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

Hyeongcheol Park[1] Jiyoung Seo[1] Jaewon Mun[1] Hogun Park[2]
Wonmin Byeon[3] Sung June Kim[1] Hyeonsoo Im[4] JeungSub Lee[4] Sangpil Kim[1]*
[1]Korea University [2]Sungkyunkwan University [3]NVIDIA Research [4]Hanwha Systems

Abstract

Retrieval-Augmented Generation (RAG) has recently been extended to multimodal settings, connecting multimodal large language models (MLLMs) with vast corpora of external knowledge such as multimodal knowledge graphs (MMKGs). Despite their recent success, multimodal RAG in the audio-visual domain remains challenging due to 1) limited modality coverage and multi-hop connectivity of existing MMKGs, and 2) retrieval based solely on similarity in a shared multimodal embedding space, which fails to filter out off-topic or redundant knowledge. To address these limitations, we propose M^{3}KG-RAG, a Multi-hop Multimodal Knowledge Graph-enhanced RAG that retrieves query-aligned audio-visual knowledge from MMKGs, improving reasoning depth and answer faithfulness in MLLMs. Specifically, we devise a lightweight multi-agent pipeline to construct multi-hop MMKG (M^{3}KG), which contains context-enriched triplets of multimodal entities, enabling modality-wise retrieval based on input queries. Furthermore, we introduce GRASP (Grounded Retrieval And Selective Pruning), which ensures precise entity grounding to the query, evaluates answer-supporting relevance, and prunes redundant context to retain only knowledge essential for response generation. Extensive experiments across diverse multimodal benchmarks demonstrate that M^{3}KG-RAG significantly enhances MLLMs' multimodal reasoning and grounding over existing approaches. Project website: `https://kuai-lab.github.io/cvpr2026m3kgrag/`

1. Introduction

The advancements in Retrieval-Augmented Generation (RAG) have substantially improved the factual accuracy and faithfulness of large language models (LLMs) by connecting them to vast external knowledge corpora [27, 40]. Recently, graph-based RAG methods [11, 17, 56] have further pushed the progress by supporting structured reasoning and precise, query-relevant retrieval. However, extending these schemes to multimodal settings—jointly handling audio, visual, and textual signals—is non-trivial as heterogeneous inputs raise complexity, motivating designs that explicitly account for multimodal structure. Recent work [26, 39, 54] addresses these challenges with multimodal knowledge graphs (MMKGs) that organize cross-modal knowledge as entities and relations, thereby delivering query-relevant evidence to multimodal large language models (MLLMs) [1, 52].

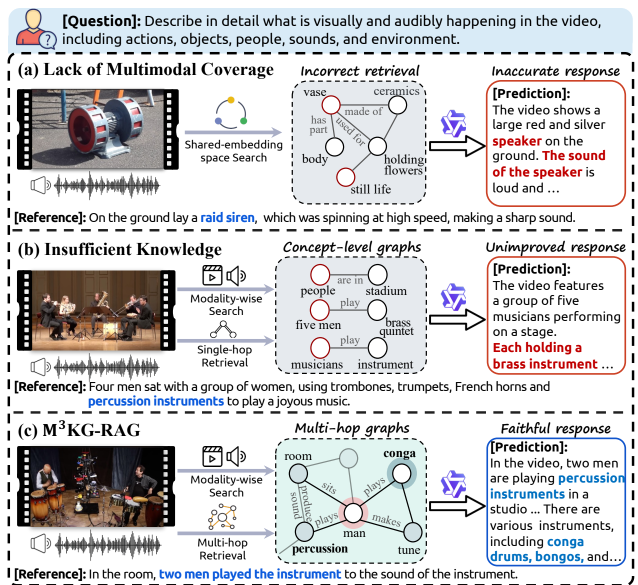

Figure 1. **Illustration of multimodal RAG scenarios.** Incorrect answers are shown in red, correct answers in blue. (a) Shared embedding search misaligns with the audio-visual query. (b) Noisy, single-hop facts provide little answer support. (c) M^{3}KG-RAG uses modality-wise multi-hop retrieval for answer-supporting context.

However, existing MMKG-enhanced RAG methods exhibit key limitations. **First**, existing MMKGs [25, 33, 54] largely emphasize image-text and provide limited audio-visual coverage, which hampers temporal and causal reasoning across modalities. Moreover, the modality gap [38] in unified multimodal embedding spaces makes cross-modal retrieval inaccurate [53]. As illustrated in Fig. 1-(a), matching an audio-visual query directly against a text-only knowledge base often fails to retrieve truly related evidence, which motivates modality-wise retrieval. While recent work [36] builds an audio-visual MMKG for precise retrieval, it in-

*Corresponding author.

duces concept-level, single-hop graphs that rarely capture temporal or causal dependencies. Therefore, constructing a multi-hop, modality-aware knowledge source across audio and visual streams is essential for query-relevant retrieval and reliable spatio-temporal reasoning.

Second, most multimodal retrieval strategies [14, 22, 36, 53] rely on similarity search in shared embedding spaces, which captures the query's broad semantics but misses fine-grained cues. They often fail to select fine-grained, query-relevant knowledge, retrieve off-topic content, and add redundant context. Even when retrieved knowledge aligns with the query, facts that do not contribute to the answer introduce noise in the MLLM context. As shown in Fig. 1-(b), such simple RAG frameworks, even when they retrieve context-matched knowledge for audio-visual queries, inject noisy evidence that fails to improve response.

To address these limitations, we propose M^{3}KG-RAG, an end-to-end, graph-enhanced RAG framework that constructs a multi-hop MMKG for modality-wise retrieval and supplies only query-aligned, answer-supportive knowledge. Specifically, we transform raw multimodal corpora into a multi-hop MMKG (M^{3}KG) with a lightweight, collaborative multi-agent pipeline in three steps. First, we perform **(i) Context-Enriched Triplet Extraction**, which captures knowledge-intensive entities and relations containing temporal and cross-modal cues. As triplets alone lack enough context for reliable reasoning [36, 54], we perform **(ii) Knowledge Grounding** to obtain canonical entity identifiers and descriptions using external resources and tools. Finally, **(iii) Context-Aware Description Refinement** aligns entity descriptions with the surrounding multimodal context to ensure consistency and specificity. In addition, we incorporate a **Self-Reflection Loop** to prevent possible hallucinated or misaligned descriptions during construction.

