QUERY-CONDITIONED RESIDUAL RETRIEVAL FOR MULTIMODAL RAG

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal retrieval-augmented generation (RAG) extends external knowledge beyond text to heterogeneous sources such as images, documents, and videos. A practical strategy represents such content using query-independent textual descriptions, whereas recent multimodal retrievers directly index the original multimodal content. These alternatives expose a basic representation problem: a textual description can be semantically correct yet omit a visual or temporal detail required by a particular query. We study this problem through *query-dependent description sufficiency* and introduce *Query-Conditioned Residual Retrieval* (QCR^2), a training-free retrieval approach. For each multimodal item, QCR^2 uses its textual descriptions to construct an item-specific description subspace in a shared multimodal representation space and retains the orthogonal component as a description-conditioned residual. At inference time, QCR^2 independently retrieves from the textual description and residual representations. The first route captures relevance directly expressed by the descriptions, while the second exposes representation directions not contained in their span. Candidates from both routes are jointly ranked, and the selected original items are provided to the answer model. Across heterogeneous text, table, image, and video knowledge sources, QCR^2 consistently improves over description-only, direct multimodal, and description–multimodal fusion retrieval. Extensive analysis shows that QCR^2's advantage grows as textual descriptions become less sufficient for the downstream query, together with increased recovery of supporting evidence missed by description retrieval.

1 INTRODUCTION

Retrieval-augmented generation (RAG) grounds model responses in knowledge retrieved from external sources (Lewis et al., 2020; Karpukhin et al., 2020). Recent multimodal RAG systems extend these sources beyond text to images, visual documents, tables, and videos (Chen et al., 2022; Hu et al., 2023; Jeong et al., 2025; Yeo et al., 2026). A central question is therefore how non-text content should be represented for retrieval. One practical approach converts multimodal content into query-independent textual descriptions and builds a text-based index (Jeong et al., 2025; Liu, 2026), while recent multimodal retrievers directly encode the original content into a shared representation space (Faysse et al., 2025; Yu et al., 2025; Zhang et al., 2025; Meng et al., 2026; Li et al., 2026b). The former benefits from mature text retrieval and a common representation across heterogeneous sources, whereas the latter preserves information that may be difficult to express in text.

We argue that the key limitation of description-based retrieval is not simply whether a description is accurate, but whether it is *sufficient for the current query*. For example, a description stating that the USNS Carl Brashear was decorated with balloons and streamers during its launching ceremony is sufficient for asking what decorations were displayed, but insufficient for asking what colors the balloons were, as illustrated in Figure 1. This ambiguity is inherent to the many-to-many relationship between visual content and language: a multimodal item admits multiple valid descriptions, each exposing only part of its content (Chun et al., 2021; Chun, 2024; Chun et al., 2025). Existing approaches address related but different problems. Caption-index auditing targets incorrect or unreliable descriptions (Liu, 2026) and direct multimodal retrieval avoids textual conversion altogether (Faysse et al., 2025; Yu et al., 2025), while UniversalRAG (Yeo et al., 2026) selects an appropriate modality and granularity for retrieval. However, explicitly distinguishing between in-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

formation already represented in an item's textual descriptions and information available only in its original multimodal representation remains unexplored.

To address this problem, we introduce *Query-Conditioned Residual Retrieval* (QCR^2). Given a multimodal item and its query-independent textual descriptions, QCR^2 identifies the subspace of the multimodal representation explained by those descriptions and retains the orthogonal component as a description-conditioned residual. At inference time, the description and residual representation spaces are retrieved independently. Description retrieval therefore preserves strong semantic matching when relevant information is already verbalized, while residual retrieval provides an additional path for items whose query-relevant content is not captured by their descriptions. Candidates recovered through either route are subsequently ranked jointly, and the corresponding original multimodal evidence is provided to the answer model. Figure 2 illustrates the overall framework of QCR^2.

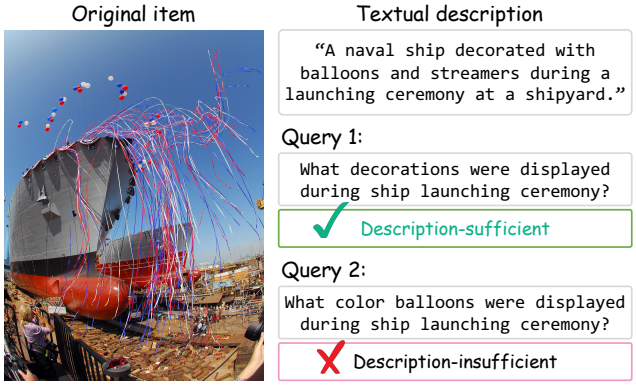


Figure 1: **Illustration of query-dependent description sufficiency.** A query-independent description can correctly summarize the general content of an image while omitting a visual detail required by a particular query.

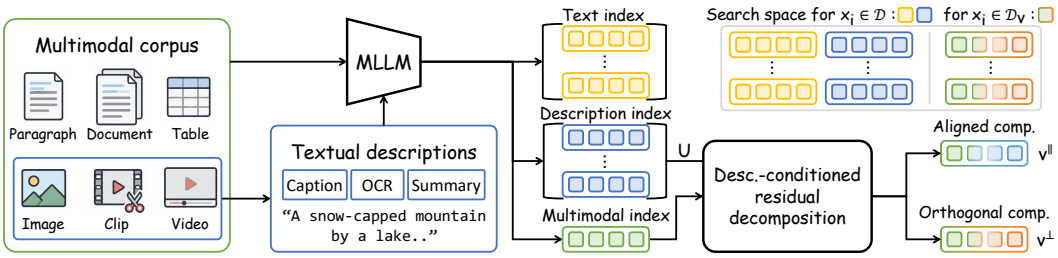
We evaluate QCR^2 under the heterogeneous multimodal RAG benchmark and evaluation protocol, following Yeo et al. (2026), spanning text, tables, images, short video clips, and full videos. In addition to end-to-end performance, we directly examine the proposed mechanism by measuring description sufficiency and identifying supporting evidence recovered exclusively through the residual index. The results show that the advantage of residual retrieval increases as textual descriptions become less sufficient for the query. These results suggest that multimodal retrieval should account for the query-dependent sufficiency of textual representations rather than treating description-based and direct multimodal retrieval as fixed alternatives.

