SV-RAG: LoRA-Contextualizing Adaptation of MLLMs for Long Document Understanding

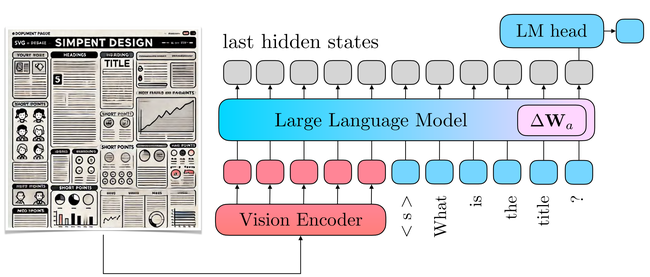
Jian Chen, Ruiyi Zhang, Yufan Zhou, Tong Yu, Franck Dernoncourt,
Jiuxiang Gu, Ryan A. Rossi, Changyou Chen, Tong Sun
University at Buffalo, Adobe Research

Introduction

We present SV-RAG, a two-step framework that empowers multimodal LLMs to understand long, visually-rich documents. Existing approaches suffer from high memory costs due to the long token sequences required by self-attention, as well as latency from OCR pipelines and the limitations of text-only understanding. SV-RAG addresses these by customizing the MLLM as both a retriever and reader—first selecting relevant pages, then answering questions—achieving state-of-the-art results on public benchmarks.

Preliminary (MLLM)

An MLLM consists of a vision encoder and a language model. The vision encoder divides the image into patches and encodes them as visual tokens, which are prepended to the text prompt. The combined sequence is processed by the LLM transformer, producing hidden states at each layer. The last hidden state is used for next-word prediction.

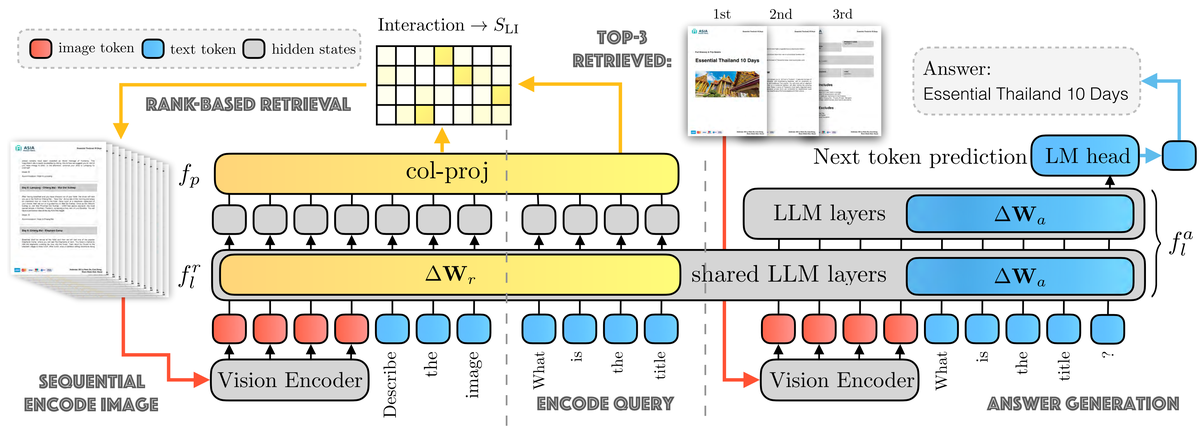


Methods

SV-RAG consists of two stages: retrieval and answer generation.

Retrieval: We finetune the MLLM as an encoder using contrastive learning. Given an image and a query string, we separately input them into the MLLM and extract hidden states as embedding sequences. Sequence-to-sequence similarity is computed as the matching score, and the model is trained to maximize similarity between relevant pages and queries.

Answer Generation: We apply LoRA-based finetuning on the MLLM, using the retrieved evidence page and query as input, and the answer as output.



Contextualized late interaction

Unlike single-vector encoders like CLIP, we uses a sequence interaction score to capture fine-grained similarity between text and image feature seqs. Given **Eq** and **Ev**, the score is defined as:

$$s_{\mathrm{LI}}(\mathbf{E}_q, \mathbf{E}_v) = \sum_{i=1}^{n} \max_{j \in \{1,..,m\}} \mathbf{e}_{q_i} \cdot \mathbf{e}_{v_j}^T$$

Contrastive Loss

The model is trained to maximize the score between a question and its evidence image (positive pair), while minimizing the score for the most similar yet unrelated image in the batch (hard negative). Since there is only one negative pair per sample, the InfoCE loss can be simplified to the Softplus form.

$$\mathcal{L} = \log(1 + \exp(s_{\mathrm{LI}}(\mathbf{E}_q, \mathbf{E}_v^-) - s_{\mathrm{LI}}(\mathbf{E}_q, \mathbf{E}_v^+)))$$

accuracy	SlideVQA top1	SlideVQA top5	MMLong top1	MMLong top5	VisR-B top1	VisR-B top5	SP-DocVQA top1	SP-DocVQA top5
Text-based Methods								
BM25	69.3	91.1	25.3	47.6	32.2	57.5	30.9	61.7
SBERT	73.0	91.0	44.7	70.2	38.8	72.1	47.4	74.0
BGE-M3	74.3	92.0	42.7	66.6	47.7	78.1	47.8	77.5
Bge-large	81.3	93.3	47.4	71.5	53.7	80.3	56.7	81.5
NV-Embed-v2	82.2	94.3	47.4	69.0	55.2	82.7	51.7	80.2
Encoder Models								
CLIP	58.4	86.9	32.4	63.4	33.4	62.1	37.1	69.4
SigLip	66.2	90.1	44.9	69.4	53.2	81.3	39.3	71.9
Col-Retrieval Modules								
Col-Paligemma	89.0	98.7	60.7	82.0	67.9	90.8	62.3	85.9
Col-InternVL2	88.5	98.3	61.3	83.0	69.3	90.7	63.2	85.9
Col-Phi-3-vision	90.6	98.8	64.8	84.8	71.9	91.8	65.1	87.0

Retrieval Results

We evaluated Col-retrieval on SlideVQA, MMLongBench-Doc, SPDocVQA, and VisR-Bench, comparing it with OCR-based text-only baselines and multimodal encoders. Retrieval accuracy results on four datasets indicate that Col-retrieval outperforms all baselines. achieving more than 98% in top-5 retrieval accuracy on the SlideVQA dataset.

