%. More importantly, this pruning strategy does not degrade performance; instead, it helps the model focus on more critical regions of the images, leading to improved accuracy. A direct comparison against ToMe (Bolya et al., 2023) and sparse attention (Child et al., 2019) under the same budget is reported in Appendix A.6.

Robustness of Token Pruning on the Small-Lesion Subset. A natural concern with RGTP is whether it harms perception of small but diagnostically critical structures. To probe this, we curate from M3D-VQA a *lesion-focused subset* containing only abnormality- and location-targeted questions whose referenced regions occupy $< 5\%$ of the volume (1,832 questions in total). As shown in Table 14, applying RGTP ($\tau = 0.1$, 60% token reduction) costs only -0.93 points in average closed-end accuracy on this lesion-centric subset, confirming that the simple L_1 -distance heuristic remains robust even under lesion-focused evaluation.

Compared with Modality-Specific Encoder. To evaluate our unified encoding strategy, we replace

Model	CT	FP	MRI	OCT	Der	Mic	X-Ray	US	Avg.
Med-Flamingo	34.6	33.3	27.5	26.0	28.3	28.1	30.1	33.2	30.2
RadFM	33.3	35.0	22.0	31.3	36.3	28.0	31.5	26.1	30.5
LLaVA-Med-7B	25.3	48.4	35.9	42.1	45.2	44.0	31.7	83.7	44.5
Qwen-VL-Chat	51.5	45.4	43.9	54.0	55.4	49.5	63.1	33.5	49.5
Yi-VL-34B	39.8	57.2	51.4	70.5	54.5	61.4	64.2	40.5	54.9
LLaVA-v1.6-7B	40.1	39.5	54.8	58.4	54.0	48.8	53.3	47.9	49.6
LLaVA-v1.6-13B	40.0	43.6	47.4	63.2	58.0	50.5	59.6	42.6	50.6
LLaVA-v1.6-34B	50.6	63.4	60.9	68.4	65.7	62.8	74.7	44.5	61.4
LLaVA-v1.5-LLaMA3-8B	33.0	49.7	53.8	76.0	63.1	48.4	56.6	31.2	48.8
+ LLaVA_Med	60.8	68.5	66.3	79.0	66.6	60.3	73.3	49.3	65.5
HuatuoGPT-Vision-34B	60.8	85.5	66.5	90.0	74.0	71.3	83.8	81.7	76.7
HuatuoGPT-Vision-7B	65.7	82.1	72.5	86.7	70.7	75.7	80.5	77.8	74.0
OmniV-Med-Tiny-1.5B	78.8	78.4	77.1	82.1	71.1	69.9	82.6	73.6	74.5
OmniV-Med-7B	75.0	77.5	84.6	82.3	76.4	72.7	80.9	76.7	78.3

Table 3: Accuracy on the held-out OmniMedVQA split (Hu et al., 2024). **FP**: fundus photography; **MRI**: magnetic resonance imaging; **OCT**: optical coherence tomography; **Der**: dermoscopy; **Mic**: microscopy; **US**: ultrasound.

Model	BMS	CM	DLM	P	PH	MMMU Health & Medicine
Med-Flamingo	29.6	28.1	24.8	25.3	31.2	28.3
RadFM	27.5	26.8	25.8	24.7	29.1	27.0
LLaVA-Med-7B	39.9	39.1	34.6	37.4	34.0	36.9
Qwen-VL-Chat	36.5	31.7	32.7	28.4	34.6	32.7
Yi-VL-34B	49.4	48.9	43.2	40.5	32.0	41.5
LLaVA-v1.6-7B	40.5	36.9	32.1	32.3	26.9	33.1
LLaVA-v1.6-13B	53.6	46.7	33.3	22.2	40.0	39.3
LLaVA-v1.6-34B	56.4	56.0	46.9	46.7	41.7	48.8
LLaVA-v1.5-LLaMA3-8B	42.3	44.0	37.0	34.7	35.2	38.2
+ LLaVA_Med	42.9	46.7	36.7	40.7	46.7	42.8
HuatuoGPT-Vision-7B	57.1	66.7	40.0	59.3	46.7	53.8
HuatuoGPT-Vision-34B	64.6	62.5	50.6	54.1	44.2	54.4
OmniV-Med-Tiny-1.5B	46.4	43.3	46.7	29.6	50.0	43.4
OmniV-Med-7B	64.2	70.0	50.0	48.1	53.3	57.2

Table 4: Accuracy on the MMMU Health & Medicine validation track (Yue et al., 2024). **BMS**: basic medical science; **CM**: clinical medicine; **DLM**: diagnostics and laboratory medicine; **P**: pharmacy; **PH**: public health.

Method	Rouge-L (↑)				GPT-4.1-Judge (↑)			
	10–20 min	20–40 min	40–60 min	Avg.	10–20 min	20–40 min	40–60 min	Avg.
Claude-Sonnet-4.5	20.42	17.18	13.85	17.15	3.45	3.08	2.72	3.08
Qwen-VL-Max	23.05	19.74	16.42	19.74	3.78	3.42	3.06	3.42
GPT-5-Thinking-High	26.42	23.18	19.96	23.19	4.15	3.82	3.48	3.82
Gemini-2.5-Pro	27.85	24.62	21.34	24.60	4.28	3.95	3.62	3.95
OmniV-Med-7B	31.42	30.18	29.06	30.22	4.55	4.45	4.32	4.44

Table 5: Performance on MedLVBench.

