M^{3}KG-RAG: Multi-hop Multimodal Knowledge Graph-enhanced Retrieval-Augmented Generation

Supplementary Material

Overview

This supplementary material provides additional implementation details, experimental results, and qualitative analyses for our proposed framework, M^{3}KG-RAG.

- In Sec. A, we describe the implementation of M^{3}KG-RAG in detail, including the construction of the M^{3}KG and the multimodal RAG framework.
- In Sec. B, we present additional analyses of our multimodal RAG framework, including hyperparameter sensitivity and ablations over key components with their computational cost.
- In Sec. C, we provide additional qualitative results of M^{3}KG-RAG on multimodal benchmarks, including win-rate evaluations.
- In Sec. D, we discuss the limitations of M^{3}KG-RAG and potential directions for improving robustness.

A. Extended Implementation Details

A.1. M^{3}KG Construction

In this section, we describe how M^{3}KG is built using a lightweight multi-agent pipeline, outlining the roles and designs of each agent. We also summarize the multimodal corpora used for construction.

A.1.1. Multi-Agents

We provide detailed descriptions and prompt designs for the multi-agent system used in M^{3}KG construction, including *rewriter*, *extractor*, *normalizer*, *searcher*, *selector*, *refiner*, and *inspector*.

Rewriter The *rewriter* transforms generic textual descriptions from the raw multimodal corpus into knowledge-intensive captions that are more informative for MLLMs. For each data point, it leverages the crawled YouTube title and description to inject unfamiliar concepts and background knowledge that are not explicitly captured in the original text caption. The detailed prompt design for the *rewriter* agent is provided in Table 1.

Rewriter agent
`<system prompt>` You refine video captions using the video's Title and Description. ORIGINAL CAPTION always has priority. If Title/Description are not clearly referring to the SAME scene/object/action, output the ORIGINAL CAPTION exactly. Allowed edits ONLY when clearly aligned: replace generic nouns with specific terms (breed/species/instrument/model/place/role), or add 1 short factual attribute. Keep the meaning and keep the length roughly similar (±20%). Disallowed: inventing new events, numbers, counts, or speculative facts; adding ads/URLs/hashtags. Keep the original style. Make the merge natural (not a concat). Output: ONLY the final caption in English (no labels or explanations).
`<user prompt>` Title: {TITLE} Description: {DESCRIPTION} ORIGINAL CAPTION: {ORIGINAL_CAPTION} Output:

Table 1. Prompt template for the *rewriter* agent.

Extractor The *extractor* takes the rewritten, knowledge-intensive captions produced by the *rewriter* and extracts structured knowledge in the form of triplets. Building on the LLM-based open information extraction prompt used in VAT-KG [11], we design the prompt for the *extractor* agent, as shown in Table 2.

Normalizer The *normalizer* operates on the head and tail entities in the triplets extracted by the *extractor* and standardizes them into canonical, searchable concepts. The prompt design for the *normalizer* agent is provided in Table 3.

Searcher The *searcher* takes the normalized entities produced by the *normalizer* and queries external knowledge resources (e.g., Wikipedia, Wiktionary) to obtain encyclopedic descriptions for each concept. If it cannot find a description from these sources, it invokes an LLM callback [17] to generate a brief description for the entity, leveraging context-enriched caption. The prompt designs for the LLM callback of the *searcher* agent are provided in Table 4.

Selector The *selector* takes multiple candidate descriptions for each entity and uses the context-enriched caption produced by the *rewriter* as context to select the most appropriate description. The prompt design for the *selector* agent is provided in Table 5.

Refiner The *refiner* takes the descriptions selected for each entity's canonical form and refines them to better match the semantics and surface form of the original entity mention, while preserving the underlying factual content. The prompt design for the *refiner* agent is provided in Table 6.

Inspector The *inspector* serves as a quality-control agent that implements the self-reflection loop in our construction pipeline. For each entity description produced either by the

Extractor agent
`<system prompt>` You are an expert in extracting structured knowledge from text. Given a video caption, extract all subject-relationship-object triples in the form (h, r, t). Extract multiple (h, r, t) triples if applicable. Each triple must be meaningful and correctly represent relationships in the text. Output format: ONE triple per line as (h, r, t). No extra text or explanation. Use concise surface forms that appear (or are directly implied) in the caption. Do not invent entities or facts not supported by the caption. Language: English only.
`<user prompt>` Caption: {CAPTION} Output:

Table 2. Prompt template for the *extractor* agent.

