UniversalRAG: Retrieval-Augmented Generation over Corpora of Diverse Modalities and Granularities

Woongyeong Yeo[1*] **Kangsan Kim**[1*] **Soyeong Jeong**[1] **Jinheon Baek**[1†] **Sung Ju Hwang**[1,2†]
[1]KAIST [2]DeepAuto.ai
{wgcyeo, kangsan.kim, starsuzi, jinheon.baek, sungju.hwang}@kaist.ac.kr
https://universalrag.github.io

Abstract

Retrieval-Augmented Generation (RAG) has shown substantial promise in improving factual accuracy by grounding model responses with external knowledge relevant to queries. However, most existing approaches are limited to a text-only corpus, and while recent efforts have extended RAG to other modalities such as images and videos, they typically operate over a single modality-specific corpus. In contrast, real-world queries vary widely in the type of knowledge they require, which a single type of knowledge source cannot address. To address this, we introduce UniversalRAG, an any-to-any RAG framework designed to retrieve and integrate knowledge from heterogeneous sources with diverse modalities and granularities. Specifically, motivated by the observation that forcing all modalities into a unified representation space derived from a single aggregated corpus causes a modality gap, where the retrieval tends to favor items from the same modality as the query, we propose modality-aware routing, which dynamically identifies the most appropriate modality-specific corpus and performs targeted retrieval within it, and further justify its effectiveness with a theoretical analysis. Moreover, beyond modality, we organize each modality into multiple granularity levels, enabling fine-tuned retrieval tailored to the complexity and scope of the query. We validate UniversalRAG on 10 benchmarks of multiple modalities, showing its superiority over various modality-specific and unified baselines.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable performance across various tasks, and have been widely adopted to assist users in everyday life (Gemini Team, 2023; OpenAI, 2025). However, LLMs often generate factually incorrect or misleading information, especially on topics they were less or not exposed to during training (Zhang et al., 2025d; Huang et al., 2025). To address this, Retrieval-Augmented Generation (RAG) has emerged as a promising approach, which allows the model responses to be grounded in the query-relevant knowledge retrieved from external knowledge sources, enhancing factual accuracy (Lewis et al., 2020; Gao et al., 2023; Chen et al., 2024a).

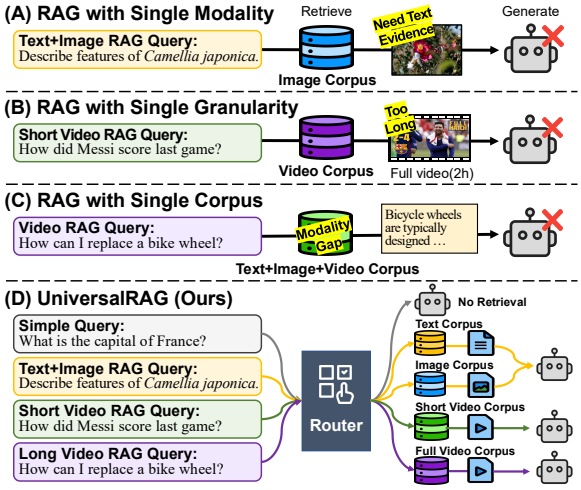


Figure 1: Conceptual illustration comparing existing RAG strategies with our proposed UniversalRAG.

Despite its effectiveness, existing approaches are typically designed for a single corpus and modality, limiting their ability to address queries that require diverse knowledge sources. In practice, as shown in Figure 1, user queries vary widely in the type of knowledge they require: some are best answered using text (e.g., surface-level facts), others demand visual understanding from images or videos (spatial or temporal cues), and still others require combinations of these modalities. Yet, the field of RAG primarily originates with a textual corpus (Lewis et al., 2020; Jiang et al., 2023; Yan et al., 2024), and although recent efforts have expanded it to modalities beyond text (such as images and videos) (Abootorabi et al., 2025; Jeong et al., 2025; Shalev-Arkushin et al., 2026), existing RAG methods (when considered individually) are typically

[*]Equal contribution; [†]Equal advising

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3843–3871
July 2-7, 2026 ©2026 Association for Computational Linguistics

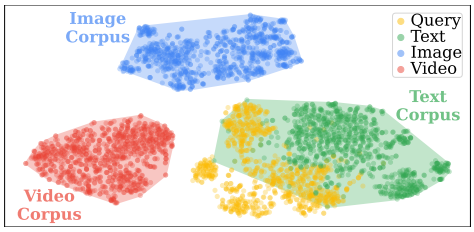


Figure 2: t-SNE plot of the unified embedding space.

modality- and corpus-specific; therefore, they may be suboptimal to serve as a universal, one-for-all framework that can flexibly handle the wide range of queries, whose knowledge requirements vary.

In this work, we present UniversalRAG, a novel any-to-any RAG framework that brings together knowledge distributed across multiple modality-specific corpora, and leverages them to generate grounded responses to queries in a universal workflow. To operationalize this, one straightforward approach might be to aggregate all entries from the collected, heterogeneous knowledge corpora, and embed them into a unified space using a multimodal encoder (which is typically trained to align inputs from different modalities if they are semantically similar). However, despite such alignment efforts, we find that this strategy suffers from modality gaps (Liang et al., 2022; Meng et al., 2026), the tendency that inputs are clustered based on their modality rather than their semantic relevance (visualized in Figures 2 and 7). As a result, retrieval becomes biased toward knowledge sources that share the same modality as the query, overlooking relevant content from other modalities.

To address this challenge, rather than forcing all modalities into a single embedding space, we take a different direction and introduce *modality-aware routing*. UniversalRAG predicts its modality requirements and routes retrieval to the corresponding modality-specific corpora (potentially multiple, when the query calls for cross-modal evidence), after which the retrieved knowledge is jointly used for grounding. Notably, this strategy not only sidesteps modality gaps by avoiding every cross-modal comparison but also enables seamless integration of new modalities by extending the routing logic without modifying existing modality-specific retrievers.

Beyond modality, data granularity (i.e., the size or unit of each entry in the corpus) also affects retrieval precision and generation quality (Chen et al., 2024b; Zhong et al., 2025), since different queries benefit from different granularities even within the same modality: overly fine-grained entries can dilute context, while overly coarse ones may bundle unrelated information. For example, complex analytical questions may require full documents or videos, while simple factoid questions are better served with a single paragraph or short video clip.

To accommodate this, we further decompose each modality into multiple granularity levels, organizing them into distinct corpora. For example, documents are additionally segmented into paragraphs and stored in a paragraph-level corpus, and similarly, videos are divided into short clips and stored, while images are kept intact since they are inherently piecemeal. Overall, with these modality- and granularity-aware corpora (including paragraphs, documents, tables, images, clips, and videos) in place, as well as an additional no-retrieval option to efficiently handle straightforward queries (that require no external knowledge), our UniversalRAG dynamically routes each query to the most relevant knowledge sources, ultimately supporting the diverse information needs of real-world users.

We validate UniversalRAG on 10 datasets spanning diverse modalities and granularities, where it outperforms all baselines by large margins on average, confirming its effectiveness in handling diverse types of queries. Moreover, UniversalRAG improves efficiency via modality-aware retrieval and appropriate granularity selection, while maintaining robustness on out-of-distribution datasets.

2 Method

We begin by describing the preliminaries.

2.1 Preliminaries

Large Vision Language Models Let us first define LLMs, which take an input sequence of tokens $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ and generate an output sequence of tokens $\boldsymbol{y} = [y_1, y_2, \ldots, y_m]$, as follows: $\boldsymbol{y} = \texttt{LLM}(\boldsymbol{x})$, where x_i and y_i are represented in text. Building on top of LLMs, Large Vision-Language Models (LVLMs) extend their capability to support multimodal understanding by incorporating visual encoders (Bai et al., 2023; Chen et al., 2024c; Liu et al., 2024; Li et al., 2025a), to process both the textual and visual inputs. Formally, similar to LLMs, LVLMs can be functionalized as $\boldsymbol{y} = \texttt{LVLM}(\boldsymbol{x})$, whose input token x_i is extended to either textual or visual. However, although they are extensively trained, LVLMs themselves are limited to their parametric knowledge, and often struggle with queries that require (fine-grained or up-to-date) information, less or not exposed for training.

Retrieval-Augmented Generation To address the aforementioned limitations of using only the parametric knowledge, RAG has been widely used, whose core idea is to retrieve query-relevant information from a large corpus and incorporate it into the generation process. Formally, in response to a query $\boldsymbol{q}$, a retrieval model $\mathcal{T}$ fetches the relevant context $\boldsymbol{c}$ from a corpus $\mathcal{C}$: $\boldsymbol{c} = \mathcal{T}(\boldsymbol{q}; \mathcal{C})$. Then, in the subsequent generation step, LVLM generates a response $\boldsymbol{a}$ conditioned on the query and retrieved context: $\boldsymbol{a} = \texttt{LVLM}(\boldsymbol{q}, \boldsymbol{c})$. However, most existing RAG approaches are restricted to retrieving from a single corpus consisting of entries from a single modality (such as only the textual documents), limiting their ability to handle diverse queries with knowledge requirements that vary across them.

2.2 UniversalRAG

We introduce UniversalRAG that dynamically identifies and routes queries to the most appropriate modality and granularity for targeted retrieval.

Challenges in Multi-Corpus Retrieval To accommodate the diverse knowledge needs of real-world queries, which may involve heterogeneous sources spanning different modalities, we consider a set of modality-specific corpora, where each corpus $\mathcal{C}_m$ contains items of modality m. Notably, one straightforward approach to operationalize this is to aggregate all corpora into a unified corpus $\mathcal{C}_{\texttt{unified}} = \bigcup_{m \in M} \mathcal{C}_m$ and embed all items into a shared space using a multimodal encoder, as for retrieval over a single corpus: $\boldsymbol{c} = \mathcal{T}(\boldsymbol{q}; \mathcal{C}_{\texttt{unified}})$. However, we find this approach suffers from modality gap (Figures 2 and 7), where queries, being textual, align more closely with elements in the text corpus regardless of the modality required. Therefore, instead of forcing all heterogeneous elements into a unified corpus, we propose selectively engaging the most relevant corpora needed for queries.

Modality-Aware Retrieval To sidestep the issue of modality gap (introduced by handling all modalities over the unified space), we instead propose to break down the overall retrieval process into two subsequent stages: (1) identifying the most relevant set of modalities for the query; and (2) performing targeted retrieval within the selected modality-specific corpora. Specifically, instead of aggregating all modality-specific corpora, we preserve each corpus in its original form with an independent embedding space. After that, to direct queries to their best-aligned knowledge sources, we introduce a routing module $\mathcal{R}$ that dynamically predicts the modalities best suited for a query $\boldsymbol{q}$, yielding $\mathcal{R}(\boldsymbol{q}) = M_{\boldsymbol{q}}$ where $M_{\boldsymbol{q}}$ is the set of modalities for $\boldsymbol{q}$. Retrieval is then restricted to the corresponding corpora $\{\mathcal{C}_m \mid m \in M_{\boldsymbol{q}}\}$, using any off-the-shelf retriever $\mathcal{T}_m$ tailored to each modality, thereby avoiding the modality gap issue present in a unified space. Proposition 1 formalizes the advantage of modality-aware routing over unified embeddings, and we provide its proof in Section C.

Proposition 1. *Let the similarity score in a unified embedding space $\mathcal{C}_{\texttt{unified}}$ be defined as*

$$s(\boldsymbol{q}, \boldsymbol{c}) = \alpha \cdot \mathbf{1}\{m(\boldsymbol{q}) = m(\boldsymbol{c})\} + \beta \cdot r(\boldsymbol{q}, \boldsymbol{c}) + \varepsilon,$$

where $\alpha > 0$ induces modality bias and $r(\cdot, \cdot)$ measures relevance. If α dominates the variability of r, modality-aware routing retrieves items from the required modality $m^*(\boldsymbol{q})$ with higher probability than unified embedding retrieval.

However, while this routing principle mitigates the modality gap, organizing corpora solely by the modality might still be suboptimal since different queries require varying levels of granularity.

Granularity-Aware Retrieval To accommodate the varying complexity and information scope of different queries, we extend UniversalRAG to operate not only across modalities but also across different levels of granularity within each modality. To be specific, rather than treating each modality-specific corpus as a flat collection of items, we organize it into representations at multiple resolutions, enabling retrieval to target either fine-grained details or broader context as required by the query. To reflect this richer organization of corpora, the routing module $\mathcal{R}$ expands its prediction space to include modality-granularity pairs best suited to a query, as well as a no-retrieval option for cases where external context is unnecessary: $\mathcal{R} : Q \to \{\varnothing\} \cup \mathcal{P}\left(\bigcup_{m \in M} \{m\} \times G_m\right)$, where M is the set of modalities and G_m is the set of granularities available for modality m. Once the router predicts the relevant pairs, retrieval is performed over the corresponding corpora, using retrievers specialized for each modality to obtain the relevant content $\boldsymbol{c}$. Finally, the LVLM generates the answer $\boldsymbol{a}$ with $\boldsymbol{c}$, customized to the modality and granularity for each individual query, thereby enabling the universal, one-for-all RAG framework.

