MKG-RAG-Bench: Benchmarking Retrieval in Multimodal Knowledge Graph–Augmented Generation

Xiaochen Wang, Bao Hoang, Han Liu, Ting Wang, Fenglong Ma

Presenter: Bao Hoang

PSU Data Science Lab@Pennsylvania State University

Motivation: Existing Benchmarks Miss MKG Retrieval

Existing literature / benchmark families

Multimodal RAG

Examples: OK-VQA, M2RAG, CRAG-MM, Dyn-VQA

Weakness: retrieve web passages/images from unstructured sources; usually generation-focused, not explicit MK triplet retrieval.

Textual KG-RAG / KGQA

Examples: WebQSP, CWQ, GrailQA

Weakness: structured KG reasoning, but mostly text-only; cannot test image-grounded entities or multimodal triplets.

Document MM-RAG

Examples: REAL-MM-RAG, DocBench, MMLongBench, MMDocRAG

Weakness: retrieval unit is PDF/document chunk, not graph triplet or structured relation.

Table 1: Comparison of representative multimodal RAG and KG-RAG benchmarks.

Benchmark Name	Knowledge Source Type	Knowledge Structure	Retrieval Unit	Heterogeneous Query Support	Evaluation Granularity	Cross-Domain Coverage
OK-VQA [26]	Web	Unstructured	Passage	Multimodal	Generation-only	Single-domain
M^2RAG [25]	Web	Unstructured	Passage and Image	Multimodal	Generation-only	Single-domain
CRAG-MM [37]	Web	Ad-hoc MKG	Passage and Image	Multimodal	Generation-only	Multi-domain
Dyn-VQA [20]	Web	Unstructured	Passage and Image	Multimodal	Generation-only	Multi-domain
WebQSP [43]	Textual KG	Textual KG	Textual Triplet	Text-only	Generation-only	Single-domain
CWQ [35]	Textual KG	Textual KG	Textual Triplet	Text-only	Generation-only	Single-domain
GrailQA [10]	Textual KG	Textual KG	Textual Triplet	Text-only	Generation-only	Multi-domain
REAL-MM-RAG [41]	Multimodal Documents	PDF	Chunk	Text-only	Retrieval-only	Multi-domain
DocBench [46]	Multimodal Documents	PDF	Chunk	Text-only	Generation-only	Multi-domain
MMLongBench [24]	Multimodal Documents	PDF	Chunk	Text-only	Generation-only	Multi-domain
MMDocRAG [6]	Multimodal Documents	PDF	Chunk	Text-only	Retrieval & Generation	Multi-domain
MKG-RAG-Bench	Multimodal KG	Explicit MKG	Multimodal Triplet	Multimodal	Retrieval & Generation	Multi-domain

Key gap: **No** benchmark systematically evaluates **both** retrieval and generation in **multimodal KG-RAG**, where queries and evidence can involve text, images, and graph relations.

Why Do We Need MKG-RAG-BENCH

- Naive MKG retrieval can **underperform** RAG-free generation.
  - Task-critical knowledge may be missing from the MKG.
  - Irrelevant retrieved triplets can inject noise into the MLLM input.

=> This motivates an aligned benchmark connecting KG, retrieval targets, and downstream evaluation.

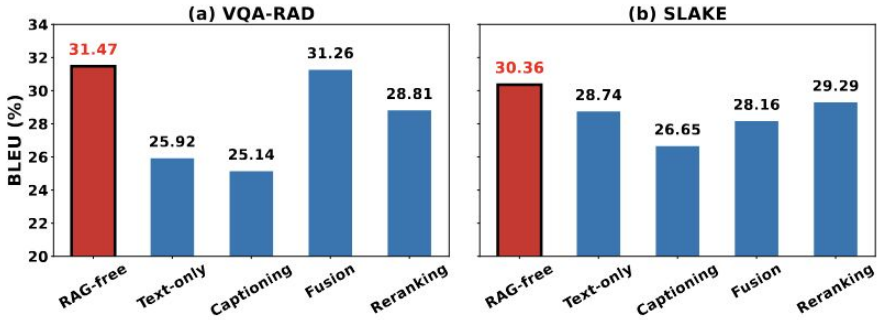
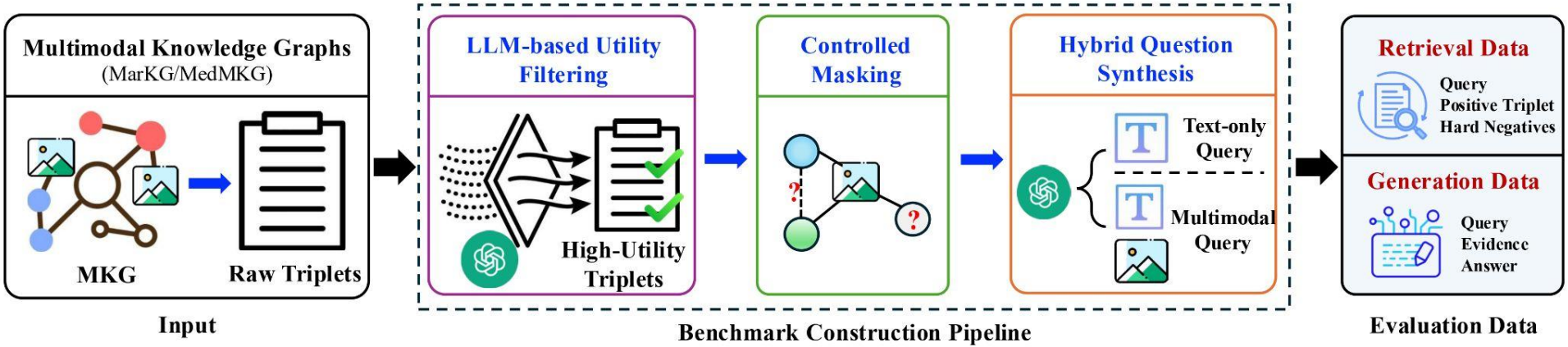