Normalizer agent
`<system prompt>` You output exactly ONE KB-searchable concept noun phrase (Wikipedia-title-like). Plain text only, no quotes or extra words. Output MUST be in English. If `CONCEPT` is not in English, translate the noun phrase into an English Wikipedia-style title; transliterate proper names if needed (do not add extra words). Otherwise, use ONLY words from `CONCEPT`; keep order; you may DROP words (no inventions/translation). If the output noun phrase is in plural form, convert it to its singular form (e.g., "dogs" → "dog", "empires" → "empire"). Must be a NOUN PHRASE; remove wrappers like "how to", "what is", guides/tips, articles, years. Prefer inner object NP (e.g., "history of jazz music" → "jazz music"). Prefer canonical/proper names; preserve original casing.
`<user prompt>` CONCEPT: {CONCEPT} Output:

Table 3. Prompt template for the *normalizer* agent.

refiner or by the LLM callback of the *searcher*, the *inspector* assigns a plausibility score on a 0–10 scale, conditioned on the context-enriched caption from Step 1. Descriptions scoring below 7 are sent back to the corresponding agent for regeneration and re-scoring; we allow at most three such iterations, after which persistently low-scoring descriptions are discarded to avoid injecting low-quality facts into the graph, while higher-scoring ones are accepted. The prompt design for the *inspector* agent is provided in Table 7.

A.1.2. Corpora for M^{3}KG Construction

We construct M^{3}KG from the training splits of three multimodal corpora: AudioCaps [6], ActivityNet [1], and

Searcher agent (LLM callback)
`<system prompt>` You are an encyclopedic writer. Write a neutral, generic, 1–2 sentence encyclopedic description of a concept. Use the caption ONLY to disambiguate the intended sense (do not describe the scene). Return ONLY the final description sentences in plain text—no labels, no lists, no quotes, no extra commentary.
`<user prompt>` Concept: {CONCEPT} Caption (sense disambiguation only): {CAPTION} Output:

Table 4. Prompt template for the LLM callback of the *searcher* agent.

Selector agent
`<system prompt>` You are a selector for concept descriptions. The CAPTION is from the same video, and the CONCEPT refers to the concept that appears or is mentioned in this video. Choose ONE candidate whose encyclopedic sense best matches that concept in this video's caption. Use ONLY the CAPTION to resolve meaning (sense); do not import outside facts. Return EXACTLY one candidate's text verbatim; do not edit, merge, summarize, quote, or label it. If multiple candidates are similarly valid, prefer the most specific non-speculative candidate. If none clearly fits, choose the safest generic candidate (least speculative). Return ONLY the chosen candidate text (no extra text).
`<user prompt>` CONCEPT: {CONCEPT} CAPTION: {CAPTION} CANDIDATES: {ENUMERATED_CANDIDATES} Output:

Table 5. Prompt template for the *selector* agent.

VALOR [8]. For graph construction, we use only the raw audio-visual content and their associated captions, without accessing any QA annotations.

AudioCaps AudioCaps is an audio captioning corpus built on 10-second clips from AudioSet [3] YouTube videos, where each clip is paired with human-written natural language descriptions of the acoustic scene and salient sound events.

ActivityNet ActivityNet is a large-scale video benchmark of untrimmed videos covering diverse human activities, annotated with temporal activity boundaries and class labels. In our construction pipeline, we segment each untrimmed video into temporally localized clips using the provided an-

Refiner agent

<system prompt>
You are a refiner for concept descriptions.
Adapt the selected description so it fits the original concept phrasing, preserving the original meaning and keeping the content of the selected description as intact as possible (minimal wording changes only—e.g., adjust possessives like "my/our/their", determiners, and surface phrasing to align with the original concept).
Do NOT add, remove, or invent facts beyond what is in the selected description.
Keep the meaning unchanged; only adapt phrasing to match the original concept.
Concise: 1–2 sentences, plain text
(no lists/quotes/markdown/meta).
Do NOT output any reasoning.
Return ONLY the rewritten description sentences.

<user prompt>
Concept (original phrasing): {ORIGINAL_CONCEPT}
Searchable concept (KB term): {SEARCHABLE_CONCEPT}
Selected description (about the searchable concept):
{SELECTED_DESCRIPTION}
Output:

Table 6. Prompt template for the *refiner* agent.

Inspector agent

<system prompt>
You are a judge that scores how well an encyclopedic DESCRIPTION matches the intended sense of a CONCEPT.
Sense scoring only.
Score 0–10 how well the DESCRIPTION's encyclopedic sense matches the intended sense of the CONCEPT (0 = different/irrelevant sense, 10 = perfect sense match).
Output a single integer 0–10 with no extra text.