2.3 Router Implementation Strategies

A key component of UniversalRAG is the router, which is responsible for determining the optimal modality and granularity of knowledge for a query.

Training-based Router To perform the routing task, we first consider training the available models to predict the appropriate modality–granularity pair for each query. However, since ground-truth labels (for the modality and granularity the query should be routed to) are not available, we leverage inductive biases in existing benchmarks, mapping each dataset to routing targets that match its task characteristics, allowing us to automatically obtain a labeled corpus without manual annotation. We then train open-source LVLMs to serve as the router using a multi-hot label representation and cross-entropy loss. At inference time, the router produces a sigmoid distribution over modality-granularity pairs and returns all configurations whose scores exceed a predefined threshold, enabling cross-modal and multi-granularity retrieval when necessary.

Training-free Router Alternatively, we also explore a training-free approach that leverages the broad knowledge and robust reasoning capabilities of modern frontier models, such as Gemini (Gemini Team, 2023). Instead of learning from labeled data, the model is directly prompted to act as a router. To achieve this, we first design the prompt template (used to elicit routing), which describes the objective and includes examples demonstrating how different types of queries correspond to specific retrieval targets (See Figure 8 for details). Then, at inference, the model is prompted with this template to predict the most suitable modality-granularity pairs from a predefined set. This eliminates the need for supervised labels or task-specific training, offering the flexibility to adapt to new domains.

3 Experiment

3.1 Experimental Setup

We now explain the experimental setup, including datasets, models, and implementation details.

Datasets To evaluate UniversalRAG, we compile a comprehensive benchmark covering RAG tasks across seven modalities and granularities. For the no-retrieval setting, we use **MMLU** (Hendrycks et al., 2021). For text-based RAG, we include **Natural Questions (NQ)** (Kwiatkowski et al., 2019) for single-hop, paragraph-level retrieval, and **HotpotQA** (Yang et al., 2018) for multi-hop, document-level retrieval. To consider diverse scenarios, we include **HybridQA** (Chen et al., 2020) for reasoning over text and tables, **MRAG-Bench (MRAG)** (Hu et al., 2025) for image RAG, and **WebQA** (Chang et al., 2022) and **InfoSeek** (Chen et al., 2023a) for cross-modal RAG over text and images. Lastly, for RAG with videos, **LVBench** (Wang et al., 2025a) is used for queries over short or localized video segments, as well as **VideoRAG-Wiki** and **VideoRAG-Synth** (Jeong et al., 2025) for queries grounded on long-form or complete videos. Please refer to Section A for more details.

Knowledge Corpora To support the aforementioned, diverse RAG scenarios with various modalities and granularities, we consider their corresponding corpora. Recall that we define seven routing pathways: **None**, **Paragraph**, **Document**, **Table**, **Image**, **Clip**, and **Video**, with cross-modal routing allowing queries to span multiple modalities. For the paragraph and document corpora, we use Wikipedia at the levels of paragraphs (Karpukhin et al., 2020) and documents (Jiang et al., 2024b). The table corpus is built by collecting tables from the HybridQA benchmark. For the image, we adopt corpora from MRAG-Bench, WebQA, and InfoSeek datasets. Lastly, we construct two video corpora at different scales: a video-level corpus consisting of full-length videos from LVBench and VideoRAG datasets, and a clip-level constructed by segmenting these videos into multiple short clips.

Methods We compare our UniversalRAG to a diverse set of 12 baselines, grouped into four categories. The first is **Naïve**, which directly answers queries without retrieving external knowledge. In addition, the group of **Unimodal RAGs** includes **ParagraphRAG**, **DocumentRAG**, **TableRAG**, **ImageRAG**, **ClipRAG**, and **VideoRAG** methods, which retrieve information exclusively from their respective corpora and leverage it for response generation. The third group of **Unified Embedding Multimodal RAGs** uses multimodal encoders to align different modalities into a shared embedding space for retrieval, including **UniRAG** (Sharifymoghaddam et al., 2025), **GME** (Zhang et al., 2025b), **PE$_{\text{core}}$** (Bolya et al., 2025), and **VLM2Vec-V2** (Meng et al., 2026). **MultiRAG** is included in the last group of **Multi-corpus Multimodal RAGs**, which performs retrieval over all the available corpora and incorporates the retrieved results

Table 1: Results of diverse RAG methods with Qwen3-VL-8B-Instruct across modalities. **Bold** denotes the best performance and underlined indicates the second-best among UniversalRAG variants, using either trained or training-free routers. R-L and BERT correspond to ROUGE-L and BERTScore, respectively.

Models	MMLU	NQ		HotpotQA		HybridQA		MRAG	WebQA		InfoSeek	LVBench	VideoRAG-Wiki		VideoRAG-Synth		Avg
	Acc	EM	F1	EM	F1	EM	F1	Acc	R-L	BERT	Acc	Acc	R-L	BERT	R-L	BERT	
Naïve	74.39	18.85	28.98	21.10	29.53	2.80	7.81	49.22	58.12	93.78	18.10	28.83	19.78	86.51	35.86	90.76	35.59
ParagraphRAG	74.39	39.25	51.32	23.40	31.45	5.10	9.21	46.71	51.70	92.53	19.75	24.07	17.62	85.94	32.86	89.97	37.26
DocumentRAG	71.29	21.95	30.26	26.35	34.72	3.75	7.40	43.68	45.57	91.50	16.80	15.70	16.54	85.60	31.18	89.61	32.26
TableRAG	72.51	11.80	18.73	16.45	22.28	9.65	13.86	43.39	44.75	91.47	9.15	16.47	12.16	84.04	30.67	89.45	29.45
ImageRAG	73.33	17.15	25.37	19.15	26.15	2.20	5.69	52.55	67.96	95.65	20.15	25.35	19.50	87.06	36.10	90.77	35.04
ClipRAG	73.33	16.70	24.56	19.30	26.75	2.35	6.16	48.93	65.68	94.83	9.85	33.72	21.10	87.67	39.39	91.47	35.18
VideoRAG	74.91	15.85	23.69	20.00	27.02	2.30	5.78	48.04	64.97	94.67	11.25	32.05	20.89	87.65	40.05	91.54	35.01
UniRAG	70.06	19.30	29.71	19.35	26.89	2.85	7.89	44.86	53.26	92.89	19.05	22.65	18.05	86.11	32.41	89.68	32.93
GME	70.41	20.05	29.91	19.50	26.93	3.00	8.00	49.45	55.03	93.26	19.20	23.68	18.01	86.03	33.02	89.95	33.88
PE_{core}	72.11	19.65	29.77	19.00	26.32	3.05	8.02	49.15	54.79	93.07	19.10	23.04	18.24	86.64	32.75	89.80	33.86
VLM2Vec-V2	71.70	19.95	29.88	18.50	25.24	2.95	8.04	46.78	52.35	92.60	18.80	23.55	18.03	86.07	33.38	90.19	33.31
MultiRAG	70.82	20.90	30.02	22.65	30.74	4.35	8.47	45.01	56.73	93.31	19.05	23.55	17.89	85.91	34.24	90.33	34.07
UniversalRAG (Ours)																	
Trained Routers																	
Qwen3-VL-2B-Instruct	74.39	38.65	50.61	**26.10**	**34.61**	**11.05**	**16.23**	**52.55**	**70.22**	**95.86**	23.20	**33.72**	**20.86**	**87.63**	**39.95**	91.51	**42.40**
InternVL3.5-1B	74.39	**38.70**	50.60	25.85	34.29	10.25	14.79	**52.55**	69.14	95.72	**23.35**	**33.72**	20.85	**87.63**	39.90	**91.52**	42.12
T5Gemma 2 270M	**74.62**	38.65	**50.62**	25.90	33.94	9.95	14.70	50.33	69.03	95.66	21.95	33.59	20.81	87.61	39.43	91.38	41.68
Training-free Routers																	
GPT-5	74.27	34.50	46.21	24.35	32.71	4.95	8.79	50.11	62.38	94.52	21.45	32.30	19.61	86.42	35.94	90.69	39.26
Qwen3-VL-8B-Instruct	74.09	35.20	47.09	24.65	33.12	5.25	9.44	50.04	65.27	94.77	20.65	32.43	18.24	86.07	34.77	90.11	39.46
Oracle	74.39	39.25	51.32	26.35	34.72	10.55	15.20	52.55	71.17	96.02	23.35	33.72	20.89	87.65	40.05	91.54	42.45

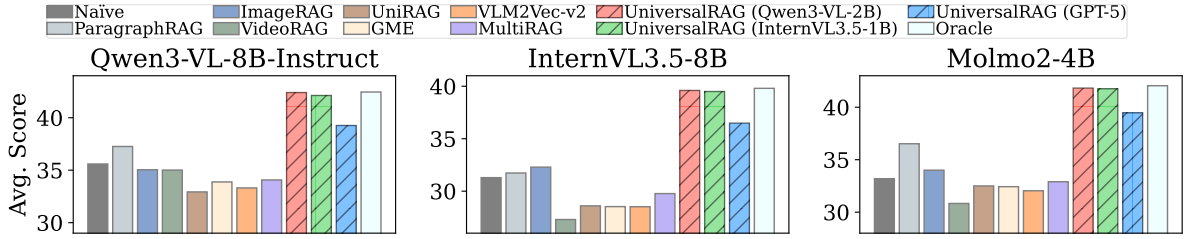


Figure 3: Comparison of averaged evaluation results across different RAG methods and LVLMs.

for response generation. Notably, as **UniversalRAG** can be operationalized with different routing strategies, we implement **training-based variants**, which leverage Qwen3-VL-2B-Instruct (Bai et al., 2025), InternVL3.5-1B (Wang et al., 2025b), and T5Gemma 2 270M (Zhang et al., 2025a) (fine-tuned on the automatically constructed routing dataset), as well as **training-free variants**, which prompt GPT-5 (OpenAI, 2025) and Qwen3-VL-8B-Instruct (Bai et al., 2025) to select appropriate modality-granularity pairs. Finally, we include an oracle setup (**Oracle**), which routes each query to its ideal corpora, non-comparable with others.

Implementation Details For response generation, we utilize multiple LVLMs, Qwen3-VL-8B-Instruct (Bai et al., 2025), InternVL3.5-8B (Wang et al., 2025b), and Molmo2-4B (Clark et al., 2026). Also, to take advantage of UniversalRAG in routing the retrieval process to the modality-specific corpus, we use modality-specific encoders: Qwen3-Embedding-4B (Zhang et al., 2025c) for text, VLM2Vec-V2 (Meng et al., 2026) for vision, and dense row-level embedding (Ji et al., 2024) with the text encoder for tables. We provide further details (including router training) in Section B.

3.2 Experimental Results and Analyses

Now we present the overall results across diverse RAG scenarios, followed by a detailed analysis.

Overall Results We present the modality- and granularity-specific results in Table 1, along with the averaged results with different LVLMs in Figure 3, from which we observe that UniversalRAG consistently achieves the best performance on average. Specifically, in Table 1, the results compared against the unimodal RAG baselines corroborate our hypothesis that retrieving from the modality (or granularity) that aligns best with the information needs of the queries achieves the highest accuracy; however, mismatches between the query and retrieval source results in significant degradation, which supports our claim that considering diverse modalities in the universal workflow is necessary for realistic RAG. Also, the level of granularity within each modality affects performance, suggesting that fine-grained retrieval and generation are necessary. In addition to them, UniversalRAG significantly outperforms unified embedding multimodal RAG baselines, confirming the issue of the modality gap inherent within them (See Figures 2 and 7). Lastly, when compared with the MultiRAG

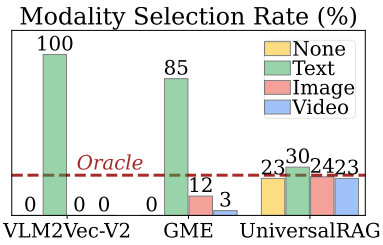


Figure 4: Distribution of the retrieved data modalities.

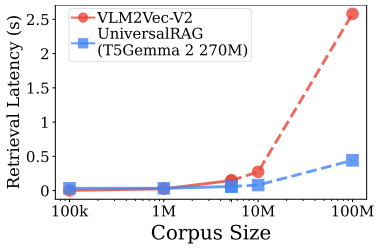


Figure 5: Retrieval latency per query across corpus sizes.