Methods	Plane	Phase	Organ	Abnormality	Location	Avg.
Closed-End (Accuracy)						
RadFM (Wu et al., 2025)	19.65	28.70	16.80	18.92	14.88	19.79
M3D-LaMed (Bai et al., 2024)	<u>98.80</u>	79.75	74.75	<u>66.65</u>	58.94	75.78
OmniV-Med-Tiny-1.5B	98.95	<u>91.35</u>	<u>75.15</u>	66.24	60.78	78.49
OmniV-Med-7B	98.75	91.65	76.35	66.80	61.86	79.08
Open-End (Recall)						
RadFM (Wu et al., 2025)	14.24	14.25	14.24	15.64	23.58	16.39
M3D-LaMed-7B (Bai et al., 2024)	<u>98.37</u>	74.41	34.20	15.91	24.00	49.38
OmniV-Med-Tiny-1.5B	98.18	88.80	40.62	16.80	26.04	54.09
OmniV-Med-7B	98.84	<u>86.70</u>	41.36	17.26	26.75	54.18

Table 6: Performance on in-domain M3D-VQA. RadFM is shown only as a reference because it was not trained on this data distribution.

OmniV-Med’s rotary position-adaptive encoder while keeping all other components and data unchanged (videos excluded due to the encoder’s lack with the RadFM 3D encoder (Wu et al., 2025),

Methods	Cholec80-VQA	EndoVis18-VQA	PSI-AVA-VQA
VisualBert (Li et al., 2019)	89.7	61.4	58.5
Block (Ben-Younes et al., 2019)	89.5	60.1	59.9
MFH (Yu et al., 2018)	87.5	58.8	47.8
Surgical-VQA (Seenivasan et al., 2022)	89.8	63.2	65.6
LV-GPT (Seenivasan et al., 2023)	87.5	68.1	59.3
Surgical-LLaVA (Jin and Jeong, 2024)	92.2	68.7	67.1
OmniV-Med-Tiny-1.5B	91.7	78.8	64.8
OmniV-Med-7B	92.9	80.3	69.3

Table 7: Performance Comparison of Methods on Different Video Benchmarks.

Embedding Type	Model	VQA-RAD		SLAKE		PathVQA	
		Recall	Accuracy	Recall	Accuracy	Recall	Accuracy
Fixed Positional Embedding	OmniV-Med-Tiny-1.5B	55.50	82.93	80.23	83.12	33.26	90.19
Rotary Position Embedding	OmniV-Med-Tiny-1.5B	59.22 (+3.72)	84.93 (+2.00)	84.59 (+4.36)	87.98 (+4.86)	38.87 (+5.61)	92.66 (+2.47)

Table 8: Ablation of fixed positional embedding and rotary position embedding.

of temporal support). As shown in Figure 7, the modality-specific 3D encoder underperforms our unified encoder on both 2D Med-VQA and 3D VQA. We attribute this to two factors: (i) medical 3D encoders are pretrained on much smaller volumetric corpora than CLIP-style ViTs, limiting their representation capacity; and (ii) strong 2D understanding is foundational for 3D reasoning, as volumes consist of correlated 2D slices whose evidence must be reliably aggregated.

Cost Analysis. Training the OmniV-Med-1.5B model requires less than 24GB of GPU memory, due to our high-quality data, unified visual encoder, and RGTP method. This enables training on eight RTX 3090 GPUs, while the 7B version supports inference on a single RTX 3090 GPU. Such efficiency enhances accessibility and underscores the practical value of our models for real-world clinical applications.

Empirical Transfer from 2D Training. Stages 1–2 contain only 2D data, permitting a clean test of whether the resulting patch representations transfer to video before mixed-modal tuning. On the complete official test sets of three video-VQA benchmarks, this 2D-only OmniV-Med-7B reaches 65.5% average zero-shot accuracy, compared with 50.2% for LLaVA-Med-LLaMA3.2-8B and 55.6% for HuatuoGPT-Vision-7B. We treat these aggregate results as empirical transfer evidence rather than an established emergent mechanism.

Effect of Data Scaling. By incorporating diverse modalities in stage 3 on a unified visual encoder, we observe that introducing 3D images and videos alongside 2D images does not degrade performance

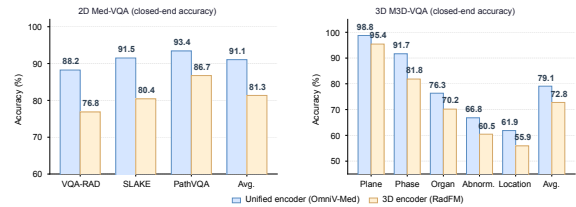


Figure 7: Comparison of our unified rotary position-adaptive encoder with the modality-specific RadFM 3D encoder. Left: closed-end accuracy on 2D Med-VQA; right: closed-end accuracy across M3D-VQA sub-tasks.

on 2D benchmarks. Instead, it enhances performance, as shown in Figure 12. This indicates that our training paradigm enables Med-VLMs to follow the data scaling law (Kaplan et al., 2020) in multimodal training scenarios, where increasing training data consistently improves performance.

Progressive Training Yields the Best Performance. We investigate the effect of jointly training stage 2 (2D images) and stage 3 (mixed-modal data) compared to our progressive training strategy. As shown in Figure 13, the best performance is achieved when 2D image understanding is robustly established before extending to 3D images and videos. This underscores the importance of 2D image understanding as a foundation and validates the effectiveness of our progressive strategy. Additional schedules (mixed-from-scratch, joint SFT, and 3D/video-first) are reported in Appendix A.5.