Figure 2: BLEU (%) comparison on (a) VQA-RAD and (b) SLAKE datasets. RAG-free consistently achieves the best performance across both benchmarks.

What We Built: MKG-RAG-Bench



Dataset recipe

(1) Curate high-utility MKG triplets
(2) mask relation/tail
(3) synthesize text-only or image-grounded questions
(4) use original triplet as retrieval positive.

Two Domains

(1) General: MarKG
(2) Medical: MedMKG

Examples

(1) (penicillin, [Mask], bacterial infection) corresponds “What is the relationship between penicillin and bacterial infection?”
(2) (Eiffel Tower, located_in, [Mask]) corresponds “Which city is the landmark in the image located in?” 

Benchmark Statistic

Table 2: Statistics of MKG-RAG-BENCH, where %MM refers to the percentage of multimodal data.

Dataset	Split	Retrieval						Generation				
		Query			Triplet			Question			Answer	
		# of Queries	%MM	AvgLen	# of Triplets	%MM	AvgLen	# of Questions	%MM	AvgLen	# of Answers	AvgLen
MKG-RAG-BENCH-G	Train	48,908	30.1	12.2	25,517	28.8	6.8	48,908	30.1	12.2	61,001	2.5
	Val	6,113	31.0	12.2	25,517	28.8	6.8	6,113	31.0	12.2	7,586	2.5
	Test	6,115	30.3	12.2	25,517	28.8	6.8	6,115	30.3	12.2	7,795	2.5
MKG-RAG-BENCH-M	Train	4,781	40.6	16.2	18,468	49.0	8.7	4,781	40.6	16.2	16,564	3.7
	Val	597	42.2	16.1	18,468	49.0	8.7	597	42.2	16.1	2,030	3.5
	Test	599	43.1	16.2	18,468	49.0	8.7	599	43.1	16.2	2,074	3.6

Evaluation settings

Retrieval

Rank the correct KG triplet among text-only and multimodal candidates.

Metrics: NDCG@K, Precision@K, Recall@K

Generation

Answer open-ended text or visual questions using the retrieved evidence.

Metrics: EM, F1, Contains@1, BLEU-1

Five modality settings cover all/text-only/multimodal queries and triplets.

Key Findings from Experiments

1 A modality gap persists in retrieval

Adding multimodal triplets does not automatically help text-only queries; mixed-modality indexing alone is not enough for cross-modal matching.

2 Multimodal embeddings are critical

For visually grounded queries, text-only retrieval and captioning degrade sharply; effective retrieval needs representations that preserve visual evidence.

3 Retrieval is domain-dependent

Fusion is strong on the general-domain benchmark, while reranking helps more in the medical domain with subtle images and specialized terminology.

4 RAG gives genuine but uneven utility

Retrieved triplets improve generation in most settings, validating the benchmark alignment, but gains depend on whether evidence is relevant.

5 Multimodal generation gains are weaker

Retrieval helps most in text settings; visually grounded questions remain difficult unless strong multimodal retrievers provide useful evidence.

6 Generation tracks retrieval quality

End-to-end MKG-RAG performance closely follows retrieval quality, motivating graph-aware retrievers and better evidence formatting.

Takeaway: MKG-RAG needs graph-aware, domain-sensitive multimodal retrieval and clearer triplet evidence integration.

Thanks!

https://psudslab.github.io/index.html

Acknowledgement: This research was partially supported by a 2025/2026 Rising Researcher Grant from Penn State's Institute for Computational & Data Sciences (RRID:SCR_025154) and the National Science Foundation under Grant No. 2333790 and 2238275