SLQ: Bridging Modalities via Shared Latent Queries for Retrieval with Frozen MLLMs

Haoran Lou[1], Ziyan Liu[1], Chunxiao Fan[1,†], Yuexin Wu[1], Yue Ming[1], Hao Wu[2], Kai Zuo[2], Yibo Chen[2], Xu Tang[2]

[1] Beijing University of Posts and Telecommunications, [2]Xiaohongshu Inc.

Code link: https://github.com/CnFaker/SLQ

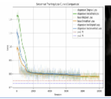
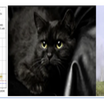
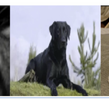
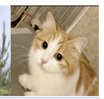
Motivation and Contribution

Motivation

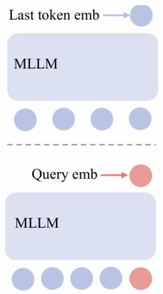
- Invasive adaptation updates MLLM parameters and may distort the pre-trained semantic space, and cost more computationally prohibitive overhead.

- **MLLMs have already learned an aligned multimodal representation space; retrieval adaptation should elicit latent representations, not retrain the model.**

Diagnostic pilot

Text \ Image					
A black cat.	Last token	0.77	0.83 √	0.77	0.77
	Query	0.57	0.84 √	0.78	0.82
An animal that barks	Last token	0.72	0.79	0.81 √	0.71
	Query	0.66	0.77	0.84 √	0.76
An animal has (2+7) lives	Last token	0.80 ×	0.71	0.68	0.68
	Query	0.61	0.75 √	0.70	0.73 √

(a) Retrieval cases across three levels. Row 1: Pattern matching. Row 2: Knowledge Retrieval. Row 3: Logical Reasoning.



(b) Last token vs. Query

Figure 2. **Diagnostic pilot study.** We compare the zero-shot retrieval performance of the last token baseline against query token method on the InternVL3-1B backbone. The retrieval score based on cosine similarity is reported as the metric. The results suggest that the query token method better aggregates global context, enabling implicit reasoning for retrieval.

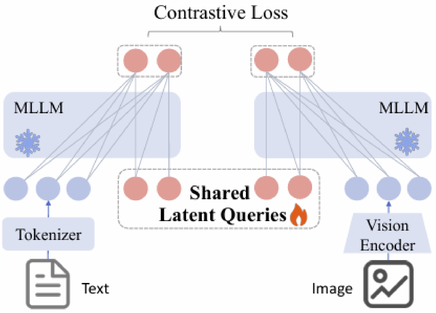
By freezing the MLLM, using queries can better elicit the MLLM's intrinsic knowledge and reasoning for retrieval.

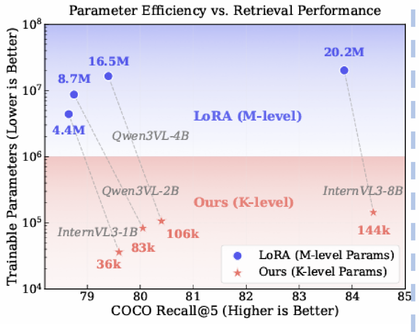
Contribution

- Efficient MLLM-to-Retriever Adaptation.
- Knowledge-Aware Reasoning Retrieval Benchmark.
- Strong Performance with Minimal Overhead.

Method

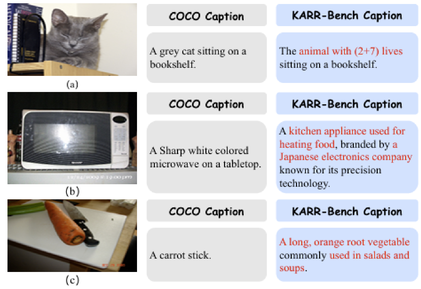
SLQ is a parameter-efficient framework that adapts frozen MLLMs for retrieval by appending a set of Shared Latent Queries to text and image tokens, using causal attention to aggregate multimodal context into a unified embedding space.



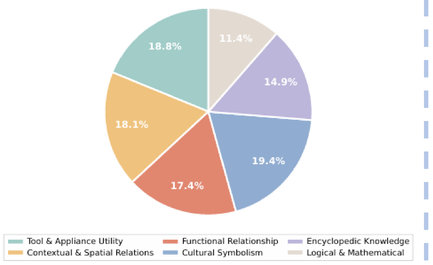


KARR-Bench

A benchmark for implicit, knowledge-aware retrieval rather than surface caption matching.



(a) Comparison between COCO and KARR-Bench



(b) Distribution of KARR-Bench Types

Figure 3. **Overview of KARR-Bench.** (a) Comparison between standard explicit captions and our knowledge reasoning captions. (b) The comprehensive distribution of categories in KARR-Bench.