Effect of Parameter Scale. The 1.5B-to-7B improvement is modest on several in-domain tests but substantially larger on held-out tasks: +3.8 on OmniMedVQA, +13.8 on MMMU Health & Medicine, and +6.72 Rouge-L on MedLVBench.

This pattern suggests that scale chiefly benefits generalization in our evaluated size range.

4.4 Clinical Reliability

HealthBench (Arora et al., 2025) uses physician-written rubrics for realistic clinical conversations and penalizes unsupported claims across its scoring axes. Table 9 reports the overall score and the axes most directly connected to safety and calibrated uncertainty. OmniV-Med-7B improves on the evaluated medical VLMs, but its absolute overall score of 0.3758 remains far from sufficient for autonomous clinical use. These results complement, rather than replace, prospective expert evaluation.

5 Conclusion

We presented OmniV-Med, a practical recipe for training unified Med-VLMs across 2D images, 3D volumes, and medical videos. Controlled studies identify the benefits of a 2D-first curriculum, a shared encoder, and redundancy-guided token pruning; held-out and reliability evaluations complement the in-domain benchmark results. The 1.5B and 7B instantiations provide an efficient foundation for future unified medical VLM research, while substantial clinical validation remains necessary before deployment.

Limitations

Our largest model has 7B parameters; the behavior of the recipe at 32B or 72B remains untested. Although the training data span many modalities and tasks, public medical datasets do not cover the full diversity of institutions, scanners, demographics, rare diseases, or clinical workflows. Distribution shift and systematic bias may therefore remain even after our decontamination procedure.

Benchmark accuracy and physician-rubric grading cannot establish clinical safety. The model can still produce unsupported findings, omit subtle or rare abnormalities, express inappropriate confidence, or respond inconsistently across patient groups. HealthBench provides a useful reliability signal but is text-centric and automatically graded; it does not replace prospective, multi-institutional evaluation by clinicians. OmniV-Med is a research model and is not suitable for autonomous diagnosis, treatment decisions, or other unsupervised clinical use. Deployment would require calibrated uncertainty, human oversight, privacy safeguards,

Model	Overall	Emergency	Hedging
HuatuoGPT-V-34B	0.1717	0.2611	0.2533
Lingshu-7B	0.1590	0.2576	0.2569
Lingshu-32B	0.1904	0.2493	0.2927
OmniV-Med-7B	0.3758	0.5231	0.4602

Table 9: HealthBench results under the official GPT-4.1 grading protocol. Higher is better.

subgroup audits, and task-specific regulatory validation.

Finally, RGTP is a simple adjacent-slice/frame heuristic. The lesion-focused test is encouraging but does not exclude failures on other small, transient, or low-contrast findings. Broader expert-reviewed evaluations are needed before applying pruning in safety-critical settings.

Acknowledgments

This work is supported by the National Key R&D Program of China (Grant No. 2024YFC3308304), the “Pioneer” and “Leading Goose” R&D Program of Zhejiang (Grant no. 2025C01128), and the National Natural Science Foundation of China (62476241).

References

- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kwon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, et al. 2024. Mimic-ext-mimic-cxr-vqa: A complex diverse and large-scale visual question answering dataset for chest x-ray images. *PhysioNet*.
- Fan Bai, Yuxin Du, Tiejun Huang, Max Q.-H. Meng, and Bo Zhao. 2024. **M3D: Advancing 3D medical image analysis with multi-modal large language models**. *Preprint*, arXiv:2404.00578.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. **Qwen2.5-VL technical report**. *Preprint*, arXiv:2502.13923.
- Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision-language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380.

- Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. [BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109.
- Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. [Token merging: Your vit but faster](#). In *The Eleventh International Conference on Learning Representations*.
- Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, and Benyou Wang. 2024. [Towards injecting medical visual knowledge into multimodal LLMs at scale](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7346–7370. Association for Computational Linguistics.
- Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. [Generating long sequences with sparse transformers](#). *Preprint*, arXiv:1904.10509.
- Adrito Das, Danyal Z Khan, Dimitrios Psychogios, Yitong Zhang, John G Hanrahan, Francisco Vasconcelos, You Pang, Zhen Chen, Jinlin Wu, Xiaoyang Zou, et al. 2025. Pitvis-2023 challenge: Workflow recognition in videos of endoscopic pituitary surgery. *Medical image analysis*, page 103716.
- Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. [PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain?](#) In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193. Association for Computational Linguistics.
- Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. [Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24108–24118.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.
- Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. [OmniMed-VQA: A new large-scale comprehensive evaluation benchmark for medical LVLM](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.
- Songtao Jiang, Sibong Song, Chenyi Zhou, Yuan Wang, Ruizhe Chen, Tongkun Guan, Ruilin Luo, Yan Zhang, Zhihang Tang, Yuchong Sun, Hang Zhang, Zhibo Yang, Shuai Bai, Junyang Lin, and Zuozhu Liu. 2026. Learning transferable temporal primitives for video reasoning via synthetic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31283–31292.
- Songtao Jiang, Yuan Wang, Sibong Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. 2025. Hulu-Med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.
- Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. [Med-MoE: Mixture of domain-specific experts for lightweight medical vision-language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860. Association for Computational Linguistics.
- Juseong Jin and Chang Wook Jeong. 2024. [Surgical-LLaVA: Toward surgical scenario understanding via large language and vision models](#). *Preprint*, arXiv:2410.09750. NeurIPS 2024 AIM-FM Workshop.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *Preprint*, arXiv:2001.08361.
- LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. 2025. [Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning](#). *Preprint*, arXiv:2506.07044.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251.
- Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. [LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 28541–28564.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. [VisualBERT: A simple and performant baseline for vision and language](#). *Preprint*, arXiv:1908.03557.
- Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. 2023b. Self-supervised vision-language pretraining for medical visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE.

- Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Song xiaohui, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. 2025. [HealthGPT: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation](#). In *Forty-second International Conference on Machine Learning*.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. [SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering](#). In *2021 IEEE 18th International Symposium on Biomedical Imaging*, pages 1650–1654.
- Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023. [Q2ATransformer: Improving medical VQA via an answer querying decoder](#). In *Information Processing in Medical Imaging*, pages 445–456. Springer Nature Switzerland.
- Zuozhu Liu, Songtao Jiang, Tianxiang Hu, Ziwei Xue, Yuhui Chen, Dengfeng Ruan, Haoxiang Xia, Yang Feng, Jin Hao, Wei Li, et al. 2026. [DestinyNet: A deep-learning framework for cell-fate analysis from lineage-tracing single-cell RNA sequencing data](#). *Patterns*, 7(4).
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. [Med-flamingo: a multimodal medical few-shot learner](#). In *Machine learning for health (ML4H)*, pages 353–367. PMLR.
- Lalithkumar Seenivasan, Mobarakol Islam, Gokul Kannan, and Hongliang Ren. 2023. [SurgicalGPT: End-to-end language-vision GPT for visual question answering in surgery](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 281–290. Springer Nature Switzerland.
- Lalithkumar Seenivasan, Mobarakol Islam, Adithya K. Krishna, and Hongliang Ren. 2022. [Surgical-VQA: Visual question answering in surgical scenes using transformer](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 33–43. Springer Nature Switzerland.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. [Towards generalist biomedical ai](#). *Nejm Ai*, 1(3):AIoa2300138.
- Natalia Valderrama, Paola Ruiz Puentes, Isabela Hernández, Nicolás Ayobi, Mathilde Verlyck, Jessica Santander, Juan Caicedo, Nicolás Fernández, and Pablo Arbeláez. 2022. [Towards holistic surgical scene understanding](#). In *International conference on medical image computing and computer-assisted intervention*, pages 442–452. Springer.
- Tom van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees G. M. Snoek, and Marcel Worring. 2023. [Open-ended medical visual question answering through prefix tuning of language models](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 726–736. Springer Nature Switzerland.
- Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. 2023. [Foundation model for endoscopy video analysis via large-scale self-supervised pre-train](#). In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 101–111. Springer Nature Switzerland.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. 2025. [Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data](#). *Nature Communications*, 16(1):7866.
- Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. [Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering](#). *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959.
- Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. [MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11941–11952.
- Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, and 4 others. 2024. [A generalist vision-language foundation model for diverse biomedical tasks](#). *Nature Medicine*, 30(11):3129–3141.
- Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh

Rao, Mu Wei, Naveen Valluri, et al. 2023a. Biomed-clip: a multimodal biomedical foundation model pre-trained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

A Appendix

A.1 Details of Hyperparameters

Our training process is divided into three stages. In Stage 1, we set the learning rates for the large language model (LLM), multimodal projector, and vision encoder to $1e-5$, $1e-3$, and $1e-5$, respectively. The weight decay is set to 0, and the warmup ratio is 0.03. We use a cosine learning rate scheduler and train for 1 epoch. In Stage 2 and Stage 3, we adjust the learning rates to $1e-5$ for the LLM, $1e-5$ for the multimodal projector, and $2e-6$ for the vision encoder, while keeping the weight decay and warmup ratio unchanged. Stage 2 is trained for 1 epoch, and Stage 3 is trained for 3 epochs. This progressive training strategy ensures stable optimization and effective convergence across all components of the model.

A.2 Details of Rejection Sampling

To construct high-quality video captions in the medical domain, as shown in Figure 8, we employ rejection sampling due to the lack of open-source Med-VLMs for video processing. Specifically, we utilize the state-of-the-art general-purpose model Qwen2.5-VL-72B (Bai et al., 2025) to generate five candidate captions for each video, which are then evaluated using the LLMs-as-Judge approach. First, we prompt Qwen2.5-VL to score the candidates based on two criteria: (1) relevance to the video content, which assesses how well the caption captures the key actions, dynamics, and contextual information, and (2) fluency of the caption, which evaluates grammatical correctness, naturalness, and avoidance of static image-like descriptions, with each criterion scored on a 5-point scale. Next, we use HuatuoGPT2-13B (Chen et al., 2024), a medical-specific LLM, to assess the medical accuracy of the captions, also on a 5-point scale, focusing on the precision of medical terminology, concepts, and logical relationships. The final score for each candidate is computed as the sum of these three scores, and the caption with the highest total score is selected as the final output, ensuring both contextual relevance and medical reliability.

A.3 Evaluation Protocol for Proprietary Commercial VLMs

To enable a reproducible comparison with leading proprietary commercial VLMs (Gemini-2.5-Pro, GPT-5-Thinking-High, GPT-4o, Claude-Sonnet-4.5, and Qwen-VL-Max) on MedLVBench and the