Our contributions are threefold:

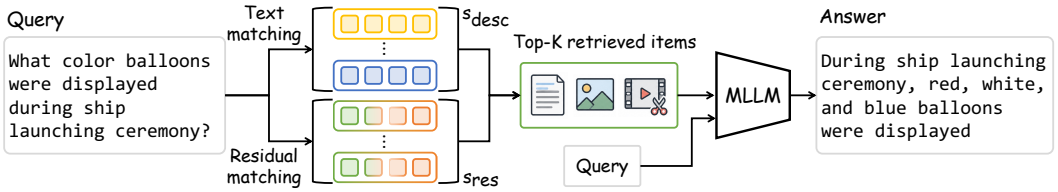
- We formulate query-dependent description sufficiency for multimodal RAG, highlighting that a valid textual description can preserve the information required by one query while omitting that required by another.
- We propose QCR^2, a training-free retrieval approach that decomposes each multimodal representation with respect to its associated textual descriptions and independently retrieves from the description and description-conditioned residual spaces.
- We provide extensive empirical evaluations across heterogeneous multimodal knowledge sources, including description sufficiency, residual-only supporting-evidence recovery, retrieval strategy comparisons, and generalization across retrieval and answer models.

2 Related Work

Retrieval-augmented generation. Retrieval-augmented generation (RAG) augments language models with information retrieved from external knowledge sources (Lewis et al., 2020; Karpukhin et al., 2020; Guu et al., 2020). Subsequent works have improved retrieval, evidence integration, and multi-step retrieval for knowledge-intensive generation (Izacard & Grave, 2021; Borgeaud et al., 2022; Jiang et al., 2023; Trivedi et al., 2023). More recent approaches further organize retrieved evidence through structured or graph-based representations and iteratively expand the supporting evidence when a single retrieval step is insufficient (Edge et al., 2024; Li et al., 2026a; O & Lee, 2026). These methods predominantly consider textual knowledge or the selection and organization of supporting items. Our work focuses instead on the representation of individual multimodal items used during retrieval.



(a) Offline indexing



(b) Inference-time retrieval and generation

Figure 2: **Overview of QCR^2.** (a) At offline indexing, description-aligned and the description-orthogonal (residual) components are precomputed and stored in separate indices. (b) At inference time, QCR^2 independently retrieves from both indices, jointly ranks their candidate union, and provides the selected original multimodal items to the answer model.

Multimodal RAG. Multimodal RAG extends external knowledge beyond text to images, tables, visual documents, and videos (Chen et al., 2022; Hu et al., 2023; Yasunaga et al., 2023; Cui et al., 2024; Jeong et al., 2025). Existing systems adopt several strategies for making such content retrievable. One approach converts multimodal content into textual descriptions, captions, OCR-derived text, transcripts, or summaries and retrieves through the resulting textual index (Jeong et al., 2025; Liu, 2026). Some works directly embed the original multimodal content into a shared retrieval space (Faysse et al., 2025; Yu et al., 2025; Zhang et al., 2025; Meng et al., 2026; Li et al., 2026b). UniversalRAG (Yeo et al., 2026) further considers heterogeneous knowledge sources and selects modality- and granularity-specific corpora before retrieval. These approaches improve which modality or representation is searched, but do not explicitly characterize whether a textual representation preserves the information required by a particular query. QCR^2 addresses this problem by retaining textual retrieval while constructing a complementary representation from the component of the multimodal embedding not represented by the associated descriptions.

Representation multiplicity and sufficiency. A multimodal item generally admits multiple valid textual descriptions, each emphasizing different aspects of its content. Prior work has modeled this multiplicity using probabilistic image–text representations rather than assuming a deterministic one-to-one correspondence (Chun et al., 2021; Chun, 2024; Chun et al., 2025). Related representation-learning approaches separate shared and domain- or modality-specific factors (Bousmalis et al., 2016; Hazarika et al., 2020), while information-theoretic formulations characterize representations according to the task-relevant information they preserve (Achille & Soatto, 2018; Federici et al., 2020). Our formulation differs in two respects. First, description sufficiency is conditioned on the downstream query: the same textual description may preserve the information required by one query while omitting that required by another. Second, rather than learning a global shared–private decomposition, we construct an item-specific residual with respect to the concrete textual descriptions used to index each multimodal item.

3 Method

3.1 Problem Formulation

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote a heterogeneous knowledge corpus consisting of two disjoint subsets: a text-driven subset $\mathcal{D}_t$ and a non-textual subset $\mathcal{D}_v$, such that $\mathcal{D} = \mathcal{D}_t \cup \mathcal{D}_v$. Here, $\mathcal{D}_t$ contains items represented directly through text, including text paragraphs and benchmark-provided textual

serializations of tables, while $\mathcal{D}_v$ includes items represented in non-text modality, including images, video clips, and videos. For each item x_i, we define a set of textual retrieval representations $\mathcal{C}_i$ as

$$\mathcal{C}_i = \begin{cases} \{\tau(x_i)\}, & x_i \in \mathcal{D}_t, \\ \{c_i^{(1)}, \ldots, c_i^{(M)}\}, & x_i \in \mathcal{D}_v, \end{cases} \qquad c_i^{(m)} = T_m(x_i), \tag{1}$$

where $\tau(x_i)$ denotes the native textual representation of a text-driven item and T_m is a description generator. For multimodal items, T_m produces M query-independent textual descriptions that may encode captions, OCR-derived content, summaries, or transcripts depending on the underlying modality (Jeong et al., 2025; Liu, 2026), or emphasize complementary aspects such as overall content, object attributes and relations, fine-grained details, or temporal information.

Given a query q, the retriever selects at most K corpus items from $\mathcal{D}$, and their corresponding benchmark-native content is provided to a multimodal answer model G. Thus, textual representations are used to determine which corpus items to retrieve, whereas the answer model receives the original items corresponding to the selected textual representations. Our goal is to retain the strong semantic matching of this text-driven retrieval while constructing a complementary retrieval representation from directions in the original multimodal embedding that are not represented by $\mathcal{C}_i$.

3.2 Query-dependent description sufficiency

Following the information theory of sufficient representations (Achille & Soatto, 2018; Federici et al., 2020), we define description sufficiency conditionally on the query. Let $Q, X, C = T(X)$, and Y denote the query, multimodal item, textual description, and target answer. A textual description is *query-sufficient* with respect to Y if

$$I(Y; X \mid Q, C) = 0. \tag{2}$$

Equivalently, the conditional distribution of the target satisfies $p(y \mid q, c, x) = p(y \mid q, c)$ for all values in (Q, C, X). Under this condition, the original multimodal item contains no additional information about the answer once the query and description are given.