Additionally, we introduce *Grounded Retrieval And Selective Pruning (GRASP)* to keep only query-relevant and answer-useful subgraphs. GRASP first leverages off-the-shelf multimodal grounding models [34, 50] to drop triplets not appearing in the query, and then applies a light LLM [42] to prune triplets that do not contribute to answering the question. As shown in Fig. 1-(c), our framework retrieves knowledge tightly linked to the query from the constructed multi-hop MMKG and passes only answer-relevant evidence to the MLLMs. Extensive experiments across diverse audio, video, and audio-visual QA demonstrate that M^{3}KG-RAG achieves substantial performance gains over existing methods. Our contributions are summarized as follows:

- We present M^{3}KG-RAG, an end-to-end framework that integrates a multi-hop MMKG with RAG to enhance audio-visual reasoning in MLLMs.
- We propose a three-step, multi-agent pipeline that builds a multi-hop MMKG from raw multimodal corpora, enabling scalable, modality-wise retrieval.
- We introduce Grounded Retrieval And Selective Pruning (GRASP), which discards graph elements absent from the query or unhelpful for answering and retains only query-relevant, answer-useful subgraphs for the MLLMs.
- Through extensive evaluations across diverse multimodal benchmarks, we demonstrate that M^{3}KG-RAG consistently outperforms strong RAG baselines.

2. Related Work

2.1. Multimodal Large Language Model

Recent advances in large language models (LLMs) [1, 3, 10, 16, 42, 51] have showcased strong reasoning and generation capabilities within the language domain. This progress has extended to multimodal settings (*e.g.*, vision and audio), leading to the emergence of multimodal large language models (MLLMs). Early MLLMs, such as Flamingo [2] and BLIP-2 [30], focused on vision–language understanding through lightweight cross-modal interfaces built on top of frozen LLM backbones. Subsequent works broadened both the scale and reasoning capabilities of MLLMs. For instance, LLaVA [31] enables general-purpose image–text understanding through visual instruction tuning using a pre-trained vision encoder and LLM. Later variants [28, 29] further extend the framework to incorporate video inputs. Parallel efforts in the audio domain have produced MLLMs capable of reasoning over auditory inputs. SALMONN [41] integrates speech and general sound understanding to LLMs through specialized audio encoders. More recently, Kimi-Audio [9] achieves strong results across audio understanding, generation, and conversational tasks as an audio foundation model. Motivated by the success of both modalities, recent MLLMs focus on enhancing joint audio-visual understanding. Video-LLaMA2 [8] realizes this with a dual branch for spatial–temporal video and audio cues, improving event and scene comprehension under synchronized fusion. Qwen2.5-Omni [48] further targets real-time interaction, supporting perception and generation across multiple modalities in a streaming manner. Complementing open-source models, commercial models such as GPT-4o [20] extend multimodal I/O to low-latency audio–visual dialogue, reflecting a shift toward tightly integrated perception and reasoning.

2.2. Multimodal RAG

Retrieval-Augmented Generation (RAG) conditions LLM generation on retrieved knowledge, improving grounding and factuality [5, 12, 13, 19, 21, 27]. To support multi-hop compositional reasoning across entities and relations, graph-based RAG represents knowledge as entity–relation graphs. Along this line, GraphRAG [11] improves coherence and coverage via graph-aware indexing and community summaries, while LightRAG [17] uses dual-level retrieval for efficient, interpretable selection. HippoRAG2 [18] strengthens

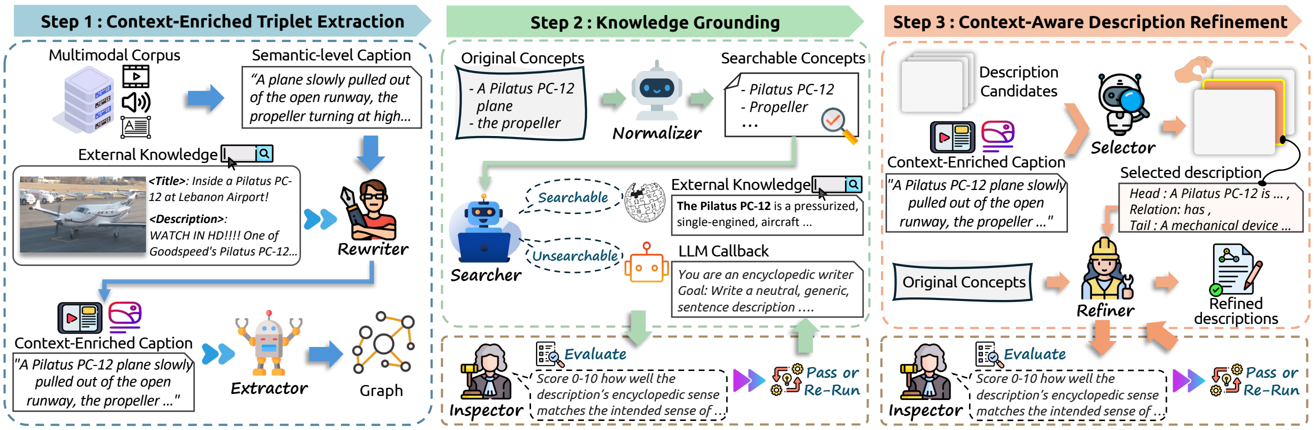

Figure 2. **An overview of the M³KG construction pipeline.** The pipeline consists of three steps: (i) **Context-Enriched Triplet Extraction**, which rewrites multimodal captions into knowledge-intensive text and extracts entity–relation triplets; (ii) **Knowledge Grounding**, linking normalized entities to open knowledge bases to obtain candidate descriptions; (iii) **Context-Aware Description Refinement**, selecting and rewriting the most context-relevant descriptions for each entity; and **Self-Reflection Loop**, where an inspector agent validates or re-runs uncertain outputs to ensure graph quality.