<user prompt>
CONCEPT: {CONCEPT}
DESCRIPTION: {DESCRIPTION}
OUTPUT:

Table 7. Prompt template for the *inspector* agent.

notations, and use the corresponding activity class label for each clip as a text signal when building M^{3}KG.

VALOR VALOR is a multimodal dataset of short video clips with synchronized audio and human-authored audio-visual captions, providing closely aligned triplets of vision, audio, and text. We use the VALOR-32K variant for constructing M^{3}KG.

A.2. Multimodal RAG Framework

In this section, we provide additional details of the multimodal RAG framework used in our experiments. Given a multimodal query, we first perform modality-wise retrieval over M^{3}KG. We then apply GRASP, which leverages multimodal grounding models [9, 16] and an LLM-based filter implemented with Qwen3-8B [17] using the instruction in Table 8 to obtain a compact set of query-relevant and answer-supportive triplets, and finally inject this evidence into the MLLM using the graph-augmented generation scheme in Eq. (6) of the main paper. In practice, Eq. (6) is realized by using the template summarized in Table 9.

LLM-based GRASP filter

You are a selector that removes only unnecessary triples for answering the query.
Keep triples that could be helpful to answer the query.
Remove triples that are clearly irrelevant, contradictory to the query, or redundant duplicates.
When uncertain, prefer KEEPING the triple.
Preserve the ORIGINAL ORDER of kept indices (do NOT rerank).
Query: {QUERY}
Triplets: {TRIPLETS}

Table 8. Instruction used for the LLM-based filter in GRASP over retrieved triplets.

Multimodal RAG Prompt

You are a multimodal QA assistant. Prioritize PRIMARY evidence from the input modalities you perceive.
Use the retrieved triples BELOW only as optional hints when they are CLEARLY observed or corroborated in the input.
Procedure:
1) Detect whether any triple's context appears in the input (entities, attributes, actions, time/place cues).
2) If matched, integrate the FULL triple (head, relation, tail) into the answer, and enrich with head_desc/tail_desc.
– Do NOT contradict the primary evidence; if conflict exists, ignore the triple.
3) If no triple is confidently matched, answer from the primary evidence only.
Query : {QUERY}
Retrieved Triples : {TRIPLES_BLOCK}
Triple Format : [i] head={h} | relation={r} | tail={t} || head_description={hd} | tail_description={td}
Answer :

Table 9. Graph-augmented generation template used to instantiate Eq. (6) in our multimodal RAG framework.

A.2.1. Baselines for Multimodal RAG

Following prior multimodal RAG work [11], we compare M^{3}KG-RAG against four knowledge-graph baselines coupled with RAG, in addition to the *None* setting where the MLLMs answer without external knowledge.

Wikidata5M Wikidata5M [14] is a million-scale text-only knowledge graph constructed from Wikidata entities and relations aligned with their textual descriptions from Wikipedia.

Reference-Aware Win-rate Prompt
You will evaluate two answers to the same question using a Reference Answer. Base every judgment solely on alignment to the Reference and the Question; do not reward verbosity or speculative content. Question: {QUESTION} Reference Answer (trusted ground truth): {REFERENCE} Answer 1: {ANSWER_1} Answer 2: {ANSWER_2} Evaluate on the following criteria and return JSON in the exact schema below. - Comprehensiveness: Which answer correctly covers more of the Reference's essential points (paraphrase allowed) with fewer mistakes or omissions? Do not reward length; penalize unsupported/contradictory claims. - Diversity: Which answer offers greater variety in organizing the Reference's facts (e.g., visual vs. audio facets) while avoiding new attribute categories not stated or trivially entailed? - Empowerment: Which answer better enables understanding or action through clear, concise, and reference-aligned guidance (no filler, no meandering)? - Overall Winner: Choose the answer that is most faithful to the Reference, with stronger correct coverage and clearer, more concise presentation. Break ties by (1) correctness/faithfulness, (2) coverage, (3) concision/clarity.

Table 10. Reference-aware win-rate comparison template used for LLM-judged preferences.

VTKG VTKG [7] is an image-text multimodal knowledge graph that augments a textual KG with visual evidence by attaching images to entities and relational triples, together with short textual descriptions. This design provides entity–relation graphs grounded in visual examples, enabling concept nodes to be linked not only by symbolic relations but also by associated images.

M^{2}ConceptBase M^{2}ConceptBase [18] is a concept-centric multimodal knowledge base that represents each concept as a node with multiple aligned visual examples and a detailed textual description. It is explicitly designed to provide fine-grained, cross-modal concept knowledge that can be passed to MLLMs as grounded external evidence, helping mitigate hallucinated or semantically inconsistent predictions.