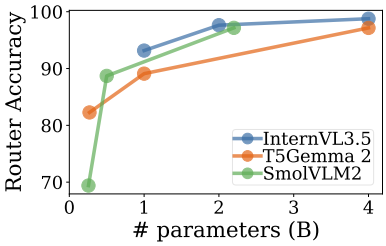


Figure 6: Router accuracy with varying the router model size.

Table 2: Performance comparison of uni-modal and cross-modal approaches across different router models. Among models, GPT-5 is the only training-free router.

Models	Retrieval	HybridQA EM	HybridQA F1	WebQA R-L	WebQA BERT
Qwen3-VL-2B	Uni-modal	9.60	14.56	67.93	95.58
	Cross-modal	**11.05**	**16.23**	**70.22**	**95.86**
InternVL3.5-1B	Uni-modal	9.65	13.86	67.90	95.49
	Cross-modal	**10.25**	**14.79**	**69.14**	**95.72**
GPT-5	Uni-modal	4.75	8.57	60.54	94.04
	Cross-modal	**4.95**	**8.79**	**62.38**	**94.52**

Table 3: Modality accuracy (in corpus selection) and recall of retrieved items for retrieval methods. Among UniversalRAG variants, GPT-5 is only training-free router.

Models	Modality Acc	Recall R@1	Recall R@3	Recall R@5
UniRAG	25.00	0.01	0.03	0.04
GME	36.27	13.84	17.79	22.16
PE_{core}	25.00	0.67	1.20	1.85
VLM2Vec-V2	25.00	2.30	3.69	4.12
UniversalRAG (Qwen3-VL-2B)	**95.28**	**21.38**	**36.29**	**44.82**
UniversalRAG (InternVL3.5-1B)	92.39	19.66	31.82	39.20
UniversalRAG (GPT-5)	68.22	16.33	23.72	31.41

baseline (within the multi-corpus multimodal RAG category), which results in suboptimal performance due to the inclusion of noise from irrelevant modalities in generation, our UniversalRAG remains effective. Its strong performance is due to its core idea around modality-aware routing, enabling the dynamic retrieval from the most relevant modalities and granularities for each query, yielding performance gains despite using several corpora.

Effectiveness of Cross-Modal Retrieval While many queries can be addressed by using a single, most prominent modality, certain tasks benefit from integrating evidence across multiple modalities. For instance, HybridQA requires reasoning that spans both structured tables and accompanying textual sources, while WebQA involves visually grounded questions that pair text with images. Table 2 shows that, compared to uni-modal retrieval, for which each query is routed to a single relevant source, cross-modal retrieval achieves consistently stronger performance. By enabling queries to be routed across multiple modalities, the cross-modal retrieval can leverage complementary evidence that would otherwise be missed by the uni-modal approach. These highlight the effectiveness of UniversalRAG's flexible routing mechanism, which dynamically retrieves information from multiple sources rather than relying on a single modality.

Effectiveness of Modality Routing To investigate the effectiveness of our routing method, we compare the distribution of retrieved modalities for VLM2Vec-V2, GME, and UniversalRAG (with Qwen3-VL-2B) in Figure 4. Using 200 sampled queries per benchmark and normalizing distributions, we find that VLM2Vec-V2 retrieves exclusively text, while GME similarly exhibits a strong bias toward text regardless of the query's required modality, reflecting the modality gap inherent to unified embedding spaces. In contrast, UniversalRAG retrieves more evenly across modalities, indicating that the router effectively mitigates modality bias and adaptively selects appropriate knowledge sources. This leads to higher modality retrieval accuracy, and consequently, higher retrieval recall, as shown in Table 3. While GME achieves comparable recall on text and image corpora, its inability to accurately retrieve from the correct modality leads to lower recall on multimodal corpora that include videos. Yet, UniversalRAG consistently retrieves from the correct modality, enabling it to achieve higher recall than baselines across all scenarios.

Effectiveness of Multigranularity Given the observed benefits of corpus selection in Table 1, we investigate its impact beyond modality by comparing UniversalRAG at varying levels of granularity[1]. Table 4 shows that incorporating granularity-aware corpus selection leads to consistent performance gains by avoiding the retrieval of context that is

[1]In our main experiments, we adopt a binary level of granularity to strike a balance between effectiveness and efficiency.

Table 4: Performance across different numbers of granularity (#Gn) for training-free router models. The prompt used to route to finer granularities is shown in Figure 9.

Models	#Gn	HotpotQA EM	HotpotQA F1	LVBench Acc
GPT-5	1	23.20	31.38	31.92
	2	24.35	32.71	32.30
	3	24.20	32.64	32.43
	4	**24.70**	**33.25**	**32.85**
Qwen3-VL-8B	1	23.85	32.54	31.53
	2	24.65	33.12	32.43
	3	24.70	33.23	32.82
	4	**25.05**	**33.70**	**33.20**

Table 5: Router accuracy and generation performance across retrieval methods on two settings. Among UniversalRAG variants, GPT-5 is the only training-free router.

Models	In-Domain Router Acc	In-Domain Avg Score	Out-Domain Router Acc	Out-Domain Avg Score
Random	14.29	31.75	14.29	37.85
PE_{core}	-	33.86	-	39.08
VLM2Vec-V2	-	33.31	-	38.99
UniversalRAG (Qwen3-VL-2B)	95.81	42.40	71.29	44.07
UniversalRAG (InternVL3.5-1B)	93.16	42.12	67.85	43.80
UniversalRAG (GPT-5)	72.33	41.68	77.38	44.39
Ensemble (Confidence-based)	96.02	42.53	**80.71**	**44.71**
Ensemble (Majority Voting)	**98.33**	**42.83**	78.56	44.54

either insufficient (e.g., a short paragraph lacking key entities for multi-hop reasoning) or excessive (e.g., a full video when only a short clip is relevant), both of which can hinder accurate response generation. Also, as additional granularity levels are introduced, we observe further improvements in some cases, though gains are not strictly monotonic across tasks, reflecting the trade-off between context sufficiency and noise. Please see Section C.2 for a theoretical analysis supporting these findings.

Efficiency of Modality-Specific Retrieval Beyond accuracy, UniversalRAG also improves efficiency by reducing the search space: it leverages modality- and granularity-aware routing to restrict retrieval to only the most relevant sources, instead of querying a unified embedding index that aggregates all modalities into a single mega-corpus. Also, the overhead for routing is small as this cost is outweighed at scale by the size of the search space, leading to sub-linear latency growth as corpus size increases, as shown in Figure 5. Here, UniversalRAG eventually achieves lower latency than unified embedding methods at large corpus sizes, with the gap widening further at very large scales (beyond 10M entries). This scalability makes UniversalRAG a practical solution for real-world applications, where corpora are significantly larger than our experimental settings. We provide an in-depth theoretical analysis of efficiency in Section C.3.

Analysis on Router Size To examine whether the routing cost can be further reduced by using smaller models as routers without sacrificing accuracy, we train three models (Wang et al., 2025b; Zhang et al., 2025a; Marafioti et al., 2025) ranging from 256M to 4B parameters and measure router accuracy. As shown in Figure 6, router accuracy consistently improves with increasing model size within each architecture, suggesting the scalability of our routing approach. While the largest models achieve near-perfect routing performance, a 1B-parameter model attains approximately 90% accuracy, indicating that compact models can serve as effective routers in UniversalRAG.

Generalizability on Out-of-Domain Scenarios As shown in Table 1, UniversalRAG with trained routers outperforms the training-free router (sometimes even approaching oracle performance), and a natural follow-up question is how these routers behave on unseen, out-of-domain (OOD) datasets. To investigate this, we evaluate on six OOD datasets (detailed in Section A.2), with results presented in Tables 5 and 10. In contrast to the in-domain setting, trained routers exhibit noticeable performance degradation, whereas the training-free router generalizes robustly and even surpasses the trained variants. Nevertheless, UniversalRAG remains effective in OOD scenarios and consistently outperforms all baselines, including those using the unified embedding spaces or random modality and granularity assignment, highlighting the benefit of adaptive, modality- and granularity-aware retrieval.

Ensemble Strategy for Robust Routing Building on the trade-off between the high in-domain accuracy of trained routers and the strong OOD generalization of training-free routers, we propose ensemble strategies that leverage their complementary strengths. Specifically, we explore confidence-based ensembling, which uses the trained router's prediction when its confidence exceeds a threshold and otherwise falls back to the training-free router, as well as majority voting, which selects the majority prediction from three routers (training-based and free) with random tie-breaking. Table 5 shows that UniversalRAG with the ensemble routing achieves a robust balance between accuracy and generalization, making it well suited for real-world scenarios with unseen or shifting distributions.

Table 6: Case study comparing unimodal RAGs with fixed modality and granularity against UniversalRAG (Ours).

Question	How many statues of people are there on the Michigan Soldiers Sailors monument?		**Answer:** Nine statues of people.
TextRAG	**Retrieved:** the next section which is surmounted by four male figures depicting the Navy, Infantry, Cavalry, and Artillery branches of the United States Army. Four female allegorical figures, resting on pedestals, are above the male statues and ... **Response:** Eight people ✗	**ImageRAG**	**Retrieved:**  **Response:** Six people ✗
VideoRAG	**Retrieved:**  **Response:** Four people ✗	**Ours**	**Routed to**: Paragraph+Image **Retrieved:** the next section which is surmounted by four male figures depicting the Navy, Infantry, Cavalry, and Artillery branches of the United States Army. Four female allegorical figures, ...  **Response:** Nine people ✓

Case Study We present a case study of UniversalRAG in Table 6. The query asks for the number of statues of people on the Michigan Soldiers and Sailors Monument. Both TextRAG and ImageRAG retrieve the relevant and correct evidence; however, each modality alone is insufficient to determine the full count. TextRAG lacks the information needed to aggregate all statues, while ImageRAG suffers from partial occlusion. VideoRAG fails to retrieve relevant evidence, as the video corpus does not contain information useful for this query. In contrast, UniversalRAG routes the query to both the "Paragraph" and "Image" corpora, allowing cross-modal reasoning and correctly identifying all nine statues. More case studies are provided in Section F.

4 Related Work

Large Vision Language Models Building on the impressive performance of LLMs (Gemini Team, 2023; OpenAI, 2024), recent studies have extended them to visual domains. Liu et al. (2023) incorporates a CLIP-based (Radford et al., 2021) image encoder to align visual inputs with language representations, followed by models using diverse encoders (Bai et al., 2023; Chen et al., 2024c; Liu et al., 2024) and extensions to video (Li et al., 2025a; Wang et al., 2025b; Bai et al., 2025). However, despite improved performance on multimodal benchmarks (Mathew et al., 2021; Yue et al., 2024; Li et al., 2024; Fu et al., 2025) from larger datasets and with improved architectures, LVLMs still often suffer from hallucinations (Huang et al., 2025) when relying solely on parametric knowledge.

Retrieval-Augmented Generation To address the aforementioned limitation of parametric-only models, RAG incorporates external knowledge during response generation. While conventional RAG focuses on the textual corpus (Lewis et al., 2020; Ram et al., 2023), recent work extends it to multimodal sources such as images and videos (Chen et al., 2022; Jeong et al., 2025; Shalev-Arkushin et al., 2026). However, these approaches assume a fixed single-modality retrieval, making them less adaptable to real-world queries that may require information from different modalities. Multimodal encoders (Radford et al., 2021; Zhang et al., 2025b; Bolya et al., 2025; Meng et al., 2026) enable unified embedding spaces across modalities, and Sharifymoghaddam et al. (2025) retrieves from such spaces, but often fails to retrieve visual content for text queries. RAG-Anything (Guo et al., 2025) sidesteps this by converting all the multimodal knowledge into textual form, at the cost of heavy preprocessing and loss of modality-specific information. Other approaches (Cui et al., 2024; Liu et al., 2025a) retrieve from all modalities, followed by extra selection mechanisms, incurring notable computational cost. Lastly, adaptive retrieval strategies (Jeong et al., 2024; Islam et al., 2024; Ding et al., 2025; Yao et al., 2025; Tang et al., 2025) address query diversity but remain restricted to a single corpus (Zhang et al., 2024; Li et al., 2025b).

Retrieval Granularity While most of the existing RAG methods operate at fixed granularity (e.g., full documents, passages, or sentences), real-world queries often require information at varying levels of specificity depending on the knowledge needed, which in turn impacts performance and efficiency in both textual (Chen et al., 2024b; Liu et al., 2025b; Zhong et al., 2025) and video-based retrieval systems (Chen et al., 2023b). In contrast, UniversalRAG performs query-level routing across modality and granularity dimensions, enabling retrieval from the most relevant source at the appropriate level.