multi-hop retrieval with PPR-style walks and on-the-fly passage integration under tighter online LLM. With the advent of MLLMs, RAG extends beyond text to multimodal grounding, yet text-only graph RAG methods do not account for intrinsic multimodal semantics, making direct transfer difficult. In response, multimodal knowledge graph (MMKG) based approaches have emerged, encoding entities and relations across modalities. MR-MKG [26] leverages MMKG with graph encoding and cross-modal alignment to improve multimodal reasoning. M²ConceptBase [54] organizes image–text corpora into a concept-centric MMKG and couples it with graph-aware retrieval, improving grounding and answer accuracy for multimodal QA and description tasks. Pushing beyond image–text, VAT-KG [36] integrates visual, audio, and text into an MMKG and proposes a RAG protocol tailored to audio–visual MLLMs, improving faithfulness under audio–visual queries. However, its graph construction is largely single-hop, and its RAG framework mainly relies on similarity-based search, which can admit off-topic neighbors and redundant context. In contrast, we build a multi-hop MMKG and execute fine-grained, query-conditioned retrieval, supplying answer-useful context.

3. Method

In this section, we detail the M³KG-RAG paradigm, emphasizing its core architecture and contributions. Sec. 3.1 presents our multi-hop multimodal knowledge graph (M³KG) construction pipeline. Sec. 3.2 introduces our multimodal RAG framework equipped with the proposed GRASP.

3.1. M³KG Construction with Multi Agents

Our M³KG-RAG improves the retrieval scheme by constructing a multi-hop MMKG that enables scalable, in-depth reasoning over multiple multimodal triplets. However, since constructing a multi-hop MMKG requires multiple stages beyond simple graph connectivity, we design a lightweight, collaborative multi-agent pipeline with our own specialized LLM agents—*rewriter*, *extractor*, *normalizer*, *searcher*, *selector*, *refiner*, and *inspector*—to balance automation and quality control, in line with recent advances in multi-agent LLM systems and self-reflection [4, 43]. The overall construction process is illustrated in Fig. 2, and the detailed role and method for each step are as follows.

Step 1: Context-Enriched Triplet Extraction Our multi-hop MMKG construction starts from a raw multimodal corpus $\mathcal{C} = \{(x_n^{text}, x_n^{audio}, x_n^{visual})\}_{n=1}^N$ of N aligned samples, where $x_n^{text}, x_n^{audio}, x_n^{visual}$ denote the text, audio, and visual data for sample n. Much of the text in $\mathcal{C}$ is semantically generic [7, 24], which limits its utility as external knowledge for MLLMs. Following prior MMKG construction studies [36], we develop a new *rewriter* that converts semantic-level caption x_n^{text} into a context-enriched caption $\tilde{x}_n^{text}$ by incorporating external knowledge—titles and descriptions collected via a crawler—to supply knowledge-intensive context. Motivated by recent advances in open information extraction leveraging LLMs [55, 57], we further introduce an *extractor* that parses $\tilde{x}_n^{text}$ and returns triplets $\mathcal{T}_n = \{(h_i, r_i, t_i)\}_{i=1}^{K_n}$, where h_i, r_i, and t_i denote the head entity, relation, and tail entity and $K_n = |\mathcal{T}_n|$ varies by input. As the rewritten $\tilde{x}_n^{text}$ is knowledge-intensive and summarizes the overall multimodal context, the extracted triplets often capture relations among long-tail or uncommon entities—cases that MLLMs commonly miss or misidentify.

Step 2: Knowledge Grounding While transforming the corpus into a graph structure enables efficient entity-level access, connections alone offer limited guidance to MLLMs.

To enrich the MMKG beyond connectivity, we ground encyclopedic descriptions to entities in this step. Head and tail entities in $\mathcal{T}$ often include modifiers or variant surface forms that hinder look-up (*e.g.*, "small brown dog" vs. "dog"). Thus, the *normalizer* first maps each entity mention to a canonical, searchable concept by removing non-essential modifiers, preserving the source word order when appropriate, and standardizing to a singular noun phrase. Given the normalized concepts, the *searcher* queries open knowledge bases (*e.g.*, Wikipedia, Wiktionary) and uses a crawler to retrieve a compact set of candidate descriptions for each entity. Since open knowledge bases cannot cover every textual concept, we include a lightweight LLM callback to fill missing descriptions. Subsequently, we have a candidate description set $\mathcal{D}$ for every normalized concept.

Step 3: Context-Aware Description Refinement A single term can carry multiple meanings, (*e.g.*, "bank": financial institution vs. a river bank). To make entities accurately informative, the *selector* chooses the most context-appropriate description from the candidate set, using the context-enriched caption from **Step 1** as guidance for each normalized concept. This keeps descriptions aligned with the context and filters out off-topic ones. Since the selected descriptions are written for a normalized concept rather than the original heads and tails in $\mathcal{T}$, the *refiner* adapts the chosen description to the original concept's phrasing to inject the original semantics while preserving the selected content. After this step, we obtain a refined description set $\hat{\mathcal{D}}$ for all entities.

Self-Reflection Loop To ensure knowledge graph quality, we introduce *Inspector* that implements a self-reflection loop within our construction pipeline. When the task extends beyond simple information extractions and instead relies on the language model's implicit knowledge (*e.g.*, Step 2 LLM Callback or Step 3 Rewriter), errors may occur. Accordingly, the inspector reviews these outputs and either passes them or returns a re-run signal to the producing agent.