VAT-KG VAT-KG [11] is a knowledge-intensive multimodal knowledge graph that jointly integrates visual, audio, and textual signals into a unified concept-centric graph. Each triplet is linked to multimodal evidence and enriched with concept descriptions, providing an audio-visual KG backbone tailored for retrieval-augmented generation under multimodal queries.

A.2.2. Evaluation Protocol

As described in the main paper, our benchmarks consist of open-ended QA with free-form responses, so we adopt an off-the-shelf Model-as-Judge (M.J.) metric [13], where an LLM judge [4] scores each generated answer on a 0-5 scale given the query and reference answer and reports the resulting score on a 0-100 scale.

In addition, we report a pairwise win-rate between M^{3}KG-RAG and each baseline, following RAG evaluation protocols based on LLM preferences [2, 5, 12]. Unlike prior work that compares two LLM-generated answers, we make the win-rate judge reference-aware by providing the reference answer alongside the two candidates. This helps reduce verbosity bias (overly favoring longer responses) and discourages rewarding merely plausible but unsupported generations, leading to a more faithful and stable preference signal. The exact evaluation instruction for the reference-aware win-rate judge is provided in Table 10. The judge compares the two answers according to three criteria—*Comprehensiveness*, *Diversity*, and *Empowerment*—and selects a preferred answer for each criterion. Based on these per-criterion preferences, it then decides which answer is preferred overall for each query.

B. Additional Analysis

B.1. Hyperparameter Sensitivity Analysis

Our multimodal RAG framework has two scalar hyperparameters: (i) the modality-wise retrieval distance threshold τ and (ii) the GRASP presence score threshold η. Both control how much knowledge is injected into the MLLMs and therefore may affect downstream QA performance. To assess the robustness of our framework to these choices, we conduct a sensitivity study on the VALOR benchmark by varying τ and η_{av} and measuring the resulting M.J. scores.

Modality-wise distance threshold τ For each query, we embed it into the modality-specific representation spaces of M^{3}KG and compute distances to candidate items. Concretely, an audio-only query is compared against audio items in M^{3}KG within the audio embedding space, and a visual-only query is compared against visual items within the visual embedding space. For an audio-visual query, we concatenate its audio and visual embeddings and match it to audio-visual items in the joint concatenated space. In each active modality m, we compute an L2 distance $d(q_m, x_m)$ between the query representation q_m and an item x_m from M^{3}KG, and use it to perform top-k retrieval. We then apply the distance threshold τ to these k candidates and keep only items with $d(q_m, x_m) \leq \tau$. The remaining retrieved items are lifted

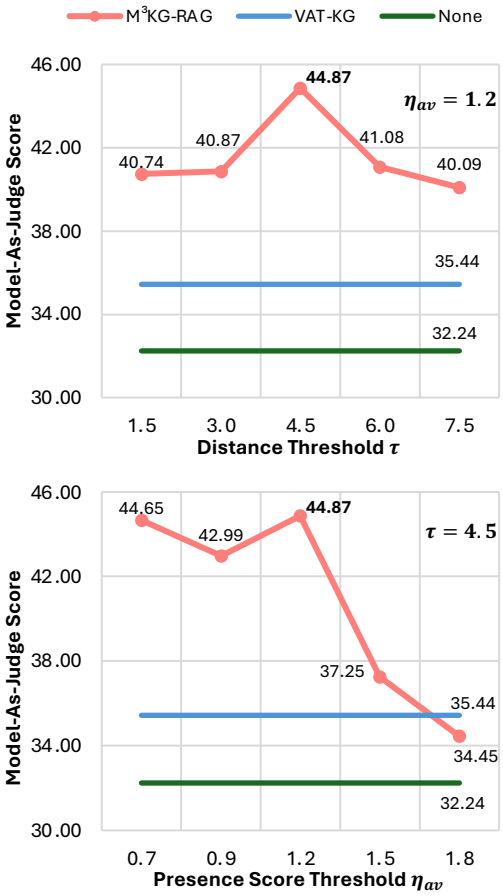


Figure 1. **Sensitivity analysis.** M.J. score on VALOR versus modality-wise distance threshold τ (top) and GRASP presence threshold η_{av} (bottom).

into the graph via Eq. (3) in the main paper.

To understand how the choice of τ affects QA performance, we conduct a sensitivity study on the VALOR benchmark by varying $\tau \in \{1.5, 3.0, 4.5, 6.0, 7.5\}$ while fixing $\eta_{av} = 1.2$, and visualize the resulting M.J. scores in the top plot of Fig. 1. As shown in the figure, the M.J. score is maximized at $\tau = 4.5$, while both smaller and larger thresholds yield lower scores. Nonetheless, across all tested values of τ, M^{3}KG-RAG consistently outperforms the baselines. When τ is set too small, only very close items are retained, so the retrieved subgraphs become overly sparse and may not provide sufficient evidence for the MLLM. Conversely, a large τ allows many more distant items to pass the filter and expands the subgraph, but also introduces multi-hop nodes that are only weakly related to the query, increasing the risk of noisy or distracting knowledge. These observations are consistent with the intended role of τ in balancing coverage and noise in modality-wise retrieval.