5 Conclusion

In this paper, we proposed UniversalRAG, a novel RAG framework designed to retrieve from corpora

of diverse modalities and granularities. Through a modality- and granularity-aware routing mechanism, UniversalRAG dynamically selects the most suitable knowledge sources for each query, effectively addressing the limitations posed by modality gaps and fixed-granularity retrieval, which we further justify through theoretical results. Empirical evaluations across 10 benchmarks demonstrate that UniversalRAG outperforms both modality-specific and unified baselines, showcasing robust performance across diverse modalities. Also, our analyses highlight the importance of fine-grained retrieval and the complementary strengths of training-free and trained routers. We believe these findings demonstrate the potential of UniversalRAG as an adaptive solution for grounding LVLMs with heterogeneous external knowledge, paving the way for the one-for-all RAG that unifies the fragmented landscape of existing corpus-specific RAGs.

Limitations

The proposed UniversalRAG is designed for leveraging heterogeneous, multimodal corpora at RAG, enabling corpus-aware routing to flexibly utilize modality- and granularity-specific corpora. It is worth noting that the routing mechanism is its central part, and to improve its accuracy, high-quality samples for training may be required; however, existing datasets or benchmarks lack ground-truth labels indicating ideal modality or granularity for each query. Nonetheless, we address this by automatically annotating queries (based on inductive biases inherent in datasets or downstream performance measured with all the available corpora), as detailed in Section A. However, since they may contain some noise, constructing high-quality, human-annotated routing datasets would be a valuable direction for future work. Also, due to similar reasons: the absence of annotated data (specifically, the query-granularity pairs), we segment each (text and video) modality into two levels of granularity to obtain supervision signals for router training. Again, collecting more fine-grained annotations that cover a wider range of query-modality and query-granularity pairs would be an exciting direction to expand the applicability of UniversalRAG.

Ethical Considerations

The proposed UniversalRAG can be seamlessly integrated with any LVLMs and compatible retrieval corpora, reducing hallucination with the corpus-specific routing. However, there can be potential private, harmful, or biased content present in the retrieved or generated outputs, depending on the nature of the underlying corpora or the internalized knowledge within LVLMs. To mitigate such risks, it is recommended to apply safeguard mechanisms and filtering techniques in retrieval and generation, to ensure the safe and responsible deployment.

Acknowledgements

This work was supported by the Institute for Information & communications Technology Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program (KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00256259 & RS-2026-25488933), the InnoCORE program of the Ministry of Science and ICT (No. N10250156), the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY) grant funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (RS-2024-00460870), the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-II220713, Meta-learning Applicable to Real-world Problems), and the Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16776–16809, Vienna, Austria. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314, Toronto, Canada. Association for Computational Linguistics.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Abdul Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. 2025. Perception encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2 - 7, 2025*.

Brandon Castellano. 2014. PySceneDetect.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 16474–16483. IEEE.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024a. Benchmarking large language models in retrieval-augmented generation. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17754–17762. AAAI Press.

Tong Chen, Hongwei Wang, Sihao Chen, Wenhao Yu, Kaixin Ma, Xinran Zhao, Hongming Zhang, and Dong Yu. 2024b. Dense X retrieval: What retrieval granularity should we use? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15159–15177, Miami, Florida, USA. Association for Computational Linguistics.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023a. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. InternVL: scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 24185–24198. IEEE.

Zhiguo Chen, Xun Jiang, Xing Xu, Zuo Cao, Yijun Mo, and Heng Tao Shen. 2023b. Joint searching and grounding: Multi-granularity video content retrieval. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 975–983, New York, NY, USA. Association for Computing Machinery.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, and 1 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. 2024. MORE: Multi-mOdal REtrieval augmented generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1178–1192, Bangkok, Thailand. Association for Computational Linguistics.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in llms. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, page 12–21, New York, NY, USA. Association for Computing Machinery.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*, pages 1–17.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025,*

Nashville, TN, USA, June 11-15, 2025, pages 24108–24118. Computer Vision Foundation / IEEE.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gemini Team. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Zirui Guo, Xubin Ren, Lingrui Xu, Jiahao Zhang, and Chao Huang. 2025. Rag-anything: All-in-one rag framework. *arXiv preprint arXiv:2510.12323*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. 2025. Mrag-bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Shayekh Bin Islam, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-RAG: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244, Miami, Florida, USA. Association for Computational Linguistics.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. 2024. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7036–7050, Mexico City, Mexico. Association for Computational Linguistics.

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

Xingyu Ji, Aditya Parameswaran, and Madelon Hulsebos. 2024. TARGET: Benchmarking table retrieval for generative tasks. In *NeurIPS 2024 Third Table Representation Learning Workshop*.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. 2024a. E5-v: Universal embeddings with multimodal large language models. *arXiv preprint arXiv:2407.12580*.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Ziyan Jiang, Xueguang Ma, and Wenhu Chen. 2024b. Longrag: Enhancing retrieval-augmented generation with long-context llms. *arXiv preprint arXiv:2406.15319*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025a. LLaVA-onevision: Easy visual task transfer. *Transactions on Machine Learning Research*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Yangning Li, Yinghui Li, Xinyu Wang, Zhen Zhang, Xinran Zheng, Yong Jiang, Hui Wang, Hai-Tao Zheng, Fei Huang, and Jingren Zhou. 2025b. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)*.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal LLMS. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025a. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 2781–2790, New York, NY, USA. Association for Computing Machinery.

Zuhong Liu, Charles-Elie Simon, and Fabien Caspani. 2025b. Passage segmentation of documents for extractive question answering. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part III*, page 345–352, Berlin, Heidelberg. Springer-Verlag.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, Pedro Cuenca, Cyril Zakka, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, and 1 others. 2025. Smolvlm: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Raghuveer Thirukovalluru, Xuan Zhang, Zeyuan Chen, Ran Xu, Caiming Xiong, Yingbo Zhou, Wenhu Chen, and Semih Yavuz. 2026. VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *Transactions on Machine Learning Research*.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2630–2640. IEEE.

OpenAI. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

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. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24*

July 2021, Virtual Event, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2025. CinePile: A long video question answering dataset and benchmark. In *Workshop on Video-Language Models @ NeurIPS 2024*.

Rotem Shalev-Arkushin, Rinon Gal, Amit H Bermano, and Ohad Fried. 2026. Imagerag: Dynamic image retrieval for reference-guided image generation. In *The Fourteenth International Conference on Learning Representations, ICLR 2026, Rio de Janeiro, Brazil, April 23-27, 2026*.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2025. UniRAG: Universal retrieval augmentation for large vision language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2026–2039, Albuquerque, New Mexico. Association for Computational Linguistics.

Xiaqiang Tang, Qiang Gao, Jian Li, Nan Du, Qi Li, and Sihong Xie. 2025. MBA-RAG: a bandit approach for adaptive retrieval-augmented generation through question complexity. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3248–3254, Abu Dhabi, UAE. Association for Computational Linguistics.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. 2025a. Lvbench: An extreme long video understanding benchmark. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, Hawaii, USA, October 19 - 23, 2025*, pages 22958–22967. IEEE.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. 2025. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. *arXiv preprint arXiv:2502.16636*.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Zijun Yao, Weijian Qi, Liangming Pan, Shulin Cao, Linmei Hu, Liu Weichuan, Lei Hou, and Juanzi Li. 2025. SeaKR: Self-aware knowledge retrieval for adaptive retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27022–27043, Vienna, Austria. Association for Computational Linguistics.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 9556–9567. IEEE.

Biao Zhang, Paul Suganthan, Gaël Liu, Ilya Philippov, Sahil Dua, Ben Hora, Kat Black, Gus Martins, Omar Sanseviero, Shreya Pathak, and 1 others. 2025a. T5gemma 2: Seeing, reading, and understanding longer. *arXiv preprint arXiv:2512.14856*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. 2025b. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 9274–9285. Computer Vision Foundation / IEEE.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025c. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang,

Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025d. Siren's song in the ai ocean: A survey on hallucination in large language models. *Computational Linguistics*, pages 1–46.

Zihan Zhang, Meng Fang, and Ling Chen. 2024. RetrievalQA: Assessing adaptive retrieval-augmented generation for short-form open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6963–6975, Bangkok, Thailand. Association for Computational Linguistics.

Zijie Zhong, Hanwen Liu, Xiaoya Cui, Xiaofan Zhang, and Zengchang Qin. 2025. Mix-of-granularity: Optimize the chunking granularity for retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5756–5774, Abu Dhabi, UAE. Association for Computational Linguistics.

A Additional Details on Dataset

Table 7 provides an overview of all datasets and their corresponding knowledge corpora used in our experiments, including the target modality type as well as the size of the queries and corpora. We divide each dataset into a 3:7 ratio for training and testing. We offer the details of each dataset below.

A.1 In-Domain Dataset

MMLU As a dataset comprising queries that can be answered without the need for retrieval, we use MMLU (Hendrycks et al., 2021), a benchmark that spans a wide range of tasks, including problem-solving abilities (e.g., elementary mathematics, computer science) and world knowledge (e.g., law, world religions). Specifically, we use questions from all tasks in the development split.

Natural Questions (NQ) We also use Natural Questions (Kwiatkowski et al., 2019), a question answering dataset consisting of real user queries issued to the Google search engine, with answers annotated based on supporting Wikipedia articles. We randomly sample 2,000 QA pairs from the dev split, and formulate the text corpus by segmenting the Wikipedia corpus into paragraphs of at most 100 words.

HotpotQA HotpotQA (Yang et al., 2018) is a Wikipedia-based QA benchmark, but it contains complex queries that are annotated to reason over multiple articles. We utilize 2,000 randomly sampled QA pairs of the test split. As it requires multi-hop reasoning over multiple documents, we formulate the text corpus by grouping multiple related documents following LongRAG (Jiang et al., 2024b), which can be longer than 4K tokens.

HybridQA HybridQA (Chen et al., 2020) is a benchmark that requires reasoning over both tabular and textual information. Each question is grounded in a Wikipedia table, but often requires linking to associated text information to locate the correct answer. We randomly sample 2,000 QA pairs from the dev split. Unlike the original benchmark, which directly connects tables and textual evidence, we separate them into distinct table and text corpora to better validate our modality-specific routing-based retrieval framework.

MRAG-Bench (MRAG) We utilize MRAG-Bench (Hu et al., 2025), a vision-centric RAG benchmark that requires only relevant images and does not rely on other modalities, and evaluate on all 1,353 questions. Unlike conventional text-only queries, each query in MRAG-Bench is multi-modal, consisting of a textual question interleaved with a query image. We construct a single image corpus by collecting all images across questions.

WebQA WebQA (Chang et al., 2022) is a benchmark designed to evaluate the ability of LVLMs to reason over multiple sources of information, including both text and images, in an open-domain setting. As the dataset is originally constructed with question-specific retrieval sources that combine text and images, we extract a subset of questions that require retrieval of image for answering. We then further filter these using GPT-4o (OpenAI, 2024) with the prompt shown in Figure 10 to make sure questions are not grounded to a certain image, resulting in a final set of 2,000 QA pairs. Finally, we construct separate text and image corpora by extracting and aggregating evidence from each modality.

InfoSeek InfoSeek (Chen et al., 2023a) is an open-domain benchmark comprising questions interleaved with images, which are best answered by retrieving relevant textual and visual information. For our experiments, we sample 2,000 QA pairs from the dev split and collect the text and image evidence associated with each question to construct the corresponding text and image corpora.

LVBench LVBench (Wang et al., 2025a) is a benchmark developed for long video understanding, featuring questions generated by annotators based on YouTube videos with an average duration of over one hour. Since the benchmark was originally designed for non-RAG tasks, we rephrase the original text-video interleaved queries into a text-only format to align with our experimental setup using GPT-4o, with video metadata and a prompt (Figure 11). Each query is associated with a specific video and a corresponding time range. Notably, the majority of queries are annotated with timestamps spanning less than five minutes, thereby focusing on short segments within the longer videos. Since some videos are currently unavailable, we conduct our evaluation on the available videos and their corresponding questions. For training, we use these short-timestamp queries as a clip-level dataset.

VideoRAG We also utilize VideoRAG-Wiki and VideoRAG-Synth benchmarks, introduced in Vide-

Table 7: Dataset summary for in-domain and out-of-domain benchmarks. Average corpus length denotes the mean token count for text corpora and the mean duration for video corpora.