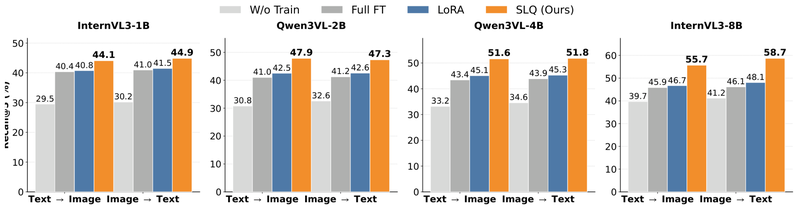
Experiment

Main result on COCO and Flickr30K

Table 1. **Performance comparison on Flickr30K and COCO retrieval benchmarks.** The best results are highlighted in **bold** and the second-best are underlined.

Model	Flickr30K (1K test set)						COCO (5K test set)					
	Image → Text			Text → Image			Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Dual-Encoder												
CLIP ViT-B (Radford et al., 2021)	77.8	95.0	98.2	58.8	83.3	89.8	51.0	74.9	83.5	30.5	56.0	66.8
CLIP ViT-L (Radford et al., 2021)	87.2	98.3	99.4	67.3	89.0	93.3	58.1	81.0	87.8	37.0	61.6	71.5
BLIP ViT-L (Li et al., 2022)	75.5	95.1	97.7	70.0	91.2	95.2	63.5	86.5	92.5	48.4	74.4	83.2
FLAME (Cao et al., 2025)	86.4	97.3	98.6	73.3	91.7	95.5	60.5	82.9	89.3	43.9	70.4	79.7
MLLM-based												
E5-V-7B (Jiang et al., 2024a)	88.2	98.7	99.4	79.5	95.0	**97.6**	62.0	83.6	89.7	52.0	76.5	84.7
TIGeR (Qu et al., 2024)	-	-	-	71.7	91.8	95.4	-	-	-	46.1	69.0	76.1
VLM2VEC-7B (Jiang et al., 2024b)	**94.6**	**99.5**	**99.8**	80.3	95.0	97.4	68.5	88.4	93.4	49.2	73.8	83.3
Ours												
SLQ (InternVL3-1B)	86.7	97.8	99.6	74.4	92.9	95.9	61.2	84.9	91.8	48.5	74.3	83.1
SLQ (Qwen3VL-2B)	85.8	97.7	99.1	76.4	93.5	96.3	62.7	84.4	90.7	50.2	75.7	83.9
SLQ (Qwen3VL-4B)	85.9	97.9	99.6	76.8	93.4	96.1	64.3	85.6	91.2	50.4	75.2	83.4
SLQ (InternVL3-8B)	92.0	99.4	**99.8**	**81.8**	**95.1**	**97.6**	**69.6**	**89.1**	**93.8**	**55.4**	**79.7**	**86.8**

Knowledge-Aware Reasoning Retrieval



As the model scale increases, SLQ achieves greater gains on knowledge reasoning and retrieval tasks.

Tuning Strategies: Efficiency vs. Capability Preservation

Table 3. **Comparison of tuning strategies.** We report Recall@5 for retrieve and assess the retention of the MLLM's inherent vision language understanding capability on MMMU, RealWorldQA, and OCRBench. Training cost is measured in GPU hours on COCO for 5 epochs using H800 GPUs.

Backbone	Dim.	Method	Param.	GPU Hour	Flickr30K		COCO		MMMU	RealWQA	OCRBench
					IR	TR	IR	TR			
InternVL3-1B	896	Full FT	0.6B	18.0	90.8	94.8	72.4	84.2	40.7	54.3	758
		LoRA	4.4M	11.4	90.4	95.7	72.9	84.4	42.1	56.8	785
		SLQ	**36k**	**3.9**	**92.9**	**97.8**	**74.3**	**84.9**	**43.4**	**58.2**	**790**
Qwen3VL-2B	2048	Full FT	1.7B	42.5	91.7	97.0	73.2	82.1	51.7	59.4	824
		LoRA	8.7M	23.8	92.1	97.2	74.8	82.7	53.1	62.8	832
		SLQ	**83K**	**6.7**	**93.5**	**97.7**	**75.7**	**84.4**	**53.4**	**63.9**	**858**
Qwen3VL-4B	2560	Full FT	4B	96.2	92.0	97.2	74.5	82.8	63.8	66.7	851
		LoRA	16.5M	48.6	92.7	96.8	75.1	83.7	65.4	68.3	857
		SLQ	**106K**	**13.5**	**93.4**	**97.9**	**75.2**	**85.6**	**67.4**	**70.9**	**881**
InternVL3-8B	3584	Full FT	7.6B	403.8	94.0	96.6	78.2	87.4	59.8	66.5	847
		LoRA	20.2M	130.1	94.4	98.7	79.4	88.3	60.9	67.2	843
		SLQ	**144k**	**38.9**	**95.1**	**99.4**	**79.7**	**89.1**	**62.7**	**70.8**	**880**