GRASP presence score threshold η GRASP assigns a query-conditioned presence score $s(\mathrm{t} \mid q)$ to each triplet t in the retrieved subgraph using an off-the-shelf multimodal grounding model [9, 16]. We then apply the presence score threshold η and prune triplets whose scores fall below it, keeping only those with $s(\mathrm{t} \mid q) \geq \eta$.

To examine how η affects QA performance, we fix $\tau = 4.5$ and vary $\eta_{av} \in \{0.7, 0.9, 1.2, 1.5, 1.8\}$ on the VALOR benchmark, visualizing the resulting M.J. scores in the bottom plot of Fig. 1. The performance is highest at $\eta_{av} = 1.2$, with only minor differences between $\eta_{av} = 0.7$, 0.9, and 1.2, but it drops noticeably once η increases to 1.5 and 1.8. This pattern suggests that moderate grounding-based pruning is beneficial, whereas overly aggressive thresholds remove many triplets that still carry useful evidence, leaving the MLLM with an under-informative subgraph.

GRASP Component	GPU VRAM (GB)	Avg time / query (s)	M.J.
None	23.0	4.30	40.91
GDino (η_v=0.8)	23.7	5.75	41.35
TAG (η_a=0.4)	23.6	4.48	41.70
GDino + TAG (η_{av}=1.2)	24.2	6.02	42.96
GDino + TAG + LLM Filter	39.8	7.02	**44.87**

Table 11. **Ablation on GRASP Components.** We report GPU VRAM usage, average inference time per query, and M.J. score.

B.2. Additional Ablation Studies

In Sec. 4.3 of the main paper, we ablate modality-wise retrieval and GRASP as a whole. We further decompose GRASP into its three submodules in the audio-visual setting: visual grounding with GroundingDINO [9] (GDino), audio grounding with TAG [16], and the final LLM-based filtering stage. Tab. 11 reports GPU VRAM, average inference time per query, and M.J. scores on the VALOR benchmark as we progressively enable these components, where GPU VRAM is measured after loading each additional module.

Starting from the configuration without GRASP, adding either GDino or TAG alone yields small but consistent gains over the base M^{3}KG-RAG model (from 40.91 to 41.35 and 41.70 M.J., respectively). Using both grounding modules together further improves performance to 42.96 M.J., indicating that audio and visual grounding provide complementary benefits. Importantly, the GPU memory footprint remains almost unchanged when moving from a single grounding module to both (about 23.6–24.2 GB), and the average latency stays within 4.3–6.0 seconds per query.

Finally, enabling the LLM-based filtering stage on top of GDino and TAG achieves the best performance of 44.87 M.J., a gain of nearly 4 points over the configuration without GRASP and about 2 points over using only the grounding modules. This improvement comes with a moderate increase in resource usage (VRAM from 24.2 GB to 39.8 GB and average time from 6.02 s to 7.02 s per query), while the LLM-based filter helps focus the retrieved subgraph on answer-supporting knowledge. Overall, these results show that each GRASP submodule contributes positively to performance, and that the full GRASP pipeline offers the best accuracy with a relatively modest overhead compared to its benefits.

C. Additional Qualitative Results

In this section, we present additional qualitative comparisons of M^{3}KG-RAG against VAT-KG [11] on multimodal QA benchmarks, including Audio QA (AudioCaps-QA [13]), Video QA (VCGPT [10]), and Audio-Visual QA (VALOR [8]), using Qwen2.5-Omni [15] as the base MLLM. For each benchmark, we show the knowledge retrieved with VAT-KG and with M^{3}KG-RAG, the corresponding answers generated from these contexts, and the win-rate judge's preference and rationale. Both methods construct MMKGs from raw multimodal corpora and support modality-wise retrieval. However, VAT-KG represents each multimedia item with a single-hop graph and relies on shallow similarity search, often yielding sparse or weakly aligned evidence. In contrast, M^{3}KG-RAG exploits multi-hop knowledge and GRASP-based pruning to serve richer, query-relevant context to the MLLM, leading to more faithful and informative responses.