Dataset	Query Modality	Target Retrieval Modality	# Queries	Corpus Size	Avg Length
In-Domain Datasets					
MMLU	Text	None	1,710	-	-
Natural Questions	Text	Paragraph	2,000	850k	100 tokens
HotpotQA	Text	Document	2,000	509k	693 tokens
HybridQA	Text	Paragraph + Table	2,000	15k	-
MRAG-Bench	Text + Image	Image	1,353	6k	-
WebQA	Text	Paragraph + Image	2,000	20k	-
InfoSeek	Text + Image	Paragraph + Image	2,000	20k	-
LVBench	Text	Clip/Video	777	89	3,865s
VideoRAG-Wiki	Text	Clip/Video	374	9k	378s
VideoRAG-Synth	Text	Clip/Video	374		
Out-of-Domain Datasets					
TruthfulQA	Text	None	790	-	-
TriviaQA	Text	Paragraph	661	661k	100 tokens
SQuAD	Text	Paragraph	2,000	1.19M	100 tokens
2WikiMultiHopQA	Text	Document	2,000	12k	562 tokens
Visual-RAG	Text	Image	374	2k	-
CinePile	Text	Clip/Video	1,440	144	158s

oRAG (Jeong et al., 2025), which are designed to evaluate RAG over a video corpus. These benchmarks are built on the HowTo100M (Miech et al., 2019) corpus (a large-scale collection of instructional YouTube videos) with queries sourced from WikiHowQA (Bolotova-Baranova et al., 2023) and synthetically generated QA pairs based on the videos. Since they lack timestamp annotations, we employ GPT-4o to identify video-level queries that are better answered through full video retrieval rather than short segments from the ground-truth video, which are then used as a video-level dataset for training the router.

A.2 Out-of-Domain Dataset

Unlike the in-domain datasets, the out-of-domain datasets are used solely for evaluation to assess the generalizability of our routing approach and consist only of test splits.

TruthfulQA TruthfulQA (Lin et al., 2022) includes general knowledge questions designed to test whether LLMs can avoid common false beliefs or misconceptions, on diverse categories, including health, law, and politics. We use the multiple-choice version of the dataset, which includes only a single correct answer per question.

TriviaQA TriviaQA (Joshi et al., 2017) is a reading comprehension dataset consisting of trivia questions paired with evidence texts sourced from Wikipedia and the web. To distinguish between queries that require text retrieval and those that do not, we categorize each query based on whether GPT-4o can produce an exact-match answer without access to external text. We randomly sample QA pairs from the dev split. Following the preprocessing strategies used in SQuAD and NQ, all supporting evidence documents are segmented into paragraphs of no more than 100 words.

SQuAD SQuAD v1.1 (Rajpurkar et al., 2016) is a benchmark dataset consisting of questions generated by crowdworkers based on a set of Wikipedia articles. Each question is answerable given the appropriate context paragraph. From the dataset's 100,000+ QA pairs, we randomly sample 2,000 pairs of dev split. For context retrieval, we utilize the full provided Wikipedia corpus, segmenting each article into paragraphs of at most 100 words.

2WikiMultiHopQA We also utilize 2WikiMultiHopQA (Ho et al., 2020), a benchmark designed to evaluate multi-hop reasoning across two Wikipedia articles. We randomly sample 2,000 QA pairs from the dev split and construct a document-level corpus by aggregating all annotated candidate paragraph-level contexts for each question.

Visual-RAG Visual-RAG (Wu et al., 2025) is a question-answering benchmark designed for visual knowledge-intensive questions, specifically tailored for text-to-image retrieval tasks. We utilize the full set of provided queries but sample five images per category to construct the image retrieval pool, ensuring efficient text-to-image retrieval.

CinePile CinePile (Rawal et al., 2025) is a long-video question-answering benchmark that features questions based on movie clips from YouTube. Since the benchmark was originally designed for video understanding tasks rather than RAG, we reformulate each query using the same procedure as LVBench. For each of the 144 available videos, we randomly select 10 questions from the test split. Since CinePile does not provide granularity annotations, we classify the questions into two categories (such as clip-level and full-video-level granularity) using GPT-4o, following the same approach used in VideoRAG.

A.3 Evaluation Metrics

We report results with standard metrics. For datasets with multiple-choice questions, we report Top-1 Accuracy (Acc), the proportion of questions answered correctly. For short-answer datasets, we use Exact Match (EM) and F1, which respectively measure exact agreement and word-level overlap between predictions and references; for InfoSeek, we use the custom accuracy metric defined in the original paper and official repository. For datasets with longer free-form answers, we use ROUGE-L, which captures the longest common subsequences between the prediction and reference (Lin, 2004), and BERTScore, which assesses their semantic similarity (Zhang et al., 2020). We report the average score by averaging first within each modality, then across modalities. Results are obtained from a single run under limited computational resources, while we validate the generality of our framework across multiple backbone models.

B Additional Implementation Details

To effectively leverage both visual and textual information for visual element retrieval, we employ an ensemble approach that combines visual and textual similarity scores with a weighting ratio of 0.8 for visual information. The textual information consists of image captions for images and scripts for videos. To handle long videos, we utilize PySceneDetect (Castellano, 2014), an open-source tool that detects scene boundaries by analyzing content changes (e.g., color histogram differences or threshold-based detection), to segment long videos into shorter clips with an average length of no more than 3 minutes. Moreover, for both the retrieval and generation stages, we uniformly sample 32 frames per video. For baseline models that do not natively support video input, specifically UniRAG (which utilizes CLIP) and GME, we average the embeddings of these sampled frames to obtain a single representative embedding vector.

Training-based routers employ a lightweight classifier head on top of the backbone model to produce logits over multi-label prediction. Multi-label targets are converted into multi-hot vectors, and training is performed via binary cross-entropy loss between these targets and the predicted logits. The router is trained for 5 epochs with a learning rate of 2e-5 and a LoRA rank of $r = 32$. At inference time, routing decisions are made using a predefined threshold of 0.8, selecting all modality-granularity combinations whose sigmoid probabilities exceed the threshold. In contrast, for the training-free variant, we prompt the model using a curated prompt that specifies task objectives and few-shot examples, as shown in Figure 8. Most experiments are conducted on NVIDIA RTX Pro 6000 Max-Q GPUs with 96GB of VRAM.

C Theoretical Analyses of UniversalRAG

In this section, we present formal analyses of each module in UniversalRAG, including the effectiveness of modality routing (Section C.1) and multi-granularity (Section C.2), as well as the efficiency of modality-aware routing (Section C.3).

C.1 Effectiveness of Modality Routing

For a rigorous analysis of the effectiveness of modality routing, we restate Proposition 1 with additional detail and provide a complete proof.

Proposition 1 (Restated). *Let the similarity score in the unified embedding space of $\mathcal{C}_{\texttt{unified}}$ be defined as*

$$s(\boldsymbol{q}, \boldsymbol{c}) = \alpha \cdot \mathbf{1}\{m(\boldsymbol{q}) = m(\boldsymbol{c})\} + \beta \cdot r(\boldsymbol{q}, \boldsymbol{c}) + \varepsilon,$$

where $\alpha > 0$ is a modality bias, $m(\cdot)$ denotes the modality, and $r(\cdot, \cdot)$ measures semantic relevance. If α is sufficiently large relative to the variance of r, the probability of retrieving items from the required modality $m^*(\boldsymbol{q})$ is less than under modality-aware routing followed by within-modality retrieval.

Proof. Without loss of generality, let us consider the top-1 retrieval, as the extension to the top-k case follows directly. Let the unified retrieval corpus

$\mathcal{C}_{\texttt{unified}}$ be decomposed into three disjoint sets:

$$\begin{aligned} S &= \{\boldsymbol{c} : m(\boldsymbol{c}) = m(\boldsymbol{q})\} \\ R &= \{\boldsymbol{c} : m(\boldsymbol{c}) = m^*(\boldsymbol{q})\} \\ O &= \mathcal{C}_{\texttt{unified}} \setminus (S \cup R). \end{aligned} \tag{1}$$

Let us consider the scenario where $m^*(\boldsymbol{q}) \neq m(\boldsymbol{q})$ and $S, R \neq \emptyset$. Define $X_c := \beta \cdot r(\boldsymbol{q}, \boldsymbol{c}) + \varepsilon_{\boldsymbol{c}}$ and suppose $\{X_{\boldsymbol{c}}\}_{\boldsymbol{c} \in \mathcal{C}_{\texttt{unified}}}$ are independent, mean-zero, sub-Gaussian with variance proxy $\sigma^2 = \beta^2 \cdot \text{Var}[r(\boldsymbol{q}, \boldsymbol{c})] + \text{Var}[\varepsilon_{\boldsymbol{c}}]$. Then the similarity scores can be expressed as

$$s(\boldsymbol{q}, \boldsymbol{c}) = \begin{cases} \alpha + X_{\boldsymbol{c}}, & \boldsymbol{c} \in S \\ X_{\boldsymbol{c}}, & \boldsymbol{c} \in R \cup O. \end{cases} \tag{2}$$

Let $M_S = \max_{\boldsymbol{s} \in S} X_{\boldsymbol{s}}$, $M_R = \max_{\boldsymbol{r} \in R} X_{\boldsymbol{r}}$, and $M_O = \max_{\boldsymbol{o} \in O} X_{\boldsymbol{o}}$. Under the unified embedding retrieval, the top-1 item lies in R if and only if

$$M_R \geq \alpha + \max\{M_S, M_O\}.$$

Hence, we can obtain the upper bound of the probability where top-1 retrieval comes from R:

$$\begin{aligned} &\mathbb{P}(\mathcal{T}_{\texttt{unified}}(\boldsymbol{q}; \mathcal{C}_{\texttt{unified}}) \in R) \\ &\quad = \mathbb{P}(M_R \geq \alpha + \max\{M_S, M_O\} \\ &\quad \leq \mathbb{P}(M_R - M_S \geq \alpha). \end{aligned} \tag{3}$$

As $\{M_R - M_S \geq \alpha\} \subseteq \cup_{(\boldsymbol{r}, \boldsymbol{s}) \in R \times S} \{X_{\boldsymbol{r}} - X_{\boldsymbol{s}} \geq \alpha\}$, by the union bound we have

$$\mathbb{P}(M_R - M_S \geq \alpha) \leq \sum_{(\boldsymbol{r}, \boldsymbol{s}) \in R \times S} \mathbb{P}(X_{\boldsymbol{r}} - X_{\boldsymbol{s}} \geq \alpha).$$

As $X_{\boldsymbol{r}} - X_{\boldsymbol{s}}$ is sub-Gaussian with variance proxy $2\sigma^2$, the Chernoff bound of the tail probability combined with Equation (3) leads to

$$\begin{aligned} &\mathbb{P}(\mathcal{T}_{\texttt{unified}}(\boldsymbol{q}; \mathcal{C}_{\texttt{unified}}) \in R) \\ &\qquad \leq |R||S| \exp\left(-\frac{\alpha^2}{4\sigma^2}\right). \end{aligned} \tag{4}$$

By contrast, if the retrieval is done at the modality-specific corpus after modality-aware routing with accuracy r, the probability where the top-1 item is in R is r. Combining this with Equation (4),

$$\begin{aligned} &\mathbb{P}(\mathcal{T}_{\texttt{unified}}(\boldsymbol{q}; \mathcal{C}_{\texttt{unified}}) \in R) \\ &\qquad \leq |R||S| \exp\left(-\frac{\alpha^2}{4\sigma^2}\right) \\ &\qquad < r = \mathbb{P}(\mathcal{T}_{\mathcal{R}(\boldsymbol{q})}(\boldsymbol{q}; \mathcal{C}_{\mathcal{R}(\boldsymbol{q})}) \in R) \end{aligned} \tag{5}$$

whenever $\alpha > 2\sigma\sqrt{\frac{\log(|R||S|)}{r}}$. Meanwhile, the right-hand side of Equation (4) decays to 0 as $\alpha/\sigma \to \infty$. Hence, for α large enough relative to the variance of r, unified embedding retrieval is strictly worse than retrieving from modality-specific corpus after modality-aware routing. □

Remark. Consider very large corpora with $|R| = |S| = 10^{12}$. In this setting, if $p = 0.8$ and $\sigma = 0.01$, then $\alpha > 2\sigma\sqrt{\frac{\log(|R||S|)}{p}} \simeq 0.17$ is sufficient to ensure that routing-based retrieval outperforms unified embedding retrieval. Given that most multimodal encoders exhibit inherent modality biases (as illustrated in Figures 2 and 7), this underscores the necessity of modality-aware routing.

C.2 Effectiveness of Multigranularity

In Sections 3.2 and D.2, we show that routing with multiple granularities within each modality improves performance (see Tables 4 and 9). We also provide a simple statement and proof that support these empirical findings.