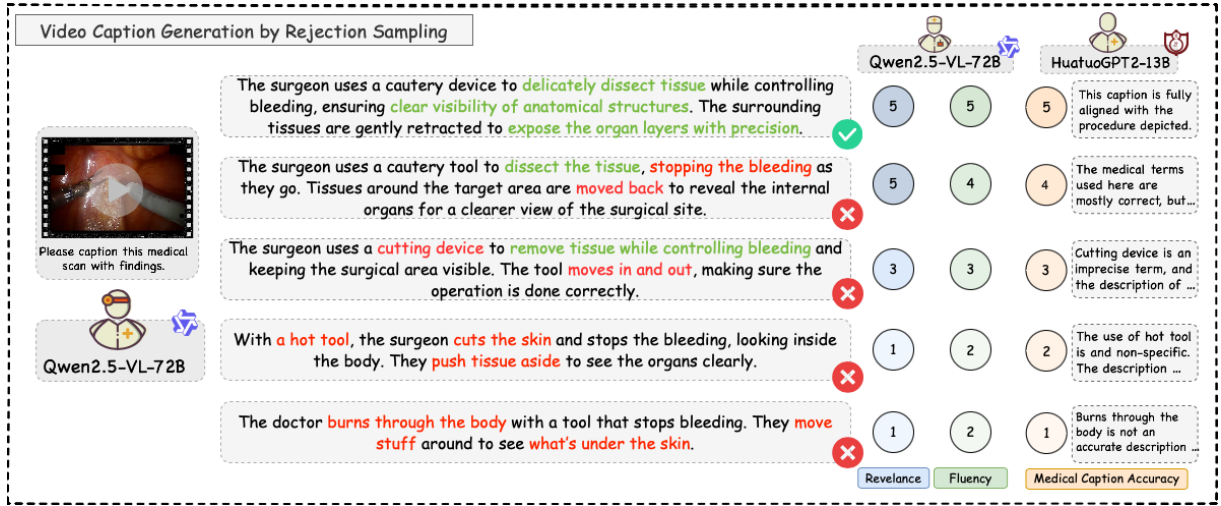


Figure 8: A framework for high-quality medical video captioning via rejection sampling: Leveraging Qwen2.5-VL-72B for candidate generation and Qwen2.5-VL-72B, HuatuoGPT2-13B for medical caption assessment based on relevance, fluency, and accuracy (5-point scale).

four 2D Med-VQA benchmarks, we standardize the evaluation protocol following Lingshu (LASA Team et al., 2025).

Video sampling. Among the four commercial base-lines, only Gemini-2.5-Pro natively supports inputs of up to one hour at full 1 fps. For the remaining models (GPT-5-Thinking-High, GPT-4o, Claude-Sonnet-4.5, Qwen-VL-Max) we sample frames at 1 fps with a per-call cap of `max_frames = 512`; on the 40–60 min duration bin where the raw 1 fps stream exceeds this cap, we uniformly sub-sample to stay under the limit while preserving temporal order. Gemini-2.5-Pro is queried with full 1 fps without sub-sampling. OmniV-Med uses its own long-video pipeline with the RGTP described in Section 3.

Generation hyperparameters. All commercial models are queried with `temperature = 1.0`, default `top-p` per API, no system prompt, and the same user prompt that asks the model to produce a free-form caption (for MedLVBench) or to answer the VQA question (for the 2D benchmarks).

Judge prompts. We use GPT-4.1 as the LLM-as-Judge for both task families, following Lingshu (LASA Team et al., 2025). For the 2D Med-VQA benchmarks (VQA-RAD, SLAKE, PathVQA, PMC-VQA, M3D-VQA) we use a binary-judge prompt that compares each prediction against the standard answer, accepting paraphrases with equivalent meaning. For MedLVBench captioning, the judge instead rates each generated caption against the canonical reference on a 5-point scale, jointly considering coverage of surgi-

cal steps/instruments, factual accuracy of medical terms, and narrative fluency. The exact prompts are reproduced below.

```
# VQA correctness (binary)
def vqa_judge_prompt(q, response,
                    answer):
    return f"""
Determine whether the user's answer is
correct given the question and standard
answer. A paraphrased answer with the
same meaning is also considered correct.
```

```
Question: {q}
Standard answer: {answer}
User answer: {response}
```

```
Output strictly in the format
(0 = incorrect, 1 = correct):
<think>your concise reasoning</think>
<judge>0/1</judge>
"""
# Caption quality (5-point)
def caption_judge_prompt(response, ref):
    return f"""
Rate the model's caption against the
reference on a 5-point scale (1=poor,
5=excellent), jointly considering:
(i) coverage of surgical steps and
instruments,
(ii) factual accuracy of medical terms,
(iii) narrative fluency and coherence.
```

```
Reference: {ref}
Model caption: {response}
```

```
Output strictly in the format:
<think>your concise reasoning</think>
<score>1-5</score>
"""
```

A.4 MedLVBench Construction

MedLVBench is built from the publicly released PitVis-2023 Challenge videos, which provide

dense frame-level step and instrument annotations of the endoscopic TransSphenoidal Approach (eTSA). For each evaluation clip, we first deterministically synthesize a paragraph-level draft caption from the underlying step/instrument annotations using a fixed rule-based template that enumerates steps in temporal order and lists the instruments active in each step. Two trained clinical annotators from a collaborating hospital in China then manually refine the draft for fluency, surgical realism, and medical accuracy; each clip is thus paired with a single canonical reference caption used by both Rouge-L and GPT-4.1-Judge. The PitVis-2023 videos used to construct MedLVBench are entirely held out from OmniV-Med-Instruct: no PitVis frames, derived QA pairs, or captions appear in the training corpus. To further rule out near-duplication with other surgical-video sources (Cholec80, PSI-AVA, EndoVis18), we drop from training any clip sharing source patient ID, procedure ID, or perceptual-hash similarity ≥ 0.8 with a PitVis frame. MedLVBench contains 158 clips spanning roughly 72 hours of surgical video, split into three duration bins: 10–20 min (61 clips), 20–40 min (58 clips), and 40–60 min (39 clips). From these clips, we derive a balanced evaluation set of 600 VQA samples (200 per duration bin) covering step identification, instrument recognition, and procedure-level reasoning, with each sample reviewed by the same clinical annotators. Because the reference captions are derived from and refined against step/instrument annotations, the recall component of Rouge-L reflects whether the model identifies the correct sequence of steps and instruments while allowing paraphrasing in textual realization. We therefore complement Rouge-L with GPT-4.1-Judge as a more lexical-invariant semantic metric and report both to avoid over-reliance on either signal alone.