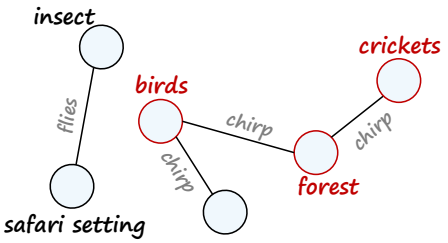
For the Audio QA case in Figure 2, the query asks what animals can be heard in a clip where birds chirp in a forest environment with background insect sounds. VAT-KG primarily retrieves a single fact about a flock of birds in the forest, leading the model to produce an answer that mentions only birds. In contrast, M^{3}KG-RAG retrieves multi-hop knowledge that links both birds and crickets chirping in a forest setting, providing richer cues about co-occurring animal sounds. Conditioned on this context, the model identifies both birds and insects as audible in the scene, which the win-rate judge prefers for covering all relevant animal sources and better matching the reference audio.

For the Video QA case in Figure 3, the query video shows a woman playing racquetball on an indoor court, and the model is asked to describe in detail what happens in the scene. VAT-KG performs coarse similarity-based retrieval that includes knowledge about both squash and racquetball, two related but distinct sports, which leads the model to produce a hedged response that refers to a game similar to squash or racquetball without clearly committing to the actual activity or capturing fine-grained details. In contrast, M^{3}KG-RAG, together with GRASP, prunes off-topic neighbors based on fine-grained query relevance and supplies racquetball-focused multi-hop evidence that matches the video. Guided by this evidence, the model correctly identifies the sport as racquetball and gives a more precise description of the player's attire, court setting, and actions, which the win-rate judge prefers for its specificity and semantic alignment with the video.

For the Audio-Visual QA case in Figure 4, the multimodal query shows a man playing an electric guitar, and the model is asked to describe the scene. VAT-KG, due to its single-hop MMKG structure, mainly connects the man to a generic guitar and to a musician–acoustic-guitar relation, providing only fragmentary, coarse knowledge. Without fine-grained audio–visual relevance checking, it treats acoustic and electric guitars as semantically interchangeable, which leads the model to describe the scene as an acoustic guitar performance and to miss surrounding contextual details. In contrast, M^{3}KG-RAG retrieves a multi-hop neighborhood around the guitar that includes electric-guitar–specific context (such as playing with an effects setup) together with local scene cues (e.g., the man sitting on a chair in an indoor room). With this richer, better-aligned evidence, the model correctly identifies the instrument as an electric guitar and produces a more detailed description of the player's appearance and environment, which the win-rate judge prefers for both factual correctness and contextual richness.

D. Limitations

Despite the advancements presented in this work, several limitations remain. First, the knowledge coverage of M^{3}KG is bounded by the raw multimodal corpora used for construction, which may underrepresent long-tail entities, rare relations, or domains not well captured in the source datasets. Second, M^{3}KG–RAG relies on the multimodal encoders and grounding models used in the retrieval phase; when their training domains do not cover the query distribution, cross-modal miscalibration can surface off-topic neighbors and degrade evidence quality. Extending M^{3}KG-RAG with better-calibrated encoders, broader and continually updated corpora, and tighter grounding mechanisms is an important direction for future work.

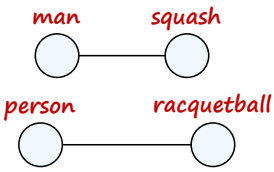
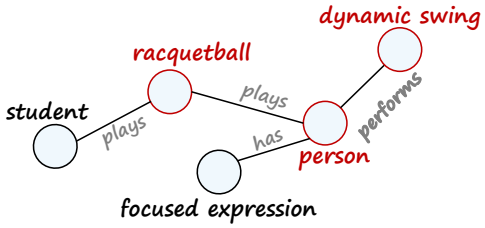
<table>
<tr><td>[Query]</td><td colspan="2">What animals can be heard in the audio?
 
(bird chirps and flies in the audio.)</td></tr>
<tr><td rowspan="2">[Retrieved Subgraph]</td><td>VAT-KG</td><td>M^{3}KG-RAG</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>[Answer 1] (VAT-KG)</td><td colspan="2">“The audio contains the sound of a bird chirping.”</td></tr>
<tr><td>[Answer 2] (M^{3}KG-RAG)</td><td colspan="2">“The animals that can be heard in the audio are birds and insects.”</td></tr>
<tr><td>[Judge Response]</td><td colspan="2">"Comprehensiveness": "Winner": "Answer 2 (M^{3}KG-RAG)"
"Explanation": "Answer 2 is more comprehensive because it provides more information about the animals that can be heard in the audio. It not only mentions birds, but also insects, giving a more detailed and thorough response to the question. Answer 1 only mentions a single bird chirping, which is a limited and incomplete description of the audio content."

"Diversity": "Winner": "Answer 2"
"Explanation": "Answer 2 shows more diversity in its response by mentioning two different types of animals: birds and insects. This provides a richer and more varied answer, offering the reader a broader understanding of the audio content. Answer 1, on the other hand, only focuses on a single aspect, which limits its diversity and range."

"Empowerment": "Winner": "Answer 2"
"Explanation": "Answer 2 empowers the reader more because it provides a more detailed and nuanced understanding of the audio content. By mentioning both birds and insects, it gives the reader a clearer picture of what they might expect to hear, allowing them to make more informed judgments about the audio. Answer 1, with its limited description, does not provide the same level of empowerment."

"Overall Winner": "Winner": "Answer 2"
"Explanation": "Answer 2 is the overall winner because it excels in all three criteria. It provides a more comprehensive and detailed description of the audio content, offers a more diverse range of information, and empowers the reader to better understand and engage with the topic. Overall, Answer 2 is a more informative, thorough, and engaging response to the question."</td></tr>
</table>

Figure 2. **Qualitative Comparison on Audio QA.** Comparing VAT-KG and M^{3}KG-RAG with Qwen2.5-Omni, including retrieved knowledge and win-rate judge preferences.

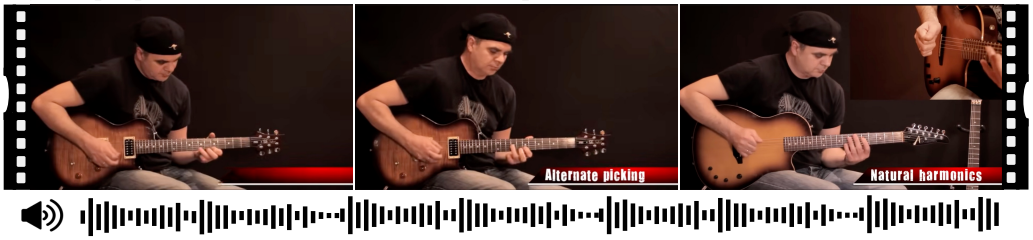
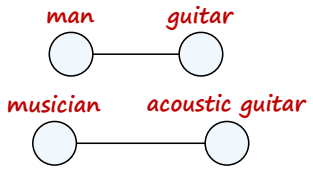
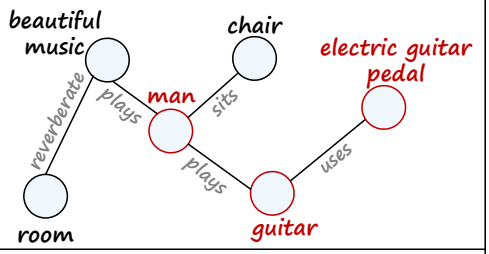
<table>
<tr><td>[Query]</td><td colspan="2">Can you describe the details of what happens in the video?
  </td></tr>
<tr><td rowspan="2">[Retrieved Subgraph]</td><td>VAT-KG</td><td>M^{3}KG-RAG</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>[Answer 1] (VAT-KG)</td><td colspan="2">"The video appears to show a person playing a sport in a walled court. The person is holding a racquet and seems to be engaged in a game similar to squash or racquetball. The setting and equipment suggest that the activity involves hitting a ball against the walls of the court. The person is actively moving around the court, indicating they are participating in a fast-paced game."</td></tr>
<tr><td>[Answer 2] (M^{3}KG-RAG)</td><td colspan="2">"The video shows a person playing racquetball in an indoor court. The individual is wearing athletic attire, including a white shirt and black shorts, and is using a racquet to hit the ball against the walls. The court's walls, floor, and ceiling serve as playing surfaces, which is typical for racquetball. The person appears to be actively engaged in the game, moving around to strike and return the ball as it rebounds."</td></tr>
<tr><td>[Judge Response]</td><td colspan="2">"Comprehensiveness": "Winner": "Answer 2 (M^{3}KG-RAG)"
"Explanation": "Answer 2 correctly covers more essential points from the Reference, including the indoor court, athletic attire, and active engagement in the game. Answer 1 omits these details and only provides a general description of the activity."

"Diversity": "Winner": "Answer 2"
"Explanation": "Answer 2 provides a more varied perspective by incorporating details about the court's walls, floor, and ceiling as playing surfaces, which is typical for racquetball. Answer 1 lacks this specific insight."

"Empowerment": "Winner": "Answer 2"
"Explanation": "Answer 2 offers clearer and more actionable guidance by providing specific details about the game, making it easier to understand the activity. Answer 1 is more general and lacks clear direction."