Proposition 2. *Let $F(Q; m, g)$ be the expected response quality when retrieving from modality m using granularity g. If there exist queries $\boldsymbol{q}_1, \boldsymbol{q}_2$ and granularities g_f, g_c such that $F(\boldsymbol{q}_1; m, g_f) > F(\boldsymbol{q}_1; m, g_c)$ and $F(\boldsymbol{q}_2; m, g_c) > F(\boldsymbol{q}_2; m, g_f)$, then a routing policy that assigns g_f to $\boldsymbol{q}_1$ and g_c to $\boldsymbol{q}_2$ attains strictly higher expected quality than any fixed-granularity policy.*

Proof. Consider any fixed policy that always uses a single granularity $g \in \{g_f, g_c\}$. If $g = g_f$, then we have

$$\begin{aligned} &F(\boldsymbol{q}_1; m, g_f) + F(\boldsymbol{q}_2; m, g_f) \\ &\qquad < F(\boldsymbol{q}_1; m, g_f) + F(\boldsymbol{q}_2; m, g_c). \end{aligned} \tag{6}$$

Similarly, if $g = g_c$, then we have

$$\begin{aligned} &F(\boldsymbol{q}_1; m, g_c) + F(\boldsymbol{q}_2; m, g_c) \\ &\qquad < F(\boldsymbol{q}_1; m, g_f) + F(\boldsymbol{q}_2; m, g_c). \end{aligned} \tag{7}$$

In both cases, the sum of response quality with the routing policy that applies g_f to $\boldsymbol{q}_1$ and g_c to $\boldsymbol{q}_2$ strictly exceeds that of any fixed granularity g. □

C.3 Efficiency of Modality-Specific Retrieval

While the empirical results in Section 3.2 demonstrate the efficiency benefits of modality-aware routing (with latency trends shown in Figure 5),

we provide a more rigorous analysis on its computational advantages. Let N denote the size of each modality- and granularity-specific corpus, assuming uniform corpus sizes for simplicity, and let k be the number of available routing choices (i.e., the number of modality-granularity pairs). Under a unified embedding approach, retrieval is performed over a single aggregated corpus of size kN, incurring a search cost that scales with the total corpus size. In contrast, UniversalRAG first performs lightweight routing to select the most relevant modality-granularity subset, and then conducts retrieval over only a small selected subset.

Proposition 3. *Let $T(m)$ denote the expected retrieval latency of a single query over a corpus of size m under a fixed retrieval backend, and let the routing cost be a fixed constant C, independent of the number of available routing choices $k > 1$. Then, UniversalRAG achieves lower latency than unified embedding space retrieval on large-scale corpora.*

Proof. Under unified embedding, all modality-granularity corpora are merged into a single index of size kN. Then, the expected per-query retrieval latency is $T_{\texttt{unified}} = T(kN)$. Under UniversalRAG, routing incurs a constant overhead C and then retrieval is executed only on a small number of routed corpora. Assuming retrieval calls of selected corpus are executed in parallel, the end-to-end latency of whole retrieval process is $T_{\texttt{routing}} = C + T(N)$. Let us first consider the case of exact retrieval with embeddings, where the backend exhibits linear scaling $T(m) = \Theta(m)$, then we obtain

$$\frac{T_{\texttt{unified}}}{T_{\texttt{routing}}} \gtrsim \frac{kN}{N + C} = \frac{k}{1 + C/N}. \tag{8}$$

Taking $N \to \infty$ yields

$$\liminf_{N\to\infty} \frac{T_{\texttt{unified}}}{T_{\texttt{routing}}} = \Theta(k), \tag{9}$$

resulting in a linear-in-k speedup. Meanwhile, many modern retrieval systems adopt approximate nearest neighbor search (Douze et al., 2025), which can achieve logarithmic query-time scaling $T(m) = \Theta(\log m)$ (in the best case). Then, for sufficiently large N,

$$\frac{T_{\texttt{unified}}}{T_{\texttt{routing}}} \gtrsim \frac{\log(kN)}{\log N + C} = \frac{\log N + \log k}{\log N + C}. \tag{10}$$

Letting $N \to \infty$, we have

$$\liminf_{N\to\infty} \frac{T_{\texttt{unified}}}{T_{\texttt{routing}}} \geq 1. \tag{11}$$

Thus, even with the approximate retrieval with logarithmic scaling, UniversalRAG achieves a constant-factor asymptotic speedup. Combining these results, UniversalRAG attains strictly lower asymptotic retrieval latency than unified embedding space retrieval for any retrieval methods. □

D Additional Experimental Results

D.1 Additional Results using Different LVLMs

Table 8 shows detailed generation results of baselines and UniversalRAG models on 10 benchmarks using InternVL3.5-8B and Molmo2-4B as generation models. In both settings, UniversalRAG outperforms all baselines and achieves average scores comparable to Oracle. These results demonstrate that UniversalRAG is robust and generalizable in various LVLM generators.

D.2 Additional Results on Multigranularity

Table 4 demonstrates the correlation between the number of granularity levels and end-to-end performance for two training-free models, leveraging the flexibility of our approach in scenarios without labeled data. We further extend this analysis to training-based routers, comparing performance with and without granularity. Table 9 reports results across three training-based router models, consistently demonstrating a performance advantage when granularity is incorporated. These findings underscore the efficacy of including granularity in routing decisions for both training-free and training-based approaches.

D.3 Detailed Results on Out-of-Domain Dataset

We provide the generation results of UniversalRAG variants and baseline methods on each out-of-domain dataset in Table 10. Overall, UniversalRAG consistently outperforms all baselines on average. Notably, the training-free router variants exhibit strong performance across all datasets, showing their outstanding generalization ability to unseen queries. In contrast, trained routers achieve relatively lower performance than on in-domain datasets; nevertheless, they remain robust and still surpass the baseline methods by a large margin.

Table 8: Results of diverse RAG methods with diverse LVLMs (InternVL3.5-8B and Molmo2-4B) across modalities. **Bold** denotes the best performance and underlined indicates the second-best among UniversalRAG variants, using either trained or training-free routers. R-L and BERT correspond to ROUGE-L and BERTScore, respectively.

	Models	MMLU	NQ		HotpotQA		HybridQA		MRAG	WebQA		InfoSeek	LVBench	VideoRAG-Wiki		VideoRAG-Synth		Avg
		Acc	EM	F1	EM	F1	EM	F1	Acc	R-L	BERT	Acc	Acc	R-L	BERT	R-L	BERT	
InternVL3.5-8B	Naïve	71.58	11.75	20.59	14.85	22.02	1.60	5.15	42.50	56.95	93.64	8.05	28.31	20.90	87.39	34.41	90.52	31.29
	ParagraphRAG	68.48	33.60	46.05	19.20	26.27	4.25	7.69	36.81	34.15	89.61	13.25	22.52	17.62	85.52	27.08	88.80	31.73
	DocumentRAG	69.30	19.40	26.85	24.90	33.40	3.35	7.37	35.03	34.36	89.54	11.25	29.60	16.37	84.86	24.07	88.04	30.57
	TableRAG	63.22	6.05	9.85	11.80	16.47	7.30	11.31	40.06	28.99	88.59	4.10	26.38	14.27	83.78	21.22	86.97	25.20
	ImageRAG	72.75	11.65	18.79	14.85	21.62	1.75	5.09	47.89	58.97	93.85	11.15	29.21	20.97	87.50	34.77	90.55	32.29
	ClipRAG	69.94	9.25	15.00	12.60	18.38	1.95	4.24	32.82	14.48	85.00	6.00	36.04	21.68	88.09	35.43	90.93	26.90
	VideoRAG	70.29	10.10	16.08	14.30	19.53	1.30	3.97	33.48	14.07	84.57	5.35	35.78	22.17	89.14	36.97	91.47	27.30
	UniRAG	69.65	14.85	23.82	17.40	25.34	2.85	6.78	34.96	34.38	89.77	10.45	23.68	18.31	86.02	25.93	88.55	28.60
	GME	69.18	15.40	24.53	17.15	25.31	2.60	6.59	35.33	34.22	89.73	11.10	23.42	17.23	85.39	25.13	88.41	28.53
	PE_{core}	69.24	14.90	23.91	17.50	25.74	2.75	6.65	34.81	31.74	89.02	10.70	24.07	17.68	85.50	25.16	88.32	28.38
	VLM2Vec-V2	69.65	15.25	24.35	16.75	24.89	3.15	7.14	35.70	32.05	89.23	10.85	23.04	17.41	85.42	26.42	88.71	28.52
	MultiRAG	68.54	18.80	28.10	18.90	26.11	3.50	7.62	37.92	37.52	90.34	11.40	22.91	18.52	86.24	26.48	88.63	29.77
	UniversalRAG (Ours)																	
	Trained Routers																	
	Qwen3-VL-2B-Instruct	71.58	33.25	45.58	24.50	33.07	**10.25**	**14.52**	**47.89**	**61.34**	**94.05**	**15.95**	**36.04**	**22.02**	**89.11**	**36.92**	**91.49**	**39.60**
	InternVL3.5-1B	71.58	33.10	45.27	**24.70**	**33.19**	10.05	14.28	**47.89**	60.98	93.86	15.70	**36.04**	21.96	88.97	36.79	91.43	39.50
	T5Gemma 2 270M	**71.87**	**33.40**	**45.70**	24.15	32.29	10.10	14.33	46.71	61.04	93.75	15.70	35.26	22.00	88.94	36.74	91.38	39.24
	Training-free Routers																	
	GPT-5	70.99	31.35	43.82	21.90	30.61	6.65	10.73	45.90	48.87	92.16	12.85	33.85	19.14	87.15	31.24	89.27	36.49
	Qwen3-VL-8B-Instruct	71.17	31.30	43.69	22.85	31.57	6.50	10.58	45.53	50.32	93.73	13.20	34.49	19.06	86.94	31.08	89.11	36.73
	Oracle	71.58	33.60	46.05	24.90	33.40	10.35	15.17	47.89	61.56	94.20	15.85	36.04	22.17	89.14	36.97	91.47	39.80
Molmo2-4B	Naïve	70.12	9.80	18.75	14.40	23.79	2.05	6.36	48.41	64.38	94.80	10.40	32.17	21.50	87.58	35.60	90.75	33.19
	ParagraphRAG	68.36	38.65	50.53	22.00	29.59	5.20	9.83	39.54	63.28	94.26	15.85	30.12	16.88	85.75	32.31	89.77	36.52
	DocumentRAG	68.42	20.50	28.45	25.50	34.51	3.95	8.38	40.28	63.17	94.32	13.20	33.72	16.49	85.39	32.05	89.66	34.53
	TableRAG	67.31	8.70	14.25	15.00	21.39	8.55	13.59	42.79	61.61	94.52	6.10	31.53	14.34	84.63	32.88	89.96	31.04
	ImageRAG	69.88	11.00	18.70	16.35	23.73	1.70	5.73	52.55	71.53	96.31	12.30	32.30	21.12	87.42	33.20	90.46	34.00
	ClipRAG	69.36	9.55	16.59	15.15	22.29	1.95	5.50	30.67	66.42	94.98	8.45	38.48	21.62	87.27	35.77	90.82	31.13
	VideoRAG	69.12	9.75	16.98	15.65	23.13	1.50	5.22	30.75	65.98	94.90	5.90	36.55	21.98	87.91	35.96	91.04	30.83
	UniRAG	67.95	12.10	21.05	16.35	24.12	3.85	8.26	41.54	62.83	94.28	13.55	33.98	17.21	86.03	32.54	89.91	32.50
	GME	68.13	12.45	21.32	16.40	24.35	3.70	8.14	41.17	63.07	94.31	13.20	33.59	17.04	85.91	32.18	89.73	32.43
	PE_{core}	68.25	13.35	21.28	16.75	24.23	3.50	7.92	41.32	63.02	94.22	13.45	32.43	16.89	85.73	32.22	89.81	32.28
	VLM2Vec-V2	68.01	12.20	21.07	16.45	24.06	3.60	8.03	40.28	62.79	94.09	12.85	32.05	17.17	85.99	32.84	90.01	32.04
	MultiRAG	68.42	13.30	22.44	18.20	25.43	4.15	8.51	42.27	64.14	94.50	14.60	32.18	16.70	85.68	32.45	89.90	32.90
	UniversalRAG (Ours)																	
	Trained Routers																	
	Qwen3-VL-2B-Instruct	**70.12**	**37.95**	**49.83**	**25.35**	**34.30**	**10.30**	**15.23**	52.55	**73.38**	**96.89**	17.20	**38.61**	21.72	**87.56**	**35.68**	**90.80**	**41.83**
	InternVL3.5-1B	**70.12**	37.85	49.62	**25.35**	**34.30**	10.15	15.08	52.55	73.27	96.81	**17.35**	38.48	**21.73**	87.54	35.57	90.79	41.76
	T5Gemma 2 270M	69.94	37.90	49.60	25.30	34.04	9.45	14.70	**52.70**	73.08	96.75	17.00	37.32	21.65	87.55	35.38	90.73	41.48
	Training-free Routers																	
	GPT-5	69.88	32.80	44.67	23.05	32.78	5.75	10.34	51.07	70.43	95.47	16.85	36.55	19.86	86.61	33.72	90.42	39.47
	Qwen3-VL-8B-Instruct	70.06	33.55	45.23	23.30	33.27	5.90	10.51	51.66	71.21	96.06	16.90	37.19	19.64	86.46	33.66	90.37	39.83
	Oracle	70.12	38.65	50.53	25.50	34.51	10.45	15.39	52.55	74.14	97.13	17.50	38.48	21.98	87.91	35.96	91.04	42.05

Table 9: Effect of granularity on the performance for training-based router models. Gn denotes Granularity.