QA Results

SV-RAG consistently outperforms InternVL2-8B (uses all pages), primarily because long documents overload LMMs with excessive visual tokens, causing high memory usage and distracted attention.

Method	#Param	Evidence	SP-SlideVQA EM	SP-SlideVQA PNLS	MMLongBench G-Acc	MMLongBench PNLS	SP-DocVQA ANLS	SP-DocVQA PNLS
Single-Page Evidence								
Cheating Baselines								
PaliGemma	3B	T	37.30	0.63	23.9	0.38	0.65	0.79
Phi-3-v	4B	T	13.72	0.80	33.7	0.52	0.65	0.85
InternVL2	4B	T	15.03	0.58	40.4	0.55	0.84	0.88
GPT-4o	-	T	30.59	0.84	56.8	0.62	0.87	0.94
Multi-image MLLMs								
InternVL2	8B	A	12.62	0.65	14.1	0.22	0.50	0.55
GPT-4o	-	A	27.28	0.81	54.5	0.57	0.69	0.80
SV-RAG Models (Proposed)								
SV-RAG-PaliGemma	3B	R1	35.03	0.60	23.9	0.35	0.56	0.69
SV-RAG-PaliGemma†	3B	R1	49.75	0.65	23.1	0.38	0.56	0.68
SV-RAG-Phi-3-vision	4B	R1	12.85	**0.78**	30.7	**0.50**	0.55	0.75
SV-RAG-Phi-3-vision†	4B	R1	**58.13**	0.77	28.4	0.44	0.68	0.73
SV-RAG-InternVL2	4B	R5	16.40	0.58	33.2	0.48	0.70	**0.76**
SV-RAG-InternVL2†	4B	R5	45.07	0.77	**34.0**	0.49	**0.71**	0.75

Our model, with only 4 billion parameters, outperforms all open-source LMMs and achieves performance comparable to private models, Claude-3 Opus and Gemini-1.5-Pro.

Method	#Param	Evidence Source TXT	LAY	CHA	TAB	FIG	Evidence Page SIN	MUL	UNA	ACC	F1
Open-source Models											
DeepSeek-VL-Chat	7.3B	7.2	6.5	1.6	5.2	7.6	5.2	7.0	12.8	7.4	5.4
Idefics2	8B	9.0	10.6	4.8	4.1	8.7	7.7	7.2	5.0	7.0	6.8
MiniCPM-Llama3-V2.5	8B	11.9	10.8	5.1	5.9	12.2	9.5	9.5	4.5	8.5	8.6
InternLM-XC2-4KHD	8B	9.9	14.3	7.7	6.3	13.0	12.6	7.6	9.6	10.3	9.8
mPLUG-DocOwl 1.5	8.1B	8.2	8.4	2.0	3.4	9.9	7.4	6.4	6.2	6.9	6.3
Qwen-VL-Chat	9.6B	5.5	9.0	5.4	2.2	6.9	5.2	7.1	6.2	6.1	5.4
Monkey-Chat	9.8B	6.8	7.2	3.6	6.7	9.4	6.6	6.2	6.2	6.2	5.6
CogVLM2-LLaMA3-Chat	19B	3.7	2.7	6.0	3.2	6.9	3.9	5.3	3.7	4.4	4.0
InternVL-Chat-v1.5	26B	14.0	16.2	7.1	10.1	16.6	14.9	12.2	**17.5**	14.6	13.0
EMU2-Chat	37B	6.1	9.7	2.6	3.8	7.7	5.7	6.1	16.5	8.3	5.5
SV-RAG Models (Proposed)											
SV-RAG-InternVL2 (R5)	4B	**26.5**	18.8	22.3	19.6	23.6	33.2	**13.1**	12.4	22.2	22.8
SV-RAG-InternVL2† (R5)	4B	26.3	**22.1**	**25.0**	**20.7**	**25.2**	**34.0**	10.6	15.7	**23.0**	**24.2**
Proprietary Models											
Claude-3 Opus	-	24.9	24.7	14.8	13.0	17.1	25.6	13.8	7.6	17.4	18.1
Gemini-1.5-Pro	-	21.0	17.6	6.9	14.5	15.2	21.1	11.1	69.2	28.2	20.6
GPT-4V	-	34.4	28.3	28.2	32.4	26.8	36.4	27.0	31.2	32.4	31.2
GPT-4o	-	46.3	46.0	45.3	50.0	44.1	54.5	41.5	20.2	42.8	44.9

Efficiency

SV-RAG maintains time and memory efficiency in both the retrieval and question answering phases.

SV-RAG-Backbone	Page	Retrieval Time	Retrieval Mem	QA Time	QA Mem
Paligemma	R1	2.3	9.2	1.0	12.4
Phi-3-vision	R1	4.1	11.6	0.9	12.9
InternVL2-4B	R1	9.2	14.2	1.4	14.6
InternVL2-4B	R5	9.2	14.2	2.8	40.8
InternVL2-4B	R12	9.2	14.2	4.1	76.4