Description sufficiency is therefore conditioned on the query rather than being an intrinsic property of the description. For the same multimodal item and textual description, the condition in Equation (2) may hold for one query and fail for another, depending on whether the description preserves the information required to determine the corresponding answer. This distinction motivates us to retain information from the original multimodal representation that is not represented by its textual descriptions.

Direct estimation of the conditional mutual information is generally intractable. Instead, we empirically examine description sufficiency in Section 4.3 using the Description Sufficiency Gap (DSG), which measures how much replacing the original supporting evidence with its textual descriptions changes the answer model's support for the reference answer.

3.3 Description-conditioned residual decomposition

We use a frozen multimodal encoder F (*e.g.*, Qwen3-VL-Embedding (Li et al., 2026b), GME (Zhang et al., 2025), or VLM2Vec-V2 (Meng et al., 2026)) that maps text and multimodal inputs into a shared d-dimensional representation space. Let $M_i = |\mathcal{C}_i|$, where $M_i = 1$ for $x_i \in \mathcal{D}_t$ and $M_i = M$ for $x_i \in \mathcal{D}_v$. For a multimodal item $x_i \in \mathcal{D}_v$ and a textual retrieval representation $c_i^{(m)} \in \mathcal{C}_i$, we encode and normalize them to obtain the following embeddings:

$$v_i = \frac{F(x_i)}{\|F(x_i)\|_2}, \qquad t_i^{(m)} = \frac{F(c_i^{(m)})}{\|F(c_i^{(m)})\|_2}. \tag{3}$$

Let $\mathbf{T}_i = [t_i^{(1)}, \ldots, t_i^{(M)}]$ be the collection of description representations of x_i, and U_i be an orthonormal basis for $\mathrm{span}(\mathbf{T}_i)$ obtained by QR decomposition. We decompose the original multimodal representation into the description-aligned component $v_i^{\parallel}$ and the description-orthogonal component $v_i^{\perp}$:

$$v_i^{\parallel} = U_i U_i^{\top} v_i, \qquad v_i^{\perp} = (I - U_i U_i^{\top}) v_i. \tag{4}$$

Here, $v_i^{\parallel}$ lies in the subspace spanned by the textual descriptions associated with the item, while $v_i^{\perp}$ contains representation directions outside that subspace. We use the normalized description-orthogonal component

$$r_i = \frac{v_i^{\perp}}{\max(\|v_i^{\perp}\|_2, \epsilon)} \tag{5}$$

as the residual retrieval representation, where ϵ is a small constant for numerical stability. Residual representations are therefore constructed only for $\mathcal{D}_v$, and items in $\mathcal{D}_t$ are retrieved directly through their textual representations.

3.4 Query-conditioned residual retrieval (QCR^2)

Given a query q, we compute its normalized embedding $z_q = F(q)/\|F(q)\|_2$ and define two relevance scores:

Textual relevance. For every corpus item $x_i \in \mathcal{D}$, we measure relevance through its textual retrieval representations:

$$s_{\text{desc}}(q, i) := \max_{m \in \{1,\dots,M_i\}} z_q^{\top} t_i^{(m)}. \tag{6}$$

For $x_i \in \mathcal{D}_t$, this reduces to similarity with its single native textual representation. For $x_i \in \mathcal{D}_v$, taking the maximum over M descriptions allows an item to receive high textual relevance when at least one description exposes information relevant to the query.

Residual relevance. For multimodal items $x_i \in \mathcal{D}_v$, we additionally measure alignment between the query and the description-orthogonal residual:

$$s_{\text{res}}(q, i) := z_q^{\top} r_i. \tag{7}$$

A high s_{desc} indicates that query relevance is exposed by the textual representation, whereas a high s_{res} indicates that the qeury aligns with directions of the original multimodal representation outside the item's description subspace.

Dual-route candidate retrieval. QCR^2 independently explores the text-driven and residual representation spaces:

$$\mathcal{K}_{\text{desc}} = \text{TopK}_{K_s}\left\{s_{\text{desc}}(q, i) : i \in \mathcal{D}\right\}, \qquad \mathcal{K}_{\text{res}} = \text{TopK}_{K_r}\left\{s_{\text{res}}(q, i) : i \in \mathcal{D}_v\right\}, \tag{8}$$

where K_s and K_r are the numbers of candidates for each subspace. The candidate set for final retrieval is their union:

$$\mathcal{K}_q = \mathcal{K}_{\text{desc}} \cup \mathcal{K}_{\text{res}}. \tag{9}$$

This independent residual route allows multimodal items with weak textual similarity to enter the candidate set without first being retrieved through their descriptions.

Final retrieval and generation. We standardize textual and residual relevance over all candidates in $\mathcal{K}_q$ and $\mathcal{K}_q \cap \mathcal{D}_v$, respectively:

$$\tilde{s}_a(q, i) = \frac{s_a(q, i) - \mu_a(q)}{\sigma_a(q) + \epsilon}, \qquad a \in \{\text{desc}, \text{res}\}, \tag{10}$$

where $\mu_a(q)$ and $\sigma_a(q)$ denote the mean and standard deviation of score s_a over $\mathcal{K}_q$ and $\mathcal{K}_q \cap \mathcal{D}_v$, respectively, and ϵ is a small constant for numerical stability.

We define the final QCR^2 relevance score as

$$s_{\text{QCR}^2}(q, i) = \tilde{s}_{\text{desc}}(q, i) + \lambda \mathbb{1}[x_i \in \mathcal{D}_v]\, \tilde{s}_{\text{res}}(q, i), \qquad i \in \mathcal{K}_q, \tag{11}$$

where λ controls the contribution of residual relevance and $\mathbb{1}[\cdot]$ is an indicator function.

Finally, we select K items from $\mathcal{K}_q$:

$$\mathcal{E}_q = \text{TopK}_K\{s_{\text{QCR}^2}(q, i) : i \in \mathcal{K}_q\}, \tag{12}$$

and the corresponding corpus items are provided to the answer model G.

The two retrieval routes play complementary roles. The textual route employs a common retrieval interface across the heterogeneous corpus, while the residual route provides an additional candidate path for multimodal items whose query-relevant representation is not sufficiently exposed by their textual descriptions. When $v_i^{\perp} = 0$, the original multimodal representation lies entirely within its description subspace and no complementary residual direction is available.

Table 1: End-to-end results on the ten in-domain benchmarks of the UniversalRAG evaluation protocol (Yeo et al., 2026), using Qwen3-VL-8B-Instruct as the answer model. R-L and BERT denote ROUGE-L and BERTScore, respectively. Avg. follows the averaging protocol of Yeo et al. (2026). Published comparison methods use their original retrieval configurations.