Resulting M^{3}KG Starting from the text data x_n^{text} in multimodal corpus $\mathcal{C}$, our multi-agent pipeline with a self-reflection loop constructs a multi-hop knowledge graph and links its triplets to the corresponding audio-visual data $(x_n^{audio}, x_n^{visual})$, yielding the following multi-hop multimodal knowledge graph:

$$\mathcal{G} = \{\mathcal{E}, \mathcal{R}, \mathcal{T}, \hat{\mathcal{D}}, \mathcal{A}, \mathcal{V}, \mathcal{L}\}, \tag{1}$$

where $\mathcal{E}$ is the set of entities; $\mathcal{R}$ the set of relations; $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ the set of triplets; $\hat{\mathcal{D}} = \{\, d_e \,\}_{e \in \mathcal{E}}$ the per-entity refined descriptions; $\mathcal{A} = \{x_n^{\text{audio}}\}_{n=1}^{N}$ and $\mathcal{V} = \{x_n^{\text{visual}}\}_{n=1}^{N}$ the audio/visual items; $\mathcal{L} \subseteq \mathcal{T} \times (\mathcal{A} \cup \mathcal{V})$ the links from triplets to associated audio/visual data. The resulting multi-hop MMKG satisfies the following coverage property:

$$\forall\, \mathsf{t} \in \mathcal{T},\ \exists\, x \in (\mathcal{A} \cup \mathcal{V}) \text{ s.t. } (\mathsf{t}, x) \in \mathcal{L}. \tag{2}$$

Consequently, since every triplet links to at least one multimodal item, all facts are eligible for retrieval under multimodal queries, providing full graph coverage.

3.2. Multimodal RAG Framework

To deliver query-relevant and answer-useful context only, we design a multimodal RAG framework composed of (i) Modality-Wise Retrieval over the MMKG to gather candidates aligned with the input modalities, and (ii) GRASP to keep only knowledge that is relevant to the multimodal query and useful for answering the question. An overview of the multimodal RAG framework is shown in Fig. 3.

Modality-Wise Retrieval Shared embedding spaces from multimodal encoders often exhibit a modality gap where cross-modal distances are not comparably calibrated [53]. Consequently, querying a knowledge base indexed in a different modality—for instance, a video query against text embeddings—often yields off-topic neighbors. To bridge the modality gap, we find the items of the same modality as the query in $\mathcal{G}$ and lift them to triplets. This procedure is enabled by Eq. 2, which guarantees that each triplet in $\mathcal{G}$ is linked to at least one audio or visual item. Concretely, we obtain query embeddings with multimodal foundation models (*e.g.*, InternVL2 [46] for video and CLAP [47] for audio) and search a FAISS [23] index built over $\mathcal{G}$'s audio/visual items using L2 distance in the embedding space. We first retrieve the top-k nearest items and then keep only candidates within a distance threshold τ of the query to avoid off-topic neighbors. When both audio and video are provided, we form a simple vector concatenation and apply the same search. Let $\mathcal{S} \subseteq (\mathcal{A} \cup \mathcal{V})$ denote the set of audio/visual items selected by the above retrieval, we then lift items to triplets to obtain the query-relevant initial graph:

$$\mathcal{G}_{init} = \{\, \mathsf{t} \in \mathcal{T} \mid \exists\, x \in \mathcal{S},\ (\mathsf{t}, x) \in \mathcal{L} \,\}. \tag{3}$$

GRASP (Grounded Retrieval And Selective Pruning) After mitigating the modality gap, we obtain a query-aligned initial graph $\mathcal{G}_{init}$. Yet, its similarity-only retrieval can lack fine-grained alignment or include knowledge that may not be useful to answer the question. For instance, if the question asks *"What instrument is being played?"*, only knowledge about the instruments present is useful, and the remaining triplets mostly become noise for the MLLMs. To address these limitations, we design a Grounded Retrieval And Selective Pruning (GRASP) to align the graph finely with the query and retain only answer-useful knowledge. Specifically, we first use off-the-shelf multimodal grounding models to verify whether entities or triplets in $\mathcal{G}_{init}$ appear in the query's audio and/or visual streams. For visual grounding, we use GroundingDINO [34] on four uniformly sampled frames F from the query video q_v, obtaining per-frame detection confidences $\Phi_v(e; f)$ for each entity $e \in \mathcal{G}_{init}$ on

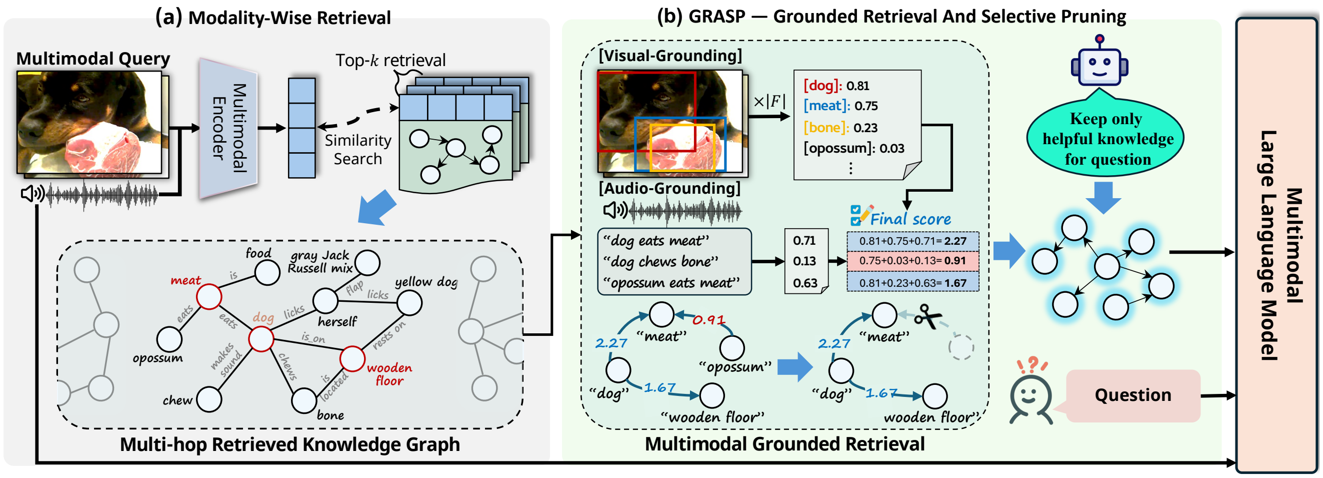

Figure 3. **Overview of the Multimodal RAG framework.** The framework consists of two components: **(a) Modality-Wise Retrieval**, which retrieves multi-hop triplets aligned with the query from the M^{3}KG; and **(b) GRASP (Grounded Retrieval And Selective Pruning)**, which uses visual and/or audio grounding models to check entity presence and prunes triplets that are off-topic or non-informative. The resulting subgraph is then provided to an MLLM for query-relevant, evidence-grounded audio-visual reasoning.

each frame f. We then take the maximum across the sampled frames as the visual presence score of each entity:

$$s_v(e \mid q_v) \;=\; \max_{f \in F} \Phi_v(e; f). \tag{4}$$

We derive a triplet-level visual presence score by summing the presence scores of the head and tail entities and prune triplets whose score falls below η_v.