A.5 Additional Curriculum Ablations

To complement Figure 13, we evaluate four alternative schedules under the same data and architecture as OmniV-Med-7B: (i) *2D only*: 3D and video data omitted; (ii) *Mixed from scratch*: stage 1 alignment skipped, all modalities trained jointly from the start; (iii) *Joint SFT*: stage 1 retained, but stages 2 and 3 are collapsed into a single joint SFT on 2D+3D+video; (iv) *3D/video first $\rightarrow$ 2D*: reversed order, training 3D and video before 2D. Table 10 shows that all four variants underperform progressive training. *Mixed-from-scratch* drops by up to

9.2 points on VQA-RAD and 7.2 on PathVQA, indicating that fitting heterogeneous modalities without a dedicated alignment stage destabilizes optimization. *Joint SFT* preserves stage 1 alignment and recovers part of this gap, but still trails progressive by 3–5 points, showing that staging the SFT itself (2D first, then mixed-modal) carries additional benefit beyond alignment alone. The clearest evidence for the modality-ordering effect comes from *3D/video-first*, which drops by 5–7 points relative to progressive across the three benchmarks: when longer-sequence modalities are introduced before 2D alignment stabilizes, patch representations drift toward longer-sequence statistics and are not fully recovered by subsequent 2D fine-tuning. Finally, *2D-only* loses only 1–2 points on each benchmark, so the 3D and video data act as a positive auxiliary signal for 2D tasks rather than as a distractor.

Schedule	VQA-RAD	SLAKE	PathVQA
2D only	65.8	84.6	39.6
Mixed from scratch	58.7	81.3	34.5
Joint SFT (stage 2+3)	63.2	82.4	38.3
3D/video first $\rightarrow$ 2D	61.5	79.1	36.2
Progressive (ours)	67.9	85.9	41.7

Table 10: Curriculum ablation under fixed data and architecture (OmniV-Med-7B). Evaluation metric for Open-ended is Recall.

A.6 Token-Reduction Strategy Comparison

We compare RGTP against two alternatives under a matched 60% token-keep budget on OmniV-Med-7B: *ToMe* (Bolya et al., 2023), which merges visually similar tokens by key-similarity, and *sparse attention* (Child et al., 2019), which keeps all tokens but restricts each token’s receptive field via a sparsified attention mask.

RGTP vs. ToMe. ToMe merges patches by visual similarity, but in medical 2D images two patches can be visually similar yet play different diagnostic roles (e.g., adjacent healthy tissue next to a lesion); merging them blurs this local context. RGTP instead removes patches whose neighbors across consecutive slices or frames are highly redundant in L_1 distance, so spatially adjacent but functionally distinct patches are preserved. This explains the 1.5–2.6 point gap on the three 2D benchmarks (Table 11); the gap is consistent in sign across benchmarks, which suggests it reflects a systematic mismatch between visual-similarity merging and lesion-level discrimination rather than a single dataset’s artifact.

RGTP vs. sparse attention. Sparse attention does not remove tokens but limits how widely each one can attend; on 2D Med-VQA this hurts global aggregation of small, distributed evidence such as scattered lesions, and the drop is larger (up to -4.2 on VQA-RAD). On the 40–60 min MedLVBench bin, sparse attention is marginally better than RGTP by 0.4 Rouge-L, because long videos contain heavy temporal redundancy that even a coarse sparsification pattern can absorb. Since preserving 2D fidelity is more critical for clinical decisions than a 0.4-point gain at the longest video bin, we adopt RGTP as the default.

Strategy	VQA-RAD	SLAKE	PathVQA
ToMe (token merge)	65.3	84.1	40.2
Sparse attention	63.7	82.6	38.8
RGTP (ours)	67.9	85.9	41.7

Table 11: Open-ended Recall under a matched 60% token-keep budget.

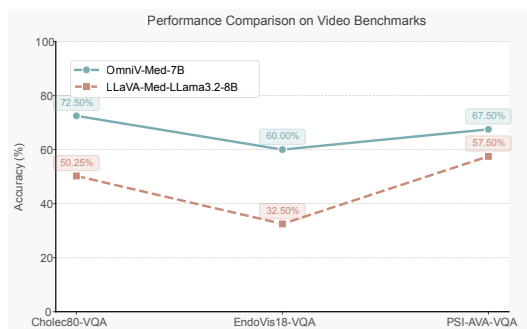


Figure 9: Task transferring capability of OmniV-Med

Method	VQA-RAD			SLAKE			PathVQA		
	EMS	Recall	BLEU	EMS	Recall	BLEU	EMS	Recall	BLEU
LLaVA-Med-7B	58.33	61.52	54.13	82.83	83.08	81.69	37.95	36.86	32.89
Med-Flamingo (Few-Shot)-9B (Moor et al., 2023)	20.00	–	–	–	–	–	31.00	–	–
PaLM-E-84B (Tu et al., 2024)	–	–	59.19	–	–	52.65	–	–	54.92
Med-MoE (Phi2)-3.6B (Jiang et al., 2024)	59.69	58.55	52.95	84.46	85.06	83.16	34.37	34.74	32.85
Med-MoE (StableLM)-2.0B (Jiang et al., 2024)	52.53	50.08	45.67	82.44	83.16	81.53	33.60	33.79	32.67
OmniV-Med-Tiny-1.5B	60.30	59.22	54.16	84.76	84.59	82.96	40.35	38.87	37.96
OmniV-Med-7B	70.72	67.87	64.62	85.99	85.92	84.63	43.06	41.68	40.23

Table 12: Detailed comparison with additional metrics in open settings. EMS denotes *Exact Match Score*; “–” marks numbers not reported by the original work.