Method	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	
No retrieval																	
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
Single-corpus retrieval																	
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
Unified multimodal retrieval																	
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
PEcore	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
Modality- and granularity-aware routing																	
UniversalRAG	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
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
Query-conditioned residual retrieval																	
QCR^2	74.39	38.80	50.90	26.00	34.50	10.90	16.00	**54.10**	**71.80**	**96.08**	**24.20**	**35.00**	**22.10**	**87.96**	**41.55**	**91.84**	**42.9**

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS AND EXPERIMENTAL PROTOCOL

Benchmark and knowledge corpora. We follow the ten-benchmark in-domain evaluation protocol of UniversalRAG (Yeo et al., 2026), covering no retrieval, paragraph, document, table, image, clip, and video settings. We use the corresponding knowledge corpora and evaluate QCR^2 over their heterogeneous union. Unlike UniversalRAG, QCR^2 does not use a modality-granularity router. Details for benchmark and knowledge corpora are provided in Section A.1.

Models and textual descriptions. We use Qwen3-VL-Embedding-2B (Li et al., 2026b) and Qwen3-VL-8B-Instruct (Bai et al., 2025) as the default retrieval and answer models, respectively. We additionally evaluate GME (Zhang et al., 2025) and VLM2Vec-V2 (Meng et al., 2026) for retrieval, and InternVL3.5-8B (Wang et al., 2025b) and Molmo2-4B (Clark et al., 2026) for answer generation to examine whether the behavior of QCR^2 generalizes across multimodal embedding models. For each non-text item, we generate $M = 4$ query-independent textual descriptions with Qwen3-VL-8B-Instruct. Text is indexed directly, and tables use the benchmark-provided textual serialization. We additionally evaluate alternative retrieval and answer models in Section 4.4. Description prompts and generation settings are given in Section A.3

Baselines. We compare QCR^2 with the retrieval approaches evaluated under the same heterogeneous RAG protocol. Following Yeo et al. (2026), these include: **Naïve**, which answers without external retrieval; **the corpus-specific** ParagraphRAG, DocumentRAG, TableRAG, ImageRAG, ClipRAG, and VideoRAG baselines; **unified multimodal retrievers** including UniRAG (Sharifymoghaddam et al., 2025), GME (Zhang et al., 2025), PEcore (Bolya et al., 2025), and VLM2Vec-V2 (Meng et al., 2026); MultiRAG, which retrieves from all available corpora; and **UniversalRAG** (Yeo et al., 2026), which first selects relevant modality–granularity-specific corpora and subsequently performs retrieval within them. We also report the **Oracle** condition of Yeo et al. (2026), which uses the benchmark-specific target corpus and therefore serves as a reference rather than a directly comparable retrieval method.

Evaluation protocol. We follow the official metric of each benchmark and the averaging protocol of UniversalRAG (Yeo et al., 2026). Retrieval quality is additionally evaluated with Recall@K when supporting-item annotations are available. Unless otherwise stated, the final retrieval budget is $K = 5$. Detailed retrieval settings, metric definitions, and implementation parameters are provided in Section A.

Table 2: Comparison of retrieval representations under the same experimental setting. All methods use Qwen3-VL-Embedding-2B, the same heterogeneous knowledge corpora and textual descriptions, a final retrieval budget of $K = 5$, and Qwen3-VL-8B-Instruct for answer generation. Recall@5 is computed on queries for which the benchmark provides annotated supporting evidence; Avg. is the end-to-end score following the UniversalRAG evaluation protocol.

Retrieval representation	Recall@5 ↑	Avg. ↑
Description	38.7	38.6
Direct multimodal	31.4	35.3
Description + multimodal	43.1	40.5
Description + multimodal (learned fusion)	44.0	40.8
QCR^2	**48.3**	**42.9**

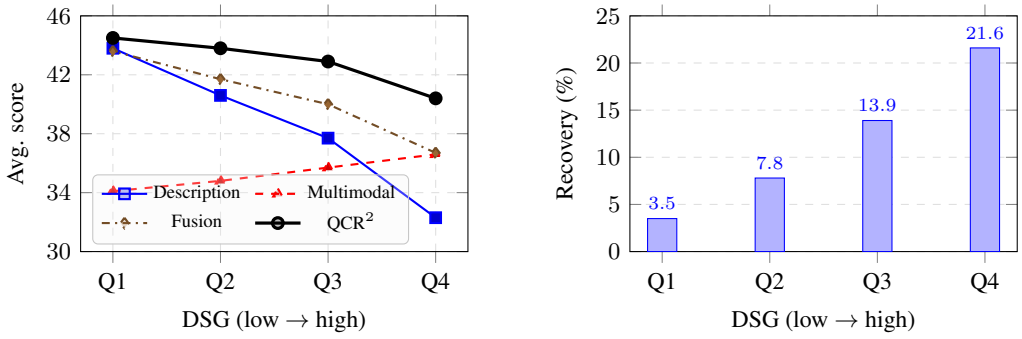
4.2 Main results

Heterogeneous multimodal RAG. Table 1 reports end-to-end performance on the ten benchmarks following the evaluation protocol of UniversalRAG (Yeo et al., 2026). Existing single-corpus RAG methods perform well when the retrieval source matches the information required by the query, but degrade substantially outside their designated modality or granularity. For example, ParagraphRAG achieves strong results on NQ, while ImageRAG and ClipRAG are more effective on image- and video-oriented tasks, respectively. This observation is consistent with prior findings that the appropriate retrieval source depends on both the modality and granularity required by the query (Yeo et al., 2026).

Unified multimodal retrievers provide a common retrieval space across heterogeneous sources (Sharifymoghaddam et al., 2025; Zhang et al., 2025; Bolya et al., 2025; Meng et al., 2026), but remain below UniversalRAG in the heterogeneous-corpus setting. UniversalRAG obtains an average score of 42.40 by explicitly selecting the relevant modality-granularity-specific corpora, compared with 33.88 for GME and 33.31 for VLM2Vec-V2. In contrast, QCR^2 performs retrieval directly over the heterogeneous corpus without a modality or granularity router and achieves the best average performance. The gains are most pronounced on tasks involving image and video evidence, including MRAG-Bench, InfoSeek, LVBench, and VideoRAG, whereas the differences on text-only benchmarks are smaller. This result demonstrates the intended response of QCR^2: the residual branch provides additional retrieval evidence when query-relevant visual or temporal information is not preserved by the textual description.