Models	Gn	HotpotQA		LVBench
		EM	F1	Acc
Qwen3-VL-2B-Instruct	✗	22.25	30.38	32.05
	✓	**26.10**	**34.61**	**33.72**
InternVL3.5-1B	✗	23.00	30.89	32.05
	✓	**25.85**	**34.29**	**33.72**
T5Gemma 2 270M	✗	22.55	30.61	31.40
	✓	**25.90**	**33.94**	**33.59**

E Modality Gap in Unified Embedding Space

Figure 7 visualizes the modality gap within the unified embedding space of six multimodal encoders (Jiang et al., 2024a; Bolya et al., 2025; Lin et al., 2025; Zhang et al., 2025b; Meng et al., 2026; Li et al., 2026). The PCA plot reveals that embeddings cluster by modality, with text embeddings (shown in green) exhibiting larger distances from those of other modalities. Recent methods like E5-V, GME, and Qwen3-VL-Embedding focus on better aligning these modalities to narrow the gap. However, despite these efforts, a noticeable separation between modalities remains, indicating that current multimodal encoders still struggle to fully unify the embedding space across text, images, and videos. Therefore, the modality routing mechanism of UniversalRAG is required to dynamically direct each query to its corresponding modality-specific embedding space, thereby effectively bridging the modality gap and enhancing retrieval performance.

F Qualitative Results

We present case studies to demonstrate the effectiveness of UniversalRAG. Table 11 compares the results of various RAG approaches, including traditional single-modality methods and UniversalRAG, on queries from the WebQA dataset. Traditional approaches such as TextRAG and VideoRAG fail to generate accurate answers: TextRAG retrieves passages lacking relevant visual details, while VideoRAG is better suited for temporal rea-

Table 10: Results of diverse RAG methods on out-of-domain dataset with Qwen3-VL-8B-Instruct across modalities. **Bold** denotes the best performance and underlined indicates the second-best among UniversalRAG variants, using either trained or training-free routers. R-L and BERT correspond to ROUGE-L and BERTScore, respectively.

Models	TruthfulQA	TriviaQA		SQuAD		2WikiMultiHopQA		Visual-RAG		Cinepile	Avg
	Acc	EM	F1	EM	F1	EM	F1	R-L	BERT	Acc	
Naïve	70.00	53.25	61.51	16.75	25.32	37.60	46.23	10.82	82.78	30.76	38.75
ParagraphRAG	68.86	55.82	63.78	34.40	44.27	41.35	50.86	8.95	80.91	30.42	42.62
DocumentRAG	68.10	52.95	61.35	18.10	27.04	48.40	58.19	8.86	80.74	30.14	41.90
TableRAG	66.08	51.13	59.27	9.35	16.12	33.50	44.01	8.20	80.23	29.72	37.14
ImageRAG	68.48	51.89	59.74	13.90	22.65	31.15	41.86	11.64	83.36	32.71	39.18
ClipRAG	69.11	51.59	59.52	14.45	23.07	34.20	45.13	10.38	82.48	35.97	40.38
VideoRAG	68.86	51.44	59.46	14.20	22.89	33.70	44.89	10.21	82.39	37.36	40.50
UniRAG	68.73	52.04	59.89	14.30	22.93	38.25	47.14	9.14	81.02	28.19	38.92
GME	67.97	53.86	61.73	14.95	23.65	39.40	48.09	8.65	80.67	28.68	39.22
PE_{core}	68.61	52.50	61.11	14.50	23.28	38.10	47.02	8.84	80.84	28.75	39.08
VLM2Vec-V2	68.10	51.89	59.99	13.85	22.66	38.85	47.95	8.70	80.72	28.89	38.99
MultiRAG	69.49	51.29	59.36	13.65	22.47	38.35	47.32	8.43	80.42	29.58	39.15
UniversalRAG (Ours)											
Trained Routers											
Qwen3-VL-2B-Instruct	69.75	54.16	62.23	31.60	41.70	45.20	54.33	10.65	82.64	33.68	44.07
InternVL3.5-1B	**69.87**	**54.46**	62.45	30.75	40.97	44.85	53.89	10.88	82.79	32.64	43.80
T5Gemma 2 270M	69.24	53.71	61.90	30.60	40.84	44.70	53.74	10.52	82.58	33.19	43.61
Training-free Routers											
GPT-5	69.62	**54.46**	**62.58**	**31.85**	**42.02**	**45.85**	**54.67**	11.27	83.21	**34.10**	**44.39**
Qwen3-VL-8B-Instruct	**69.87**	54.31	62.45	31.70	41.86	45.60	54.55	**11.33**	**83.31**	33.82	44.35
Oracle	70.00	55.82	63.78	34.40	44.27	48.40	58.19	11.64	83.36	37.36	46.24

soning tasks. In contrast, UniversalRAG correctly routes the query to the image modality, recognizing that visual information about color is necessary, and successfully generates the correct response. This highlights the advantage of modality-aware routing in leveraging the appropriate data from the correct modality corpus, demonstrating UniversalRAG's ability to adaptively select the most informative modalities and granularities for accurate answer generation.

In addition to modality routing, we observe that UniversalRAG also benefits from retrieving information at the appropriate granularity. Table 12 shows results from HotpotQA, where the query requires complex reasoning over multiple text sources. While paragraph-level granularity fails to provide sufficient context for reasoning, UniversalRAG routes the query to the document-level corpus to retrieve all the textual information necessary for accurate reasoning. Similarly, for video queries, Table 13 shows results from LVBench on the query that requires only a short segment of the full long video to answer. While full-video-level retrieval includes irrelevant content and uniformly sampled frames fail to capture the necessary information, clip-level retrieval focuses on smaller, more relevant segments of the video to ensure that only the most pertinent visual details are considered, leading to a more accurate answer.

UniversalRAG performs cross-modal retrieval, allowing the router to select multiple modality-granularity combinations when required, rather than restricting routing to a single source. Table 14 presents an example from HybridQA, where queries primarily rely on tabular data but benefit substantially from complementary textual evidence. In such cases, factual information is best captured from paragraphs, whereas structured knowledge, such as numerical values, is more effectively represented in tables. By jointly retrieving from both modalities, UniversalRAG effectively aggregates complementary evidence and provides the information necessary to answer the query correctly. In contrast, a unimodal variant that restricts retrieval to a single modality retrieves incomplete evidence and fails to support correct reasoning.

However, there are some cases where the routing mechanism fails, particularly when the query exhibits ambiguity in modality requirement or when the required information spans across multiple modalities. Table 15 shows failure cases in which UniversalRAG, employing GPT-5 as a training-free router, incorrectly routes the modality. In the first example, the router's prediction deviates from the inductive ground-truth label as GPT-5, as a modern frontier model, has prior knowledge beyond the predefined routing taxonomy. Although this results in a nominal misclassification, it does not affect the final generation quality, as the model can answer the query without external retrieval. The router also struggles to distinguish between closely related modalities. As illustrated in the second case,

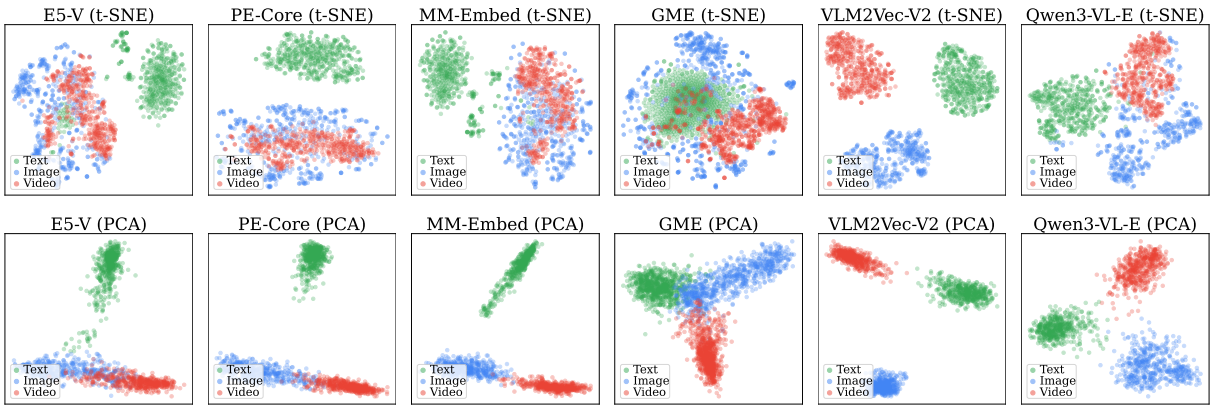


Figure 7: Visualization of the unified embedding space across various multimodal encoders.

a query requiring temporally localized visual evidence is incorrectly routed from clip-level retrieval to static image retrieval. Moreover, the router sometimes exhibits difficulty in determining the appropriate retrieval granularity. Queries that lie near the boundary between different granularity levels are sometimes misrouted, as shown in the third and fourth examples. Finally, the router occasionally fails to recognize cross-modal information needs, leading to incorrect routing decisions for queries that require joint reasoning across modalities, as illustrated in the last example.

Table 11: Case study comparing RAG on single modality against UniversalRAG (Ours). Traditional RAG models are limited by modality-specific retrieval, often missing relevant knowledge from other modalities. UniversalRAG (Ours) dynamically routes each query to the most appropriate modality, enabling more accurate retrieval and response generation.

<table>
<tr><th colspan="2">Comparison of UniversalRAG to Single-modality RAG</th></tr>
<tr><td>Question</td><td>What color balloons were displayed on the USNS Carl Brashear during its ship launching ceremony in San Diego?</td></tr>
<tr><td>Ground Truth</td><td>The balloons were red, white, and blue that were displayed on the USNS Carl Brashear during the ship launching ceremony in San Diego.</td></tr>
<tr><td>No Retrieval</td><td>The USNS Carl Brashear was adorned with blue and gold balloons during its ship launching ceremony in San Diego, reflecting the colors of the U.S. Navy and honoring the legacy of Carl Brashear. ✗</td></tr>
<tr><td>TextRAG</td><td>Retrieved: Operation Tomodachi, "Carl Brashear" loaded more than 800 pallets of humanitarian cargo at Sasebo's Juliet pier on March 20 and set sail later that day to join the Navy ships operating off northern Japan. "Brashear" completed 17 underway replenishment missions, delivering more than 1 million gallons (3,800 m³) of fuel to ...

Response: The documents provided do not mention the color of balloons displayed during the ship launching ceremony of the USNS Carl Brashear in San Diego. ✗</td></tr>
<tr><td>TableRAG</td><td>Retrieved:
<table><tr><th>Inmate Name</th><th>Register Number</th><th>Status</th><th>Details</th></tr><tr><td>Qian Xuesen
Liz Renay
⋮</td><td>Unlisted†
Unlisted†
⋮</td><td>Held at Terminal Island ...
Held at FCI Terminal Island ...
⋮</td><td>Chinese-born rocket scientist ...
Girlfriend of Los Angeles mob kingpin Mickey Cohen ...
⋮</td></tr></table>
Response: The color of balloons displayed on the USNS Carl Brashear during its ship launching ceremony in San Diego is not mentioned in the provided documents. ✗</td></tr>
<tr><td>VideoRAG</td><td>Retrieved:
 
Response: The video does not provide any information about the USNS Carl Brashear or its ship launching ceremony in San Diego, so it is not possible to determine the color of balloons displayed during that event based on the given content. ✗</td></tr>
<tr><td>UniversalRAG (Ours)</td><td>Routed to: Image

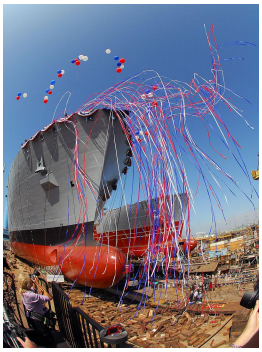
Retrieved:
 
Response: During its ship launching ceremony in San Diego, the USNS Carl Brashear was adorned with red, white, and blue balloons. ✓</td></tr>
</table>

Table 12: Case study comparing different levels of text granularity. The user query requires complex retrieval involving multiple entities. ParagraphRAG retrieves limited context centered around a single entity, leading to an incorrect answer. UniversalRAG (Ours) routes the query to the document corpus and retrieves richer document-level information, allowing it to capture both relevant entities and generate the correct response.

Comparison of Using Different Text Granularity	
Question	George Reed was, along with which Drake University player, one of the greatest running backs in CFL history?
Ground Truth	Johnny D. Bright
ParagraphRAG	**Retrieved:** Matthew Reed (born November 30, 1951) is an American former gridiron football player. He played professionally as a quarterback in the World Football League (WFL) and the Canadian Football League (CFL).A graduate of Grambling State University, Reed was a tenth-round selection (240th overall pick) of the Buffalo Bills in the 1973 NFL Draft but did not play in the league. Reed began his pro career with the Birmingham Americans of the WFL. In 1974, he backed up George Mira, completing 77 of 188 passes (41%) for 1345 yards and 11 touchdowns and 12 interceptions. He was part of their championship team. **Response:** The question contains an error: Matthew Reed is a quarterback, not a running back, and there is no mention of him being associated with Drake University. Additionally, George Reed (a famous CFL running back) is not mentioned in the provided text as being paired with a Drake University player. The text does not contain enough information to answer the question accurately. ✗
UniversalRAG (Ours)	**Routed to:** Document **Retrieved:** *George Reed (Canadian football)* George Robert Reed, CM, SOM (born October 2, 1939), is a former American college football and Canadian Football League player. Reed, along with Mike Pringle and Johnny Bright, is one of the players most often mentioned as being the greatest running back in CFL history. In November 2006, Reed was voted one of the CFL's Top 50 players (#2) of the league's modern era by Canadian sports network. ... *Johnny Bright* Johnny D. Bright (June 11, 1930 – December 14, 1983) was a professional Canadian football player in the Canadian Football League. He played college football at Drake University. He is a member of the Canadian Football Hall of Fame, the National Football Foundation's College Football Hall of Fame, the Missouri Valley Conference Hall of Fame, the Edmonton Eskimos Wall of Honour, the Alberta Sports Hall of Fame, and the "Des Moines Register's" Iowa Sports Hall of Fame. **Response:** Johnny Bright ✓

Table 13: Case study comparing different levels of video granularity. The user query requires only a segment of the video to determine the answer. VideoRAG retrieves a broad range of frames across the video, which may include irrelevant content or miss key frames, leading to an incorrect response. UniversalRAG (Ours) routes the query to the clip-level granularity, retrieving more focused and relevant visual information, enabling it to generate the correct response.

Comparison of Using Different Video Granularity	
Question	What does the protagonist observe through the window after being taken to the utility room in the full episode of Blue Eye Samurai on Netflix? (A) A group of monks sitting cross-legged in the snow (B) A group of citizens chatting together (C) A group of warriors practicing swords (D) A group of samurais eating
Ground Truth	C
VideoRAG	**Retrieved:**  (Timestamp Range: 00:00~1:01:05) **Response:** A ✗
UniversalRAG (Ours)	**Routed to:** Clip **Retrieved:**  (Timestamp Range: 33:46~36:56) **Response:** C ✓

Table 14: Case study comparing UniversalRAG across uni-modal and cross-modal scenarios. In the uni-modal setup, where only a single prominent modality is used, information can sometimes be incomplete as they require evidence across modalities. UniversalRAG (Ours), with its cross-modal capability, gathers evidence from multiple modalities to generate a more comprehensive response.

<table>
<tr><th colspan="2">Comparison of UniversalRAG across Uni- and Cross-Modal Retrieval</th></tr>
<tr><td>Question</td><td>What year did an artist known by the mid-1960s in soul and jazz circles for his recording skills release an American record company and label founded by brothers Max and Sol Weiss in 1949?</td></tr>
<tr><td>Ground Truth</td><td>2000</td></tr>
<tr><td>UniversalRAG (Uni-modal)</td><td>Routed to: Paragraph

Retrieved: David Axelrod (April 17 , 1931 [nb 1] - February 5 , 2017) was an American composer , arranger , and producer . After starting out as a staff producer for record companies specializing in jazz , Axelrod became known by the mid-1960s in soul and jazz circles for his recording skills . In 1968 , Axelrod embarked on a solo career and released several eccentric albums during the 1970s that showcased his characteristic sound , which combined heavily microphoned drums and baroque orchestration , and avant garde themes ranging from the environment to heightened mental awareness . With his early solo projects , Axelrod was one of the first recording artists to fuse elements of jazz , rock , and R & B . One of his most important records , Song of Innocence (1968) , featured instrumental interpretations of 18th-century poet William Blake 's poetry collection of the same name done in a contemporary musical vein , leading one critic at the time to coin the term jazz fusion and numerous hip hop producers to sample the album 's music decades later .

Response: 1960 ✗</td></tr>
<tr><td>UniversalRAG (Ours)</td><td>Routed to: Paragraph+Table

Retrieved: (Above Paragraph with the following table)

<table>
<tr><th>Year</th><th>Album</th><th>Artist</th><th>Genre</th><th>Label</th><th>Credit</th></tr>
<tr><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td></td></tr>
<tr><td>1998</td><td>Greatest Hits</td><td>Joe Cocker</td><td>Rock</td><td>EMI Electrola</td><td>Trombone on You Can Leave Your Hat On</td></tr>
<tr><td>2000</td><td>The Axelrod Chronicles</td><td>David Axelrod</td><td>Jazz , funk , soul</td><td>Fantasy</td><td>Trombone</td></tr>
<tr><td>2004</td><td>Ultimate Collection</td><td>Joe Cocker</td><td>Rock</td><td>Hip-O , A & M</td><td>Horn on You Can Leave Your Hat On</td></tr>
<tr><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td>⋮</td><td></td></tr>
</table>

Response: 2000 ✓</td></tr>
</table>

Table 15: Failure cases in modality routing with UniversalRAG (Ours).

Question	Ground Truth	UniversalRAG (Ours)
What language does the French word polytechnique come from?	Paragraph	No
Who is seated to the right of Kobe in the Jimmy Kimmel tribute show?	Clip	Image
Which book by William A. Dembski summarizes the concepts he introduced about intelligent design in another of his works?	Document	Paragraph
What is the main cause of Lee Chong Wei losing points in the first half of his semi-final match against Lin Dan in the Rio 2016 Olympics replay?	Video	Clip
What is at the top of Hanbit Tower at Expo Science Park?	Paragraph+Image	Paragraph

Classify the following query into one or more categories from: **[No, Paragraph, Document, Table, Image, Clip, Video]**, based on whether it requires retrieval-augmented generation (RAG) and the most appropriate modality. Consider:

- **No**: The query can be answered directly with common knowledge, reasoning, or computation without external data.
- **Paragraph**: The query requires retrieving factual descriptions, straightforward explanations, or concise summaries from a single source.
- **Document**: The query requires multi-hop reasoning, combining information from multiple sources or documents to form a complete answer.
- **Table**: The query requires information that is best represented in a tabular format, often involving comparisons or structured data.
- **Image**: The query focuses on visual aspects like appearances, structures, or spatial relationships.
- **Clip**: The query targets a short, specific moment or event within a video, without needing full context.
- **Video**: The query requires understanding dynamic events, motion, or sequences over time in a video.

Examples:

- "What is the capital of France?" → **No**
- "What is the birth date of Alan Turing?" → **Paragraph**
- "Which academic discipline do computer scientist Alan Turing and mathematician John von Neumann have in common?" → **Document**
- "Among the recipients of the Turing Award, who had the earliest birth year?" → **Table**
- "Describe the appearance of a blue whale." → **Image**
- "Describe the moment Messi scored his goal in the 2022 World Cup final." → **Clip**
- "Explain how Messi scored his goal in the 2022 World Cup final." → **Video**
- "Solve 12×8." → **No**
- "Who played a key role in the development of the iPhone?" → **Paragraph**
- "Which Harvard University graduate played a key role in the development of the iPhone?" → **Document**
- "What is the cheapest iPhone model available in 2023?" → **Table**
- "Describe the structure of the Eiffel Tower." → **Image**
- "Describe the moment Darth Vader reveals he is Luke's father in Star Wars." → **Clip**
- "Analyze the sequence of events leading to the fall of the Empire in Star Wars." → **Video**
- "Describe the visual appearance and habitat of the blue whale." → **Paragraph+Image**
- "Compare the architectural features shown in Gothic and Renaissance cathedrals." → **Image+Table**
- "Describe the moment of the moon landing and explain the mission details." → **Paragraph+Clip**

Classify the following query: `{query}`
Provide only the category or categories combined with '+'.

Figure 8: Prompt for query routing in a training-free manner. The prompt defines each category with concise criteria and illustrative examples. Specifically, examples are designed to contrast closely related cases: for example, Paragraph vs. Document for simple fact retrieval vs. multi-hop reasoning; and Clip vs. Video for short specific moments vs. long-term sequential understanding, highlighting the key aspect that differentiates each category.

Classify the following query into one or more categories from: **[No, Paragraph, Passage, Section, Document, ... , Clip, Sequence, Segment, Video]**, based on whether it requires retrieval-augmented generation (RAG) and the most appropriate modality. Consider:

- **Paragraph**: The query requires retrieving factual descriptions, straightforward explanations, or concise summaries from a single source.
- **Passage**: The query requires a detailed block of text (a few paragraphs) from a single source, with added context.
- **Section**: The query requires retrieving an extensive section of a document explaining a subtopic, possibly with examples or elaboration.
- **Document**: The query requires multi-hop reasoning, combining information from multiple sources or documents to form a complete answer.
- **Clip**: The query targets a short, specific moment or event within a video, without needing full context.
- **Sequence**: The query targets a continuous stretch of related shots (about 10 minutes) that together form a self-contained mini-narrative or process, providing more context and flow than a standalone clip.
- **Segment**: The query targets a longer portion of a video (about 30 minutes) capturing a meaningful sub-scene or subplot-rich and cohesive enough to serve as its own chapter-like unit.
- **Video**: The query requires understanding dynamic events, motion, or sequences over time in a video.

Examples:

- "What is the birth date of Alan Turing?" → **Paragraph**
- "Summarize Alan Turing's concept of the Turing Machine." → **Passage**
- "Explain Alan Turing's contributions to cryptography during WWII." → **Section**
- "Which academic discipline do computer scientist Alan Turing and mathematician John von Neumann have in common?" → **Document**
- "Describe the moment Messi scored his goal in the 2022 World Cup final." → **Clip**
- "Detail the sequence of passes and movements leading to Messi's goal in the 2022 World Cup final." → **Sequence**
- "Describe the build-up sequence during the mid-game period of the 2022 World Cup final." → **Segment**
- "Analyze how Argentina won the 2022 World Cup." → **Video**

Classify the following query: `{query}`
Provide only the category or categories combined with '+'.

Figure 9: Prompt for query routing in a training-free manner with additional granularity choices. Only the components that differ from Figure 8 are shown, including the task objective and few-shot examples.

Evaluate whether the query can be answered using general knowledge about the image's subject rather than relying solely on details unique to the provided image, and verify that the answer is obtainable from the image and the query.

- Respond "yes" if:
  1. The query can be fully answered using general knowledge about the subject.
  2. The answer can be derived solely from the image and the query, without needing image-specific details.
- Respond "no" if either condition is not met.

Example 1:

- Image: A portrait of Donald Trump
- Query: What is the color of Trump's hair?
- Answer: White
- Response: "yes"

Example 2:

- Image: A close-up photo of a light bulb
- Query: What is the color of the light bulb in this image?
- Answer: Yellow
- Response: "no"

Figure 10: Prompt to filter queries for WebQA.

You will receive a query from a video QA dataset and the title of the corresponding video on YouTube. I want you to paraphrase the query by replacing "in the video?", "of the video", or similar phrases with references to the video content naturally. The output should sound as if a human is asking ChatGPT, and should not explicitly mention the exact name of the video or even parts of the title. However, the rephrased query should contain enough implicit information about the video to allow the model to identify it. Try to reduce the chance of the model getting confused between multiple possible video candidates. If there could be multiple video matches for a given query, try to include more information in the rephrased query.

Example 1:

- Query: What year appears in the opening caption of the video?
- Video Title: Blue Eye Samurai | Hammerscale | Full Episode | Netflix
- Upload Date: 2023-11-05
- Channel Name: Netflix
- Rephrased Output: What year appears in the opening caption of the Blue Eye Samurai episode on Netflix?

Example 2:

- Query: After the vlogger sees a dog with an advertisement from the company named Smitten, camera changes to the scene with ___.
- Video Title: My ICELAND Experience | Ultimate Travel Vlog
- Upload Date: 2022-10-26
- Channel Name: Kallmekris
- Rephrased Output: After spotting a dog with a Smitten advertisement, what scene does the camera transition to in Kallmekris's Iceland travel vlog from 2022?

Figure 11: Prompt to rephrase queries using video metadata for LVBench and CinePile.