Threshold τ	Cholec80-VQA	EndoVis18-VQA	PSI-AVA-VQA
0 (no pruning)	91.43	77.05	63.53
0.1	91.69 (+0.26)	78.81 (+1.76)	64.80 (+1.27)
0.3	90.54	74.64	61.82

Table 13: Accuracy of OmniV-Med-Tiny-1.5B on the video benchmarks under varying token-reduction thresholds τ . Gains over the no-pruning baseline are shown in parentheses.

Setting	Abnorm.	Location	Avg.
W/o pruning	71.42	70.28	70.85
+Pruning ($\tau=0.1$)	70.45	69.39	69.92
Δ	−0.97	−0.89	−0.93

Table 14: Closed-end accuracy on the small-lesion subset of M3D-VQA (< 5% volumetric coverage, 1,832 questions). Pruning cuts 60% of visual tokens with negligible loss on lesion-centric questions.

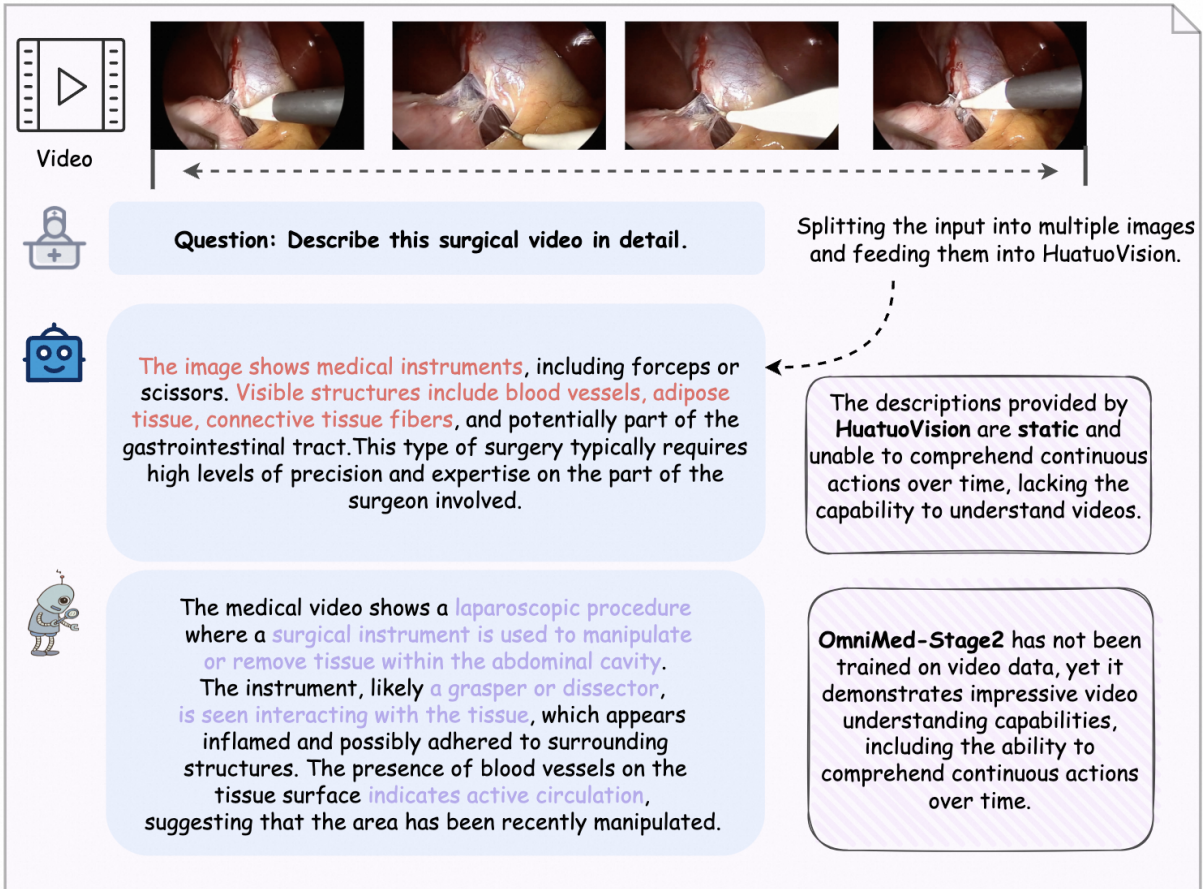


Figure 10: OmniV-Med-Stage2 demonstrates outstanding task transfer capability, extending from 2D image understanding to video comprehension, showcasing a strong grasp of temporal action dynamics.