Comparison of retrieval strategies. We next compare alternative representations under an identical retrieval setting to determine whether the improvement arises from the description-conditioned residual rather than from the retrieval backbone or the use of two indices. All methods in Table 2 use Qwen3-VL-Embedding-2B, the same heterogeneous knowledge corpora, retrieval budget, textual descriptions, Qwen3-VL-8B-Instruct answer model, and generation protocol. *Description* retrieves only from the textual-description index; *Direct multimodal* retrieves from the original multimodal embeddings; and *Equal fusion* and *Learned fusion* combine description and direct multimodal relevance scores. The latter two therefore have access to the same two sources of information as QCR^2 but do not remove description-explained directions from the multimodal representation.

Description-based retrieval performs substantially better than direct multimodal retrieval over the heterogeneous corpus. This is consistent with the modality-gap phenomenon observed in multimodal contrastive representation spaces (Liang et al., 2022) and with the cross-modal retrieval behavior reported by UniversalRAG (Yeo et al., 2026). Combining description and multimodal similarities further improves both retrieval recall and downstream performance, confirming that the original modality contains information complementary to its textual descriptions. However, QCR^2 consistently outperforms both equal-weight and learned fusion. The improvement indicates that simply adding the full multimodal similarity is insufficient: isolating the component not represented by the associated textual descriptions provides a more useful complementary retrieval signal. We examine this mechanism directly in Section 4.3.



(a) Performance across DSG groups.

(b) Residual-only supporting-evidence recovery.

Figure 3: **Description sufficiency and residual retrieval.** (a) Retrieval performance on queries partitioned into four equal-sized groups according to the Description Sufficiency Gap (DSG), where Q1 and Q4 denote the lowest- and highest-DSG groups, respectively. The advantage of QCR^2 increases as textual descriptions become less sufficient for the downstream query. (b) Fraction of annotated supporting evidence retrieved through the residual route but not through description retrieval, $\mathcal{K}_{\text{res}} \setminus \mathcal{K}_{\text{desc}}$. Residual-only recovery becomes increasingly frequent with DSG, linking description insufficiency to complementary evidence retrieval.

4.3 Understanding Query-Conditioned Residual Retrieval

When are textual descriptions insufficient? We first examine whether the benefit of QCR^2 is related to how well textual descriptions preserve the evidence required by a query. For benchmarks that provide annotated supporting evidence, let $\mathcal{S}_q$ denote the supporting set for query q and $\mathcal{C}(\mathcal{S}_q)$ its corresponding textual descriptions. We quantify description sufficiency using the *Description Sufficiency Gap* (DSG) defined by

$$\text{DSG}(q) = s_G(y^* \mid q, \mathcal{S}_q) - s_G(y^* \mid q, \mathcal{C}(\mathcal{S}_q)), \tag{13}$$

where s_G measures the answer model's support for the reference answer y^*. The precise definitions for multiple-choice and free-form answer generation are provided in Section B. A larger DSG indicates a larger reduction in reference-answer support when the original supporting evidence is replaced by its textual descriptions.

To compare retrieval performance across different levels of description sufficiency, we divide queries into four equal-sized groups according to DSG within each benchmark and pool the corresponding groups across benchmarks. Figure 3a shows the results across these groups. For low-DSG queries, description retrieval remains competitive, indicating that the textual representation already provides substantial support for the information required by the query. As DSG increases, description retrieval degrades more sharply, while the advantage of QCR^2 increases. This trend shows that the benefit of QCR^2 is increasingly concentrated on queries for which replacing the original evidence with its textual descriptions produces a larger reduction in reference-answer support.

Does residual retrieval recover supporting evidence missed by description? We next examine whether the residual route actually introduces relevant evidence that description retrieval fails to retrieve. For each query, we identify annotated supporting items contained in $\mathcal{K}_{\text{res}} \setminus \mathcal{K}_{\text{desc}}$. Figure 3b reports the fraction of supporting evidence recovered exclusively through this route across the same DSG groups. Residual-only recovery becomes increasingly frequent as DSG increases, showing that the independent residual route contributes complementary supporting evidence. We further analyze the downstream effect of residual-only recoveries in Section B.4.

4.4 Ablation Analysis

Component analysis. Table 3 examines the contribution of the principal design choices in QCR^2. Residual-only retrieval reaches 36.9, below description retrieval at 38.6, indicating that the residual representation is complementary to rather than a replacement for the textual representation. Combining the description and original multimodal representations improves performance to 40.5 and 40.8,

Table 3: Ablation of QCR^2 on the UniversalRAG benchmark suite using Qwen3-VL-Embedding-2B for retrieval and Qwen3-VL-8B-Instruct for answer generation.

Method	Avg. ↑
Description only	38.6
Direct multimodal	35.3
Description + multimodal	40.5
Description + multimodal, learned	40.8
Residual only	36.9
Global description subspace	41.4
QCR^2, description candidates only	41.7
QCR^2	**42.9**

Table 4: Performance comparison of QCR^2 across retrieval and answer models.

Retriever	Answer	Fusion	QCR^2
Qwen3-VL-Embedding-2B	Qwen3-VL-8B-Instruct.	40.5	**42.9**
GME		39.1	**41.0**
VLM2Vec-V2		38.7	**40.6**
Qwen3-VL-Embedding-2B	InternVL3.5-8B	39.8	**42.0**
	Molmo2-4B	38.9	**40.8**

but remains below the 42.9 achieved by QCR^2, showing that the gain cannot be explained solely by adding a second multimodal retrieval signal. Using a global description subspace reduces performance to 41.4, supporting the effectiveness of the item-specific decomposition. Finally, restricting residual scoring to candidates already retrieved through descriptions obtains 41.7, confirming the importance of allowing the residual route to introduce candidates independently. Additional ablations are provided in Section C.

Generalization across models. We further evaluate the effectiveness of QCR^2 across different retrieval and answer models. As shown in Table 4, QCR^2 consistently improves over description-multimodal fusion when replacing Qwen3-VL-Embedding-2B with GME (Zhang et al., 2025) or VLM2Vec-V2 (Meng et al., 2026), and when replacing the answer model with InternVL3.5-8B (Wang et al., 2025b) or Molmo2-4B (Clark et al., 2026). These results demonstrate the robustness of QCR^2 to the retrieval and answer models.

