Photon: Speedup Volume Understanding with Efficient Multimodal Large Language Models

Chengyu Fang*, Heng Guo*, Zheng Jiang, Chunming He, Xiu Li†, Minfeng Xu†

Tsinghua University · DAMO Academy, Alibaba Group · Hupan Lab · Duke University

PROBLEM SETTING

Medical 3D visual question answering needs full-volume reasoning, but naive volumetric modeling is too expensive for modern MLLMs.

- Slice-based pipelines break volumetric continuity and inject frame-selection bias.
- Fixed-length or fixed-ratio compression removes subtle pathology.
- The token budget should adapt to instruction relevance instead of staying fixed across all cases.

Overview

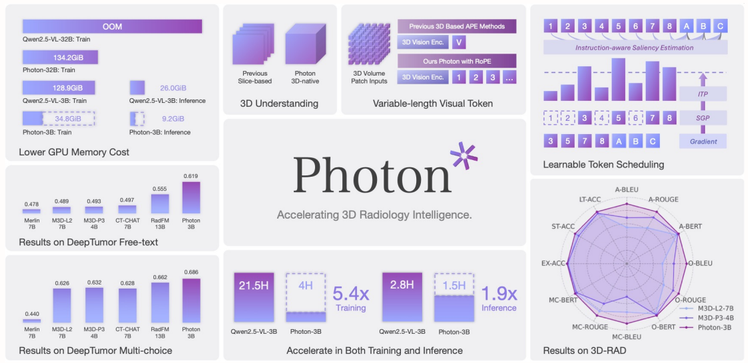

Figure 1. Photon is a 3D-native MLLM framework with variable-length visual tokens scheduling and improves both efficiency and performance.

Contributions

- Direct 3D volume modeling without slice sampling or fixed token compression.
- Instruction-conditioned Token Scheduling (ITS) predicts sample-specific retention.
- Surrogate Gradient Propagation (SGP) makes discrete token dropping trainable.
- Training and inference are both accelerated while preserving clinical grounding.

METHOD

Photon first aligns 3D patch embedding layer with Qwen2.5-VL using volume-caption pairs, then fine-tunes the full model with adaptive token scheduling on specific QA tasks. The same learned pruning logic is used during both training and inference.

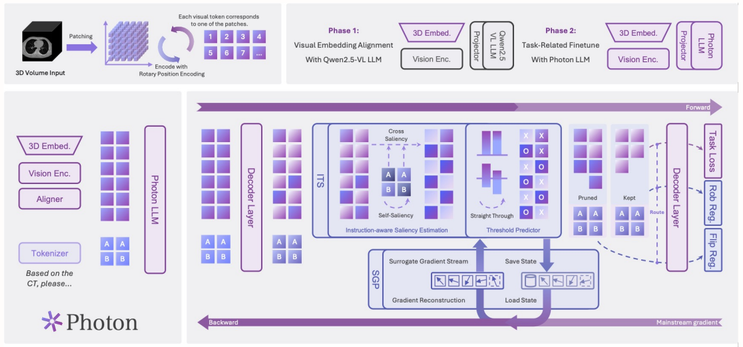

Figure 2. Phase 1 aligns visual embeddings; Phase 2 learns token reduction thresholds end to end.

In Phase 2, ITS and SGP jointly perform learnable visual token pruning.

How ITS and SGP Work

Predicted by network

ITS Saliency

$$u_j = \sum_t w_t \frac{\langle q_t, k_j \rangle}{\sqrt{\alpha}}$$

$$\rho_j = \mathrm{norm}(u_j)$$

Instruction-aware token importance.

Retention Rule

$$q_j = \sigma\left(\frac{\rho_j - \theta}{\tau_{ce}}\right)$$

$$M_j = \mathbf{1}\{q_j > 0.5\}$$

Probability first, then hard pruning.

Surrogate Gradient

$$\beta\, d_j\, s_j \max\{q_j(1-q_j),\, \epsilon_{sat}\}$$

SGP reinforces useful tokens while keeping gradients from saturating.

Read

Use the instructions to determine which instruction tokens are central.

Score

Calculate which visual tokens are relevant to those instruction tokens.

Prune

Predict a case-specific threshold and drop low-value visual tokens.

Learn

Send surrogate gradients back so the pruning rule improves during training.

ITS decides what to keep for this question; SGP makes that hard selection trainable instead of inference-only.

Qualitative Results

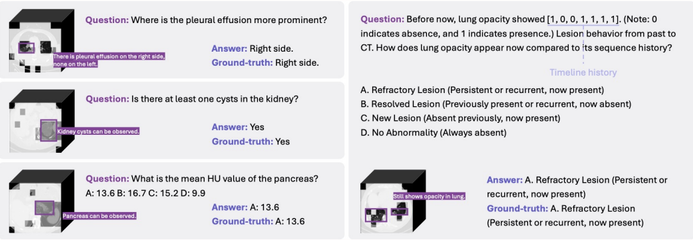

Figure 3. Question-dependent retained regions align with clinical targets such as pleural effusion, kidney cysts, pancreas, and lung opacity.

RESULTS

We evaluate our method on 3D-RAD and DeeptumorVQA (in our paper).

Task	Metric	Zero-shot						Finetuned		
		Qwen2.5-VL 3B	RadFM 13B	M3D-L2 7B	M3D-P3 4B	OmniV 1.5B	Lingshu 7B	M3D-L2 7B	M3D-P3 4B	Photon 3B
Existence Detection	Accuracy	19.52	29.20	18.00	40.25	28.66	59.60	81.09	82.43	83.07
Stat. Temp. Diag.	Accuracy	0.00	44.11	25.47	25.40	22.96	6.02	51.20	49.30	52.86
Longit. Temp. Diag.	Accuracy	0.29	42.99	24.17	24.31	24.23	12.13	74.78	74.77	77.01
Medical Measurement	BLEU	1.78	3.34	15.95	2.55	2.52	2.50	30.54	33.52	37.74
	ROUGE	4.30	6.62	23.24	5.63	7.88	5.45	36.06	36.46	39.36
	BERT Score	84.20	86.85	91.50	85.74	85.66	83.81	94.65	94.86	96.14
Image Observation	BLEU	3.51	13.48	10.69	16.31	16.42	5.34	31.28	39.66	51.59
	ROUGE	8.84	19.14	20.82	23.19	26.69	12.25	39.12	50.52	56.66
	BERT Score	84.53	87.16	86.61	86.92	88.29	85.67	90.00	92.19	93.62
Anomaly Detection	BLEU	2.93	11.00	9.10	15.06	13.47	3.71	25.25	33.28	42.33
	ROUGE	9.17	17.62	18.64	23.19	25.72	9.46	33.76	42.45	47.50
	BERT Score	84.47	86.76	86.07	87.11	88.21	84.81	89.16	90.72	91.96

Table 1. Comparison with other MLLMs on 3D-RAD

Methods	E.D.	S.T.D.	L.T.D.	Medical Measurement			Image Observation			Anomaly Detection			Infer Speed	Token Num
	Acc.	Acc.	Acc.	BLEU	ROUGE	BERT	BLEU	ROUGE	BERT	BLEU	ROUGE	BERT	(Tok/s)	(Tok/case)
Qwen2.5-VL	81.97	47.62	75.36	37.00	38.57	96.04	52.72	56.48	93.63	42.01	47.36	91.93	2.30	7.0K
VisionZip	82.00	47.19	75.42	37.04	38.69	96.06	52.51	56.36	93.63	41.96	47.39	91.93	2.32	2.1K
VisionZip	81.99	47.74	75.93	37.11	38.68	96.07	52.72	56.60	93.65	41.98	47.36	91.92	2.23	3.5K
VisionZip	82.01	47.40	75.88	37.07	38.82	96.04	52.57	56.43	93.64	42.01	47.39	91.93	2.18	4.9K
HiPrune	81.99	48.08	75.50	37.00	38.59	96.05	52.56	56.46	93.64	42.08	47.32	91.93	0.76	2.1K
HiPrune	81.97	47.84	75.58	37.14	38.75	96.07	52.65	56.60	93.63	41.92	47.31	91.91	0.75	3.5K
HiPrune	82.02	47.19	75.35	37.00	38.62	96.06	52.57	56.44	93.62	41.94	47.40	91.91	0.73	4.9K
Photon	83.07	52.86	77.01	37.74	39.36	96.14	51.59	56.66	93.62	42.33	47.50	91.96	4.12	Dynamic
Photon MAX	85.50	57.79	79.70	39.44	40.96	96.17	52.83	58.72	93.70	42.85	48.89	92.00	2.60	Dynamic

Table 2. Compare with other token pruning methods on 3D-RAD

Task	E.D.	S.T.D.	L.T.D.	Medical Measurement			Image Observation			Anomaly Detection			Computational Metrics			
	Acc.	Acc.	Acc.	BLEU	ROUGE	BERT	BLEU	ROUGE	BERT	BLEU	ROUGE	BERT	Train spd.	Train Tok.	Infer Spd.	Infer Mem.
Without task-related finetune.																
Qwen2.5-VL	19.52	0.00	0.29	1.78	4.30	84.20	3.51	8.84	84.53	2.93	9.17	84.47	—	—	—	—
Vis. Ful. Ft.	0.00	0.00	0.00	5.02	9.91	84.82	1.02	2.96	81.46	0.83	2.61	80.83	—	—	—	—
Photon Phase 1	28.10	8.41	10.90	4.81	9.69	85.35	5.98	17.62	86.39	6.84	15.64	85.98	—	—	—	—
Finetuned with 3D-RAD datasets.																
Qwen2.5-VL Ft.	81.97	47.62	75.36	37.00	38.57	96.04	52.72	56.48	93.63	42.01	47.36	91.93	0.15	7.00K	2.30	26.0GiB
Photon	83.07	52.86	77.01	37.74	39.36	96.14	51.59	56.66	93.62	42.33	47.50	91.96	0.85	0.39K	4.12	9.2GiB
Photon Max	85.50	57.79	79.70	39.44	40.96	96.17	52.83	58.72	93.70	42.85	48.89	92.00	0.19	3.08K	2.60	21.4GiB

Table 3. Efficient Comparison. (Train spd.=sample/s, Infer Spd.=tok/s, Average Performance).

Homepage LinkedIn WeChat

Email: chengyufang.thu@gmail.com
The first author is currently seeking a PhD position for Fall 2027. Any opportunities would be appreciated!