For audio grounding, we use a Text-to-Audio Grounding model (TAG) [50]. Unlike visual signals, audio is not as easily factorized into independent snapshots. Accordingly, we convert each triplet $\mathfrak{t}$ into a natural sentence $\sigma(\mathfrak{t})$ (*e.g.*, "h r t") and use the TAG scoring function Φ_a to measure how strongly $\sigma(\mathfrak{t})$ is grounded in the query audio q_a, resulting in the audio presence score:

$$s_a(\mathfrak{t} \mid q_a) = \Phi_a(\sigma(\mathfrak{t}); q_a). \tag{5}$$

Similar to the visual case, we drop triplets whose audio presence score falls below η_a. When both audio and visual streams are available, we simply sum their presence scores and remove triplets whose fused score is below η_{av}. This grounding step yields the grounded subgraph $\mathcal{G}_{grd}$.

After obtaining the grounded subgraph $\mathcal{G}_{grd}$, we remove knowledge that is not helpful for answering the question. A lightweight LLM [49] produces a binary mask over triplets under a conservative keep-or-drop policy, yielding $\mathcal{G}_{\text{GRASP}}$. This procedure filters unhelpful and off-topic triplets while keeping answer-supportive knowledge.

Graph-Augmented Generation Our M^{3}KG-RAG targets MLLMs that jointly reason over video, audio, and text. Given the retrieved subgraph $\mathcal{G}_{\text{GRASP}}$, we condition the MLLM by concatenating the multimodal query q with the graph context. For each triplet $(h, r, t) \in \mathcal{G}_{\text{GRASP}}$, we include the relation r together with the entity–description pairs $\langle h, d_h \rangle$ and $\langle t, d_t \rangle$, where d_h and d_t are the refined descriptions of h and t, respectively. Formally, our graph-enhanced generation is defined as follows:

$$p_{\text{aug}} = q \parallel \left(\bigcup_{(h,r,t) \in \mathcal{G}_{\text{GRASP}}} \langle h, d_h \rangle \xrightarrow{r} \langle t, d_t \rangle \right). \tag{6}$$

By providing p_{aug} to the MLLMs, we inject query-relevant, answer-useful knowledge and supply the inter-entity relations together with detailed entity attributions, which improves the MLLM's reasoning capabilities.

4. Experiments

4.1. Experimental Setup

Datasets To highlight the diverse modality coverage of M^{3}KG-RAG, we evaluate on three multimodal tasks: Audio QA, Video QA, and Audio-Visual QA. For Audio-QA, we use AudioCaps-QA [44], which provides human-annotated QA pairs built on AudioCaps corpus [24]. For Video-QA, we adopt the VideoChatGPT (VCGPT) benchmark [35], which is built on videos curated from ActivityNet [6]. For Audio-Visual QA, we evaluate on the VALOR [32] benchmark, which explicitly requires joint reasoning over synchronized audio and visual streams for response.

MLLMs We apply our M^{3}KG-RAG to three MLLMs capable of joint audio-visual understanding to assess its effectiveness. Specifically, we employ VideoLLaMA2 [8] and Qwen2.5-Omni [48], both advanced open-source models that jointly process audio and visual streams and serve as strong baselines for audio–visual reasoning. We further adopt GPT-4o [20], a strong commercial model with substantial implicit

MLLM	Method	Audio QA	Video QA	Audio-Visual QA
		AudioCaps-QA	VCGPT	VALOR
VideoLLaMA2	None	43.13	39.09	25.66
	Wikidata	43.58	38.58	26.43
	VTKG	43.02	38.88	25.92
	M^2ConceptBase	42.19	39.31	25.93
	VAT-KG	44.60	39.42	28.30
	M^3KG-RAG	**53.23**	**39.92**	**29.25**
Qwen2.5-Omni	None	49.00	42.21	32.42
	Wikidata	49.78	40.82	30.28
	VTKG	48.95	42.96	32.70
	M^2ConceptBase	49.78	42.78	32.31
	VAT-KG	51.30	43.50	35.44
	M^3KG-RAG	**60.77**	**44.35**	**44.67**

Table 1. **Overall performance.** We report Model-as-Judge (M.J.) scores (higher is better). Across all benchmarks and for both MLLMs, M^3KG-RAG provides the largest and most consistent gains over the no-retrieval and MMKG-based baselines. The best results are **bolded**.

knowledge and capacity, to examine whether our method remains effective even for such high-capacity models.

Baseline Methods Following prior multimodal RAG work [36], we compare M^3KG-RAG against five baselines: (i) None—the MLLM answers without external knowledge; (ii) Wikidata [45] + naïve RAG with retrieval in a shared embedding space [46, 47] between the text KG and the multimodal query; (iii) VTKG [25] + naïve RAG (image–text MMKG), matching visual queries to images in MMKG via a vision–language space (*e.g.*, CLIP [37]) and audio queries in the CLAP [47] audio–text space; (iv) M^2ConceptBase [54], following VTKG's protocol; and (v) VAT-KG [36], evaluated under its released RAG protocol that accounts for audio–visual streams.

Implementation Details We implement the multi-hop MMKG construction pipeline with a lightweight multi-agent stack built on a single backbone LLM (Qwen3-8B) [49], using only the training splits of our evaluation benchmarks. In modality-wise retrieval, we set $k = 5$ and select the top-k best-matching items per query, then expand each to its connected multi-hop subgraph. Since our benchmarks span different audio–visual distributions, we set the modality-wise distance threshold τ and GRASP presence threshold η separately per benchmark as follows: for AudioCaps-QA, $\tau = 3.0$ and $\eta_a = 0.5$; for VCGPT, $\tau = 0.15$ and $\eta_v = 1.5$; and for VALOR, $\tau = 4.5$ and $\eta_{av} = 1.2$. These values are held constant within each benchmark across all experiments. All experiments use a single NVIDIA H100 GPU. Note that further details are provided in the supplementary material.

Evaluation Metrics We implement two evaluation schemes. First, since our benchmarks consist of open-ended QA with free-form responses, we adopt an off-the-shelf Model-as-Judge (M.J.) protocol [44], where an LLM judge [15] scores each response. Second, in line with established RAG evaluation [11, 17, 39], we report a win-rate preference protocol where the LLM judge compares the two responses (ours vs. a baseline) and selects the preferred response based on multiple criteria. We make it reference-aware by providing the judge with the reference answer during comparison, which reduces verbosity bias and yields a more reliable, multi-dimensional assessment.

4.2. Quantitative Results

We compare M^3KG-RAG against text-KG and multimodal-KG baselines on Audio-QA, Video-QA, and Audio-Visual QA. The overall results are summarized in Table 1. Across all benchmarks, M^3KG-RAG yields significant gains over the base MLLMs, indicating that modality-wise retrieval and GRASP deliver knowledge that is both tightly aligned with the query and directly useful for answering. In contrast, other baselines tend not to consistently improve the MLLMs. Specifically, text KG with naïve RAG (Wikidata) yields weak or even negative deltas, as retrieval ignores the temporal nature of audio–visual queries, often retrieving off-context neighbors and injecting noisy facts that do not support the answer. Image-text KGs with naïve RAG (VTKG, M^2ConceptBase) partially account for visual cues via images in MMKGs but still miss query dynamics, leading to limited impact on response quality and occasional degradation. VAT-KG, which considers audio–visual streams, improves all baselines uniformly. However, its largely single-hop MMKG captures only local, concept-level facts. The MLLM therefore receives only shallow, fragmentary context, so the knowledge implicitly encoded in the underlying multimodal data is only partially exploited, and performance gains remain mostly marginal. In contrast, M^3KG-RAG builds multi-hop neighborhoods that aggregate temporally and se-

	AudioCaps-QA		VCGPT		VALOR	
	Baseline	**Ours**	Baseline	**Ours**	Baseline	**Ours**
Baseline: None						
Comprehensiveness	15.9%	**84.1%**	47.6%	**52.4%**	39.8%	**60.2%**
Diversity	20.3%	**79.7%**	37.8%	**62.2%**	45.5%	**54.5%**
Empowerment	14.0%	**86.0%**	42.1%	**57.9%**	40.1%	**59.9%**
Overall	15.2%	**84.8%**	47.0%	**53.0%**	39.8%	**60.2%**
Baseline: Wikidata						
Comprehensiveness	14.9%	**85.1%**	48.3%	**51.7%**	40.3%	**59.7%**
Diversity	22.4%	**77.6%**	47.4%	**52.6%**	**55.5%**	44.5%
Empowerment	12.0%	**88.0%**	39.6%	**60.4%**	40.8%	**59.2%**
Overall	13.7%	**86.3%**	44.5%	**55.5%**	40.8%	**59.2%**
Baseline: VTKG						
Comprehensiveness	20.8%	**79.2%**	49.1%	**50.9%**	39.1%	**60.9%**
Diversity	33.8%	**66.2%**	45.9%	**54.1%**	45.2%	**54.8%**
Empowerment	21.2%	**78.8%**	46.6%	**53.4%**	39.2%	**60.8%**
Overall	21.2%	**78.8%**	49.1%	**50.9%**	39.4%	**60.6%**
Baseline: M^2ConceptBase						
Comprehensiveness	21.2%	**78.8%**	41.8%	**58.2%**	38.2%	**61.8%**
Diversity	28.3%	**71.7%**	43.9%	**56.1%**	45.1%	**54.9%**
Empowerment	19.7%	**80.3%**	44.6%	**55.4%**	38.6%	**61.4%**
Overall	21.0%	**79.0%**	44.3%	**55.7%**	38.3%	**61.7%**
Baseline: VAT-KG						
Comprehensiveness	26.1%	**73.9%**	48.4%	**51.6%**	41.4%	**58.6%**
Diversity	34.8%	**65.2%**	46.6%	**53.4%**	48.3%	**51.7%**
Empowerment	24.3%	**75.7%**	43.5%	**56.5%**	42.1%	**57.9%**
Overall	25.6%	**74.4%**	47.6%	**52.4%**	41.8%	**58.2%**

Table 2. **Win-rate comparison.** Pairwise win rates (%) of each baseline versus M^{3}KG-RAG across three benchmarks and four criteria. Columns show the preference rate of the *Baseline* and *Ours*, with the higher win rate in each pair highlighted in **bold**.

mantically related evidence across modalities and, together with modality-wise retrieval and GRASP, delivers query-focused, answer-supporting knowledge that more faithfully reflects the multimodal query. Consequently, M^{3}KG-RAG achieves notable gains over VAT-KG on every benchmark. These findings become more pronounced with a stronger commercial MLLM. As shown in Table 3, even with substantial built-in knowledge, GPT-4o paired with M^{3}KG-RAG improves across all benchmarks and exhibits larger gains than with VAT-KG, reinforcing that multi-hop evidence together with GRASP provides a diverse, answer-supporting context that the model can exploit more effectively.

Win-rate preference results in Table 2 corroborate the M.J. scores, showing consistent preference for M^{3}KG-RAG over baselines across benchmarks and criteria. We observe higher **Comprehensiveness** because richer multi-hop evidence aggregates the key entities and relations needed to answer the query end-to-end, aided by refined entity descriptions for clarity. **Diversity** improves as the multi-hop MMKG offers several distinct evidence chains, while pruning removes off-topic or duplicate content. **Empowerment** benefits from a strictly query-relevant context that reduces hallucination and steers the model toward concrete, answer-supporting details rather than generic filler. Together, these effects yield stronger **Overall** preferences in pairwise comparisons.

MLLM	Method	AudioCaps-QA	VCGPT	VALOR
GPT-4o	None	56.74	49.68	46.02
GPT-4o	VAT-KG	57.70	51.49	55.86
GPT-4o	M^{3}KG-RAG	**59.17**	**53.05**	**56.53**

Table 3. **Performance on Commercial MLLM (GPT-4o).** We report M.J. scores (higher is better). The best result is **bolded**.