Retrieval complexity. All description and residual representations are precomputed. Constructing the description subspace costs $O(dM^2)$ per non-text item for $M \ll d$. At inference time, QCR^2 performs two ANN searches and ranks at most $K_s + K_r$ candidates, without additional training or generation. Detailed indexing cost, storage, and latency measurements are provided in Section D.

5 Conclusion

We introduced QCR^2, a retrieval approach for multimodal RAG motivated by the observation that a query-independent textual description may preserve the general semantics of a multimodal item while omitting information required by a particular downstream query. Instead of treating textual descriptions and original multimodal representations as interchangeable retrieval signals, QCR^2 decomposes each multimodal representation into a description-explained component and a description-conditioned residual. It then searches the description and residual representation spaces independently, allowing relevant evidence to be retrieved either through information explicitly represented in text or through information not captured by the associated descriptions. Our analyses further characterize this behavior through description sufficiency and residual-only recovery, providing a direct way to examine when text-based indexing is sufficient and when access to the original modality remains important. More broadly, our results suggest that multimodal retrieval should account for the query-dependent sufficiency of textual representations rather than committing to either text-only or direct multimodal retrieval for all queries.

AI USE STATEMENT

During manuscript preparation, large language models were used for writing assistance to improve grammar, wording, and readability of text drafted by the authors. The research ideas, methodological design, experimental setup, data analysis, and interpretation of results were carried out by the authors. The authors reviewed the final manuscript and take full responsibility for its content and claims.

ETHICS STATEMENT

Multimodal RAG systems are increasingly used to retrieve information from heterogeneous collections containing documents, images, and videos. Improving retrieval when textual descriptions omit query-relevant information can make such systems more reliable in settings where fine-grained visual, document, or temporal evidence is important. At the same time, stronger multimodal retrieval can also increase access to sensitive, private, copyrighted, or otherwise inappropriate content contained in the underlying corpus. QCR^2 does not address these risks directly and inherits the biases and failure modes of its multimodal encoder, description generator, knowledge corpus, and answer model. Deployments should therefore combine improved retrieval capability with appropriate corpus governance, access control, provenance tracking, and content-safety mechanisms.

REPRODUCIBILITY STATEMENT

Our experiments are conducted on four NVIDIA RTX A6000 GPUs. Details required for reproducing the experiments, including dataset settings, evaluation protocols, baseline implementations, hyperparameter configurations, and the implementation details of QCR^2, are provided in the main text and Appendix. We plan to release the code and supporting implementation upon publication.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *JMLR*, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, Tengyu Ma, Jiale Zhi, Jathushan Rajasegaran, Hanoona Rasheed, Junke Wang, Marco Monteiro, Hu Xu, Shiyu Dong, Nikhila Ravi, Shang-Wen Li, Piotr Dollar, and Christoph Feichtenhofer. Perception encoder: The best visual embeddings are not at the output of the network. In *NeurIPS*, 2025.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andrew Brock, Jack Rae, Oriol Vinyals, Simon Osindero, Karen Simonyan, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *ICML*, 2022.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In *NeurIPS*, 2016.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. WebQA: Multihop and multimodal QA. In *CVPR*, 2022.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of EMNLP*, 2020.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *EMNLP*, 2022.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *EMNLP*, 2023.

Sanghyuk Chun. Improved probabilistic image-text representations. In *ICLR*, 2024.

Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *CVPR*, 2021.

Sanghyuk Chun, Wonjae Kim, Song Park, and Sangdoo Yun. Probabilistic language-image pre-training. In *ICLR*, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *CVPR*, 2026.

Wanqing Cui, Keping Bi, Jiafeng Guo, and Xueqi Cheng. MORE: Multi-modal retrieval augmented generative commonsense reasoning. In *Findings of ACL*, 2024.

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

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Celine Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. In *ICLR*, 2025.

Marco Federici, Anjan Dutta, Patrick Forré, Nate Kushman, and Zeynep Akata. Learning robust representations via multi-view information bottleneck. In *ICLR*, 2020.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. REALM: Retrieval-augmented language model pre-training. In *ICML*, 2020.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM MM*, 2020.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *ICLR*, 2021.

Wenbo Hu, Jia-Chen Gu, Zi-Yi Dou, Mohsen Fayyaz, Pan Lu, Kai-Wei Chang, and Nanyun Peng. MRAG-Bench: Vision-centric evaluation for retrieval-augmented multimodal models. In *ICLR*, 2025.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A. Ross, and Alireza Fathi. REVEAL: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *CVPR*, 2023.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *EACL*, 2021.

Soyeong Jeong, Kangsan Kim, Jinheon Baek, and Sung Ju Hwang. VideoRAG: Retrieval-augmented generation over video corpus. In *Findings of ACL*, 2025.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *EMNLP*, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *EMNLP*, 2020.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 2019.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *NeurIPS*, 2020.

Minghan Li, Junjie Zou, Xinxuan Lv, Chao Zhang, and Guodong Zhou. S2G-RAG: Structured sufficiency and gap judging for iterative retrieval-augmented QA. In *ACL*, 2026a.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026b.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

Yangyang Liu. Audit-and-repair indexing: Probabilistic caption indexes for robust multimodal retrieval. In *ICMR*, 2026.

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. VLM2Vec-V2: Advancing multimodal embedding for videos, images, and visual documents. *TMLR*, 2026.

Sunah O and Jay-Yoon Lee. Subtraction gets you more: Gap-aware retrieval for multimodal multi-hop QA. *arXiv preprint arXiv:2605.28641*, 2026.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. UniRAG: Universal retrieval augmentation for large vision language models. In *Findings of NAACL*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *ACL*, 2023.

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. LVBench: An extreme long video understanding benchmark. In *ICCV*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *ICML*, 2023.

Woongyeong Yeo, Kangsan Kim, Soyeong Jeong, Jinheon Baek, and Sung Ju Hwang. Universal-RAG: Retrieval-augmented generation over corpora of diverse modalities and granularities. In *ACL*, 2026.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *ICLR*, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *ICLR*, 2020.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *CVPR*, 2025.

Table A1: Benchmarks, knowledge sources, and evaluation metrics used in our experiments.