"Overall Winner": "Winner": "Answer 2"
"Explanation": "Answer 2 wins due to its stronger correct coverage of the Reference, clearer and more concise presentation, and faithfulness to the original description. It is more comprehensive, diverse, and empowering than Answer 1."</td></tr>
</table>

Figure 3. **Qualitative Comparison on Video QA.** Comparing VAT-KG and M^{3}KG-RAG with Qwen2.5-Omni, including retrieved knowledge and win-rate judge preferences.

<table>
<tr><td>[Query]</td><td colspan="2">Describe in detail what is visually and audibly happening in the video, including actions, objects, people, sounds, and environment. Be specific and concise.
 
</td></tr>
<tr><td>[Retrieved Subgraph]</td><td>VAT-KG
 
</td><td>M^{3}KG-RAG
 
</td></tr>
<tr><td>[Answer 1] (VAT-KG)</td><td colspan="2">"The video features a man playing an acoustic guitar. He is wearing a black hat and a black shirt. The man is seated and playing the guitar with his right hand strumming the strings and his left hand pressing the strings on the fretboard. The video is split into two parts, with the top part showing a close-up of the man's hands on the guitar and the bottom part showing a wider view of the man playing the guitar. The video also includes text on the screen that reads Natural harmonics."</td></tr>
<tr><td>[Answer 2] (M^{3}KG-RAG)</td><td colspan="2">"In the video, there is a man playing an electric guitar. He is sitting on a chair and wearing a black shirt and a black cap. The guitar is a light brown color with a dark brown pickguard. The man is using his right hand to strum the strings and his left hand to press down on the frets. There is a close-up shot of his hands on the guitar. The sound of the guitar playing is beautiful music. The room where the man is playing reverberates the music."</td></tr>
<tr><td>[Judge Response]</td><td colspan="2">"Comprehensiveness": "Winner": "Answer 2 (M^{3}KG-RAG)"
"Explanation": "Answer 2 covers more essential points from the Reference, including the man playing guitar, the instrument's color, and the beautiful music. Answer 1 incorrectly states the guitar is acoustic and adds unnecessary information about the video being split into two parts and the text 'Natural harmonics'."

"Diversity": "Winner": "Answer 2"
"Explanation": "Answer 2 provides a more varied perspective, describing both visual (man, guitar, chair) and audio (beautiful music) aspects of the Reference, while staying faithful to the original context. Answer 1 focuses mainly on the visual aspects."

"Empowerment": "Winner": "Answer 2"
"Explanation": "Answer 2 is clearer and more concise, providing a direct and actionable description of the video content. It enables understanding of the scene without unnecessary details or speculative claims."

"Overall Winner": "Winner": “Answer 2”
"Explanation": "Answer 2 wins due to its stronger, more faithful coverage of the Reference's essential points, its varied yet concise perspective, and its clearer, more actionable presentation, making it the most faithful and empowering answer."</td></tr>
</table>

Figure 4. **Qualitative Comparison on Audio-Visual QA.** Comparing VAT-KG and M^{3}KG-RAG with Qwen2.5-Omni, including retrieved knowledge and win-rate judge preferences.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[2] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 4

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017. 2

[4] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 4

[5] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 10746–10761, Suzhou, China, 2025. Association for Computational Linguistics. 4

[6] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 2

[7] Jaejun Lee, Chanyoung Chung, Hochang Lee, Sungho Jo, and Joyce Whang. Vista: Visual-textual knowledge graph representation learning. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 7314–7328, 2023. 4

[8] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 6

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 3, 5

[10] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12585–12602, 2024. 6

[11] Hyeongcheol Park, Jiyoung Seo, MinHyuk Jang, Hogun Park, Ha Dam Baek, Gyusam Chang, Hyeonsoo Im, and Sangpil Kim. Vat-kg: Knowledge-intensive multimodal knowledge graph dataset for retrieval-augmented generation. *arXiv preprint arXiv:2506.21556*, 2025. 1, 3, 4, 6

[12] Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. Videorag: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025. 4

[13] Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy Chen. Audiobench: A universal benchmark for audio large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4297–4316, 2025. 4, 6

[14] Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. Kepler: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021. 3

[15] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. 6

[16] Xuenan Xu, Ziyang Ma, Mengyue Wu, and Kai Yu. Towards weakly supervised text-to-audio grounding. *IEEE Transactions on Multimedia*, 2024. 3, 5

[17] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 3

[18] Zhiwei Zha, Jiaan Wang, Zhixu Li, Xiangru Zhu, Wei Song, and Yanghua Xiao. M2conceptbase: A fine-grained aligned concept-centric multimodal knowledge base. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3113–3123, 2024. 4