MLLM	Method		M.J↑
	Modality-Wise Retrieval	GRASP	
Qwen2.5-Omni	✗	✗	36.62
	✓	✗	40.91
	✗	✓	36.96
	✓	✓	**44.67**

Table 4. **Ablation on VALOR.** Checkmarks denote enabled components. We report the M.J. score; combining Modality-Wise Retrieval and GRASP gives the best score. The best result is **bolded**.

4.3. Ablation Study

To explore the effectiveness of our design, we conduct ablations of modality-wise retrieval and GRASP on the VALOR, which requires joint audio–visual reasoning, using Qwen2.5-Omni [48] as the base MLLM. For modality-wise retrieval, we remove the cross-modal links $\mathcal{L}$ that connect multimodal items ($\mathcal{A}$, $\mathcal{V}$) to the triplet set $\mathcal{T}$ in the M^{3}KG, yielding a text-only KG. We then convert each triplet t into a natural sentence $\sigma(t)$ and index them using the text encoder of the multimodal embedding model [46, 47], enabling retrieval for audio-visual queries in a shared embedding space.

As shown in Table 4, using modality-wise retrieval alone keeps retrieval within the query's modality and reduces mismatched evidence. However, relying solely on similarity search cannot verify entity-level relevance to the query or ensure that retrieved evidence supports the answer, leading to limited performance gains. Using GRASP improves faithfulness via fine-grained pruning, yet the initial graph retrieved from a text-only KG in a shared space is weakly aligned with the audio–visual cues, yielding only modest improvements.

Combining both offers the largest gain: modality-wise retrieval supplies candidates aligned with the query's audio and visual streams, and GRASP retains only triplets that directly support the question and removes redundancy. Taken together, modality-wise retrieval and grounded pruning are complementary and jointly necessary for multimodal RAG.

4.4. Qualitative Results

Fig. 4 presents qualitative results across Audio, Video, and Audio–Visual QA. With M^{3}KG-RAG, the MLLM produces more specific, context-faithful answers by grounding generation in multi-hop evidence and concise entity descriptions from our modality-wise retrieval and GRASP.

In the case of Audio-QA, the context supplied by M^{3}KG-RAG directly supports answering. For the

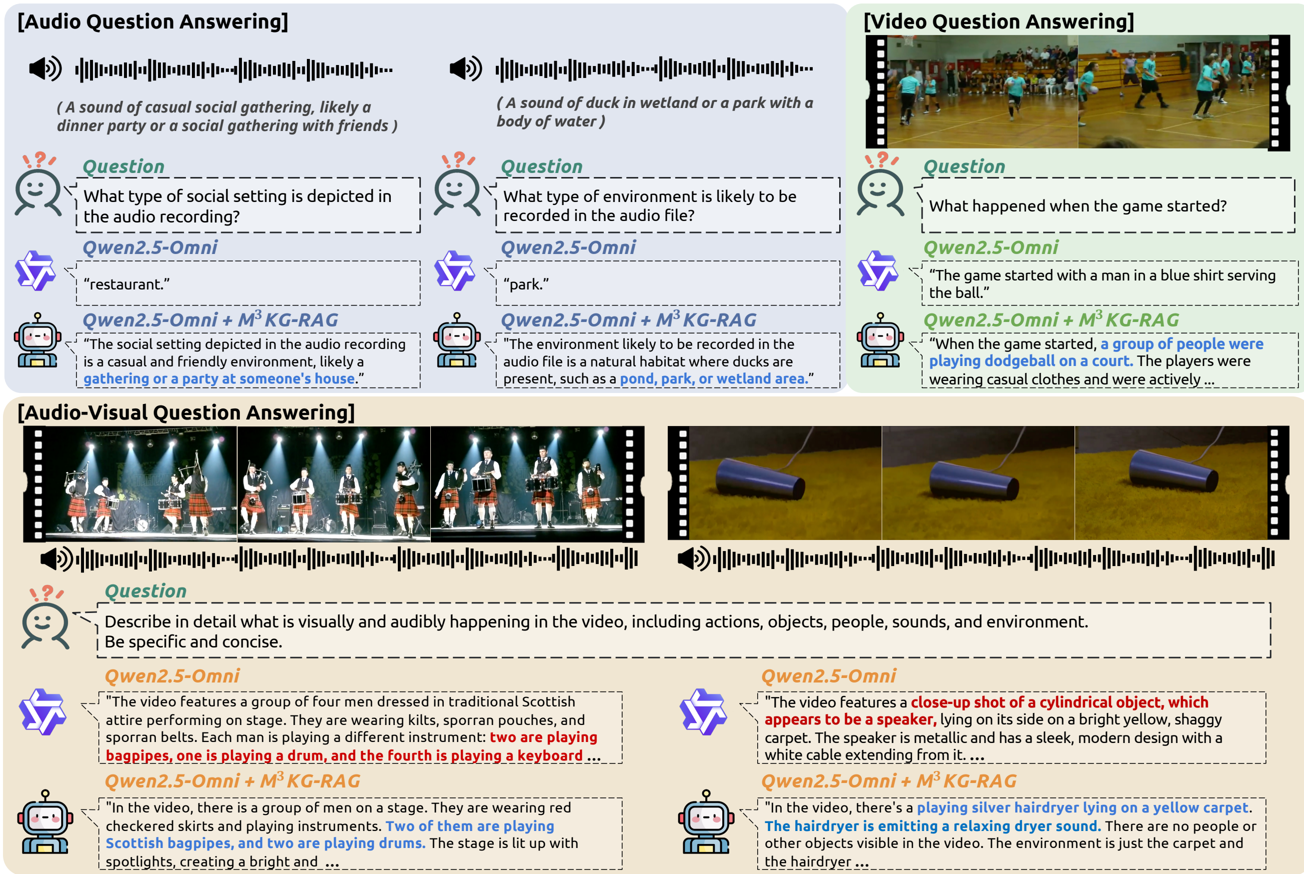

Figure 4. **Qualitative results on various Question Answering tasks.** Incorrect and insufficient model responses are highlighted in red, while correct and sufficient responses are highlighted in blue.

question "What type of social setting is depicted in the audio recording?" the base model responds "restaurant," which is loosely related yet misaligned with the asked social setting and lacks sufficient detail. With M^{3}KG-RAG, query-conditioned retrieval selects the correct context, and GRASP passes only answer-useful cues, producing "a gathering or a party at someone's house." Likewise, for the environment clip with water, the base MLLM overlooks water-related acoustic attributes, whereas our retrieval highlights duck calls and water ambience, producing "a natural habitat with ducks, such as a pond, park, or wetland."

In the case of Video-QA, the graph-enhanced context enables precise answering. While the base model misses the action occurring in the video, our RAG conditions retrieval on the visual stream, identifies the scene as dodgeball, and supplies action-aware context that enables a precise answer.

Retrieval conditioned on both audio and visual sharpens predictions in Audio-Visual QA. For the stage-performance clip, the model hallucinates a keyboard player. With M^{3}KG-RAG, based on the retrieved query-related context, the model produces correct stage context and instrument roles (two bagpipes, two drums). For the hair-dryer clip, the model tags a "cylinder object" and misclassifies it as a speaker. On the other hand, M^{3}KG-RAG conditions retrieval on the audio-visual query and passes context that links the buzzing audio to a running hair dryer, enabling a precise response. These results demonstrate that modality-conditioned retrieval with GRASP supplies on-topic, fine-grained context that improves the specificity and faithfulness of answers.

5. Conclusion

We introduce M^{3}KG-RAG, a novel graph-augmented multimodal RAG framework for enhancing audio–visual reasoning in MLLMs. Our lightweight multi-agent pipeline constructs a multi-hop multimodal knowledge graph that supports precise modality-wise retrieval and robust knowledge grounding. Furthermore, the proposed GRASP (Grounded Retrieval And Selective Pruning) scores triplets with visual and audio foundation models and retains only query-relevant, answer-supporting evidence. Extensive evaluation across diverse multimodal benchmarks highlights the effectiveness of M^{3}KG-RAG, with consistent performance gains over strong baselines. We believe M^{3}KG-RAG will serve as a practical foundation for future research in multimodal RAG.

Acknowledgement

This work was supported by Korea Research Institute for Defense Technology Planning and Advancement (KRIT) - Grant funded by the Defense Acquisition Program Administration (DAPA) (KRIT-CT-23-021, 97%), Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism (International Collaborative Research and Global Talent Development for the Development of Copyright Management and Protection Technologies for Generative AI, RS-2024-00345025, 1%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(RS-2025-00521602, 1%), Institute of Information & communications Technology Planning & Evaluation (IITP) & ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT) (No.RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 1%), and the NVIDIA Academic Grant Program using NVIDIA RTX PRO 6000 Max-Q Workstation Edition.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 2

[3] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 2

[4] Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. Reflective multi-agent collaboration based on large language models. *Advances in Neural Information Processing Systems*, 37:138595–138631, 2024. 3

[5] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR, 2022. 2

[6] 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. 5

[7] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020. 3

[8] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 2, 5

[9] Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025. 2

[10] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407, 2024. 2

[11] 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. 1, 2, 6

[12] Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501, 2024. 2

[13] Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, 2023. 2

[14] Sreyan Ghosh, Sonal Kumar, Chandra Kiran Reddy Evuru, Ramani Duraiswami, and Dinesh Manocha. Recap: Retrieval-augmented audio captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1161–1165. IEEE, 2024. 2

[15] 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. 6

[16] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2

[17] 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. 1, 2, 6

[18] Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Forty-second International Conference on Machine Learning*, 2025. 2

[19] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3929–3938, 2020. 2

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 5

[21] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020. 2

[22] Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21278–21298, Vienna, Austria, 2025. Association for Computational Linguistics. 2

[23] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 4

[24] 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. 3, 5

[25] 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. 1, 6

[26] Junlin Lee, Yequan Wang, Jing Li, and Min Zhang. Multimodal reasoning with multimodal knowledge graph. *arXiv preprint arXiv:2406.02030*, 2024. 1, 3

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 1, 2

[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2

[29] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun MA, and Chunyuan Li. LLaVA-neXT-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742, 2023. 2

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 34892–34916, 2023. 2

[32] 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. 5

[33] Junming Liu, Siyuan Meng, Yanting Gao, Song Mao, Pinlong Cai, Guohang Yan, Yirong Chen, Zilin Bian, Ding Wang, and Botian Shi. Aligning vision to language: Annotation-free multimodal knowledge graph construction for enhanced llms reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 981–992, 2025. 1

[34] 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. 2, 4

[35] 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. 5

[36] 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, 2, 3, 6

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[38] Sameera Ramasinghe, Violetta Shevchenko, Gil Avraham, and Ajanthan Thalaiyasingam. Accept the modality gap: An exploration in the hyperbolic space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27263–27272, 2024. 1

[39] 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. 1, 6

[40] Chaitanya Sharma. Retrieval-augmented generation: A comprehensive survey of architectures, enhancements, and robustness frontiers. *arXiv preprint arXiv:2506.00054*, 2025. 1

[41] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023. 2

[42] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024. 2

[43] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O'Sullivan, and Hoang D Nguyen. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*, 2025. 3

[44] 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. 5, 6

[45] 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. 6

[46] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 4, 6, 7

[47] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 4, 6, 7

[48] 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. 2, 5, 7

[49] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. 5, 6

[50] Xuenan Xu, Ziyang Ma, Mengyue Wu, and Kai Yu. Towards weakly supervised text-to-audio grounding. *IEEE Transactions on Multimedia*, 2024. 2, 5

[51] 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. 2

[52] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 1

[53] Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universalrag: Retrieval-augmented generation over corpora of diverse modalities and granularities. *arXiv preprint arXiv:2504.20734*, 2025. 1, 2, 4

[54] 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. 1, 2, 3, 6

[55] Bowen Zhang and Harold Soh. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9820–9836, 2024. 3

[56] Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Hao Chen, Yilin Xiao, Chuang Zhou, Junnan Dong, et al. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*, 2025. 1

[57] Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. A survey of generative information extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, 2025. 3