Benchmark	Source	Granularity	Metric
MMLU	–	No retrieval	Acc.
NQ	Text	Paragraph	EM / F1
HotpotQA	Text	Document	EM / F1
HybridQA	Text/Table	Table	EM / F1
MRAG-Bench	Image	Image	Acc.
WebQA	Text/Image	Image	R-L / BERT
InfoSeek	Text/Image	Image	Acc.
LVBench	Video	Clip	Acc.
VideoRAG-Wiki	Video	Video	R-L / BERT
VideoRAG-Synth	Video	Video	R-L / BERT

APPENDIX

A EXPERIMENTAL DETAILS

A.1 BENCHMARKS AND KNOWLEDGE CORPORA

We follow the in-domain evaluation protocol of UniversalRAG (Yeo et al., 2026) on the following benchmarks:

MMLU (Hendrycks et al., 2021) evaluates general knowledge and reasoning without external retrieval. Following UniversalRAG (Yeo et al., 2026), we use MMLU as the no-retrieval setting and report accuracy.

Natural Questions (Kwiatkowski et al., 2019) evaluates open-domain question answering over Wikipedia. Following UniversalRAG, evidence is retrieved from the paragraph-level Wikipedia corpus, and performance is measured using Exact Match (EM) and F1.

HotpotQA (Yang et al., 2018) contains multi-hop questions requiring evidence from multiple Wikipedia passages. Under the UniversalRAG protocol, it evaluates document-level text retrieval and is scored using EM and F1.

HybridQA (Chen et al., 2020) requires reasoning over tabular and textual evidence. We follow UniversalRAG and retrieve from its table corpus using the benchmark-provided textual serialization. Performance is evaluated with EM and F1.

MRAG-Bench (Hu et al., 2025) evaluates retrieval-augmented multimodal reasoning in which relevant visual evidence must be identified from an image corpus. We use the image corpus adopted by UniversalRAG and report accuracy.

WebQA (Chang et al., 2022) contains information-seeking questions requiring textual and visual evidence. We follow the preprocessing protocol of UniversalRAG (Yeo et al., 2026), which constructs a 2,000-example evaluation subset by selecting image-retrieval questions and applying its query-filtering procedure. The images are included in the shared image corpus, and answer quality is measured using ROUGE-L and BERTScore.

InfoSeek (Chen et al., 2023) evaluates visual information-seeking questions that require external knowledge beyond the input image. We follow the UniversalRAG image-retrieval setting and use the official accuracy metric.

LVBench (Wang et al., 2025a) evaluates understanding of long videos. Under the UniversalRAG protocol, we use the text-only reformulation of the original video-grounded queries, constructed using the associated video metadata. Temporally localized segments form the clip-level retrieval corpus, and performance is measured using accuracy.

VideoRAG-Wiki (Jeong et al., 2025) evaluates question answering with knowledge retrieved from a corpus of full videos. We follow the VideoRAG/UniversalRAG evaluation protocol and report ROUGE-L and BERTScore.

Table A2: Agreement between HNSW and exact inner-product retrieval. Overlap@K is averaged across evaluation queries.

Index	Overlap@5 (%)	Overlap@10 (%)
Description	99.7	99.8
Direct multimodal	99.5	99.7
Residual	99.4	99.6

VideoRAG-Synth (Jeong et al., 2025) provides a complementary video-RAG evaluation set with full-video retrieval. We use the corresponding UniversalRAG video corpus and evaluate generated answers with ROUGE-L and BERTScore.

Table A1 summarizes the ten benchmarks, their corresponding knowledge sources, retrieval granularities, and evaluation metrics.

The paragraph- and document-level corpora are constructed from Wikipedia, while the table corpus contains tables from HybridQA. The image corpus combines images from MRAG-Bench, WebQA, and InfoSeek. For video retrieval, we use the full videos from LVBench and VideoRAG together with their temporally segmented clips, following UniversalRAG (Yeo et al., 2026).

For methods operating over a unified heterogeneous collection, including QCR^2, we take the union of the paragraph, document, table, image, clip, and video corpora while preserving each item at its original granularity. QCR^2 does not use the modality–granularity router of UniversalRAG. For MMLU, no external evidence is retrieved.

A.2 Retrieval settings

All corpus representations are precomputed before evaluation. Unless otherwise specified, QCR^2 retrieves $K_s = 5$ candidates from the description index and $K_r = 5$ candidates from the residual index. After duplicate removal, their union is jointly ranked and the top $K = 5$ items are provided to the answer model.

We independently standardize the description and residual relevance scores within the candidate union and set the residual weight to $\lambda = 1$. The sensitivity to λ is examined in Section C.2.

For all comparison methods using two retrieval representations, we use the same per-channel candidate budget and final top-K budget. Retrieved non-text items are provided to the answer model in their original modality; textual descriptions and residual representations are used only for retrieval.

Nearest-neighbor search. We use FAISS with an HNSW flat index (`IndexHNSWFlat`) and inner-product similarity for nearest-neighbor retrieval. All query and corpus embeddings are ℓ_2-normalized before indexing. We set the HNSW graph degree to $M = 32$, `efConstruction`$=$ 200, and `efSearch`$=$ 128 for all description, multimodal, and residual indices. We use no vector quantization or representation compression. The same configuration is used for all methods and representation types. Additional indexing, storage, and latency details are reported in Section D.

Exact-search verification. We verify that the differences among retrieval representations are not caused by representation-dependent approximation error in HNSW. For each query, we compare HNSW retrieval with exact inner-product search over the same indexed vectors and compute

$$\text{Overlap@}K = \frac{\left|\mathcal{K}^{\text{HNSW}} \cap \mathcal{K}^{\text{exact}}\right|}{K}. \tag{14}$$

HNSW recovers more than 99% of the neighbors returned by exact search for all three representation spaces. The slightly lower agreement for residual representations does not materially affect retrieval behavior, indicating that the gains of QCR^2 are not an artifact of approximate nearest-neighbor search.

A.3 TEXTUAL DESCRIPTION CONSTRUCTION

Description generation. For each non-text corpus item x_i, we use Qwen3-VL-8B-Instruct (Bai et al., 2025) to generate four query-independent textual descriptions. The downstream question is never provided during description generation. This construction follows the practical setting of caption- or description-based indexing, where multimodal content is converted into text before the user query is observed (Jeong et al., 2025; Liu, 2026). The four descriptions are generated independently with complementary instructions rather than by repeatedly sampling from an identical prompt.

Images and visual documents. For images, the four instructions emphasize (1) the overall scene and entities, (2) object attributes and relations, (3) fine-grained visual details, and (4) visible textual content when present. For visual documents, the same procedure additionally asks the model to preserve document structure, including headings, tables, figures, labels, and spatial relations between elements. We do not provide the downstream query or benchmark answer when producing any description.