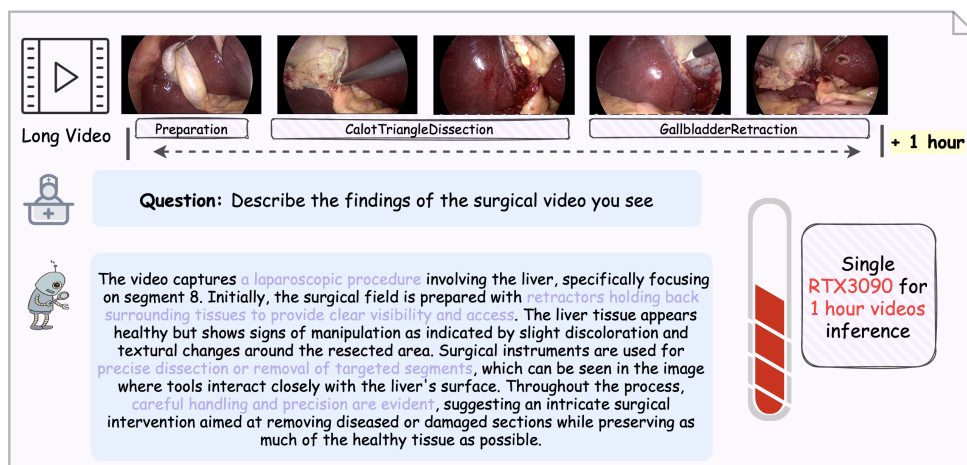


Figure 11: OmniV-Med demonstrates robust performance on long-video inference with one single RTX 3090.

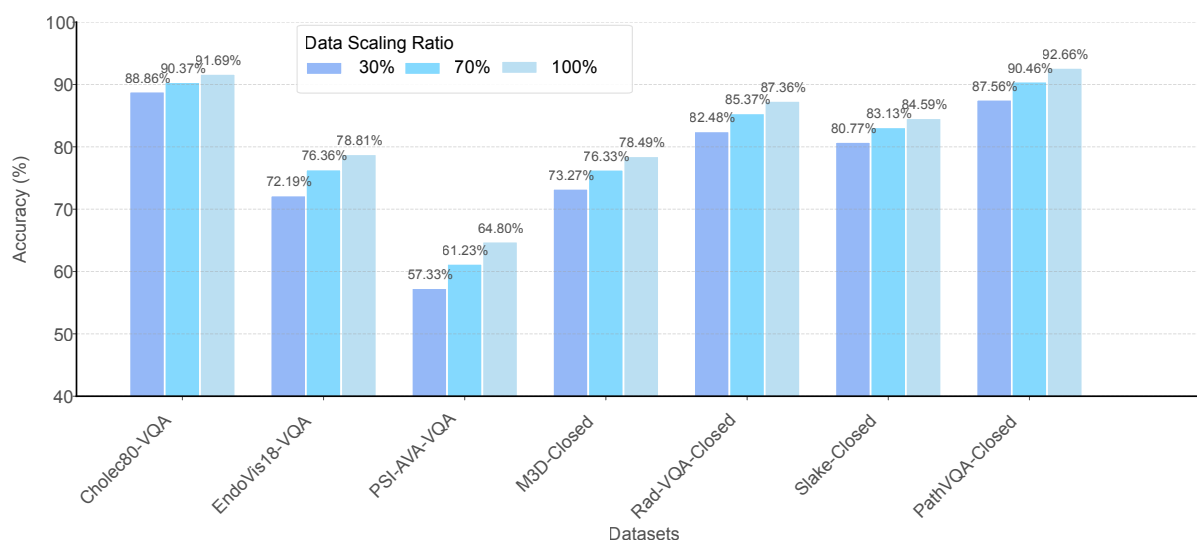


Figure 12: Performance under different data scaling ratio.

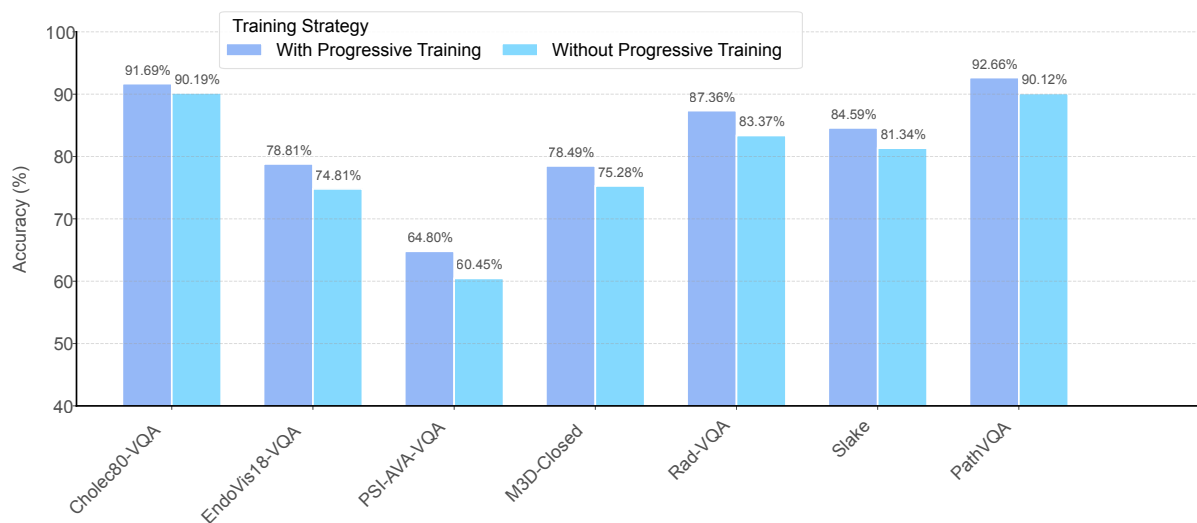


Figure 13: Performance comparison with and without progressive training strategy.