Videos. For videos, the descriptions emphasize (1) the overall event, (2) actions and interacting entities, (3) temporal order and state changes, and (4) fine-grained visual and textual details. The full video input used by the corresponding benchmark is supplied to the description model. This is important because a generic description may correctly summarize the main event while omitting a temporally localized detail required by a later query, which is one of the failure modes studied by QCR^2.

Text and tables. Text passages require no translation and are indexed in their original form. For tabular evidence, we use the textual serialization supplied by the benchmark rather than generating an additional visual description. Accordingly, the residual branch is constructed only for corpus items for which an original multimodal representation and its textual descriptions are both available.

Prompts. The description prompts use the following templates:

> **Image / visual document, description 1:** Describe the overall content of the input, including the main entities, scene, and events. Do not assume any downstream question.
>
> **Description 2:** Describe the attributes of the visible entities and their spatial or semantic relations. Include details that may be omitted by a short caption.
>
> **Description 3:** Describe fine-grained visual details of the input, including small objects, attributes, quantities, and other visually distinguishable information.
>
> **Description 4:** Describe textual and structured information visible in the input, including text, labels, headings, tables, plots, and their relevant relations whenever present.

For videos, we replace the four instructions with:

> **Video description 1:** Describe the overall content and main event of the video.
>
> **Video description 2:** Describe the actions, entities, and interactions occurring in the video.
>
> **Video description 3:** Describe the temporal progression of the video, including event order, motion, and changes of state.
>
> **Video description 4:** Describe fine-grained visual and textual details that may be relevant for understanding the video but are not captured by a high-level summary.

All prompts explicitly prohibit access to the downstream question. We use the same generated descriptions for all retrieval methods that require textual descriptions, ensuring that differences among these methods are not caused by different description-generation procedures.

A.4 EVALUATION METRICS

We follow the official evaluation metric of each benchmark. MMLU, MRAG-Bench, InfoSeek, and LVBench are evaluated using accuracy. Natural Questions, HotpotQA, and HybridQA are evaluated

using Exact Match (EM) and F1. WebQA, VideoRAG-Wiki, and VideoRAG-Synth are evaluated using ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020).

For the overall score reported as Avg., we follow the evaluation protocol of UniversalRAG (Yeo et al., 2026): benchmark metrics are first averaged within each modality and then averaged across modalities.

When benchmark-annotated supporting items are available, we additionally report Recall@K to measure retrieval quality independently of answer generation.

A.5 Baselines

The comparisons in Table 1 follow the configurations reported under the UniversalRAG evaluation protocol (Yeo et al., 2026).

For the controlled retrieval-strategy comparison in Table 2, all methods use Qwen3-VL-Embedding-2B, the same heterogeneous knowledge corpus, the same textual descriptions, the same candidate budget, and Qwen3-VL-8B-Instruct for answer generation.

Description. Retrieval uses only the textual-description representations.

Direct multimodal. Retrieval uses the original multimodal embeddings without textual descriptions.

Description + multimodal. The description and original multimodal representations are searched independently, and their normalized relevance scores are combined for final ranking.

Description + multimodal, learned fusion. This baseline combines the standardized description and direct multimodal scores as

$$s_{\text{fusion}}(q,i) = \alpha\, \widetilde{s}_{\text{desc}}(q,i) + (1-\alpha)\, \widetilde{s}_{\text{mm}}(q,i).$$

A single global $\alpha \in [0,1]$ is selected on a small held-out calibration set and fixed for all test queries.

Residual only. Retrieval uses only the description-orthogonal residual representation.

QCR^2, description candidates only. Candidates are retrieved only from the description index; residual scores are subsequently used for reranking within that candidate set. This variant removes the independent residual retrieval path while retaining the same residual scoring function.

B Description Sufficiency Analysis

B.1 Description sufficiency gap

For a query q with benchmark-annotated supporting evidence $\mathcal{S}_q$, let $\mathcal{C}(\mathcal{S}_q)$ denote the textual descriptions associated with the same evidence. The Description Sufficiency Gap (DSG) compares the answer model's support for the reference answer under the original evidence and under its textual descriptions.

For multiple-choice tasks, we use the log-probability assigned to the correct answer:

$$\text{DSG}(q) = \log p_G(y^* \mid q, \mathcal{S}_q) - \log p_G(y^* \mid q, \mathcal{C}(\mathcal{S}_q)), \tag{15}$$

where y^* denotes the correct answer representation under the same answer format used during evaluation. If the answer representation consists of multiple tokens, its probability is computed autoregressively over the complete token sequence.

For free-form tasks, let $y^* = (y_1, \ldots, y_L)$ denote a tokenized reference answer. We use its length-normalized conditional log-likelihood as DSG:

$$\text{DSG}(q) = \frac{1}{L}\sum_{t=1}^{L}[\log p_G(y_t^* \mid y_{<t}^*, q, \mathcal{S}_q) - \log p_G(y_t^* \mid y_{<t}^*, q, \mathcal{C}(\mathcal{S}_q))]. \tag{16}$$

A positive DSG therefore indicates that the answer model assigns greater support to the reference answer when conditioned on the original supporting evidence than its textual descriptions. Importantly, DSG is used exclusively as a post-hoc analysis quantity and is never available to the retriever or used for candidate ranking, model selection, or hyperparameter selection.

Table B3: Downstream outcomes conditioned on residual-only supporting-evidence recovery. Outcomes compare QCR^2 with description-only retrieval on the same query.

Residual-only recovery	Improve (%)	Unchanged (%)	Hurt (%)
No	9.4	84.8	5.8
Yes	**30.8**	63.0	6.2

Table C4: Effect of the number of query-independent textual descriptions.

# Descriptions M	Description only	QCR^2
1	37.1	41.8
2	38.0	42.4
4	38.6	**42.9**
8	38.9	43.0

B.2 DSG GROUPING

To compare retrieval behavior across different levels of description sufficiency, we partition queries into four equal-sized groups according to DSG. The partition is performed independently within each benchmark to avoid differences in benchmark-specific answer-score scales. The resulting groups are denoted Q1–Q4, where Q1 contains queries with the lowest DSG and Q4 those with the highest DSG. We subsequently pool the corresponding groups across benchmarks for the analysis in Figure 3.

B.3 RESIDUAL-ONLY SUPPORTING-EVIDENCE RECOVERY

We additionally measure whether the residual retrieval branch recovers supporting evidence missed by description retrieval. For each query, residual-only recovery is defined over annotated supporting items retrieved in $\mathcal{K}_{\text{res}} \setminus \mathcal{K}_{\text{desc}}$. We report the fraction of supporting items recovered exclusively through this residual route within each DSG group. As shown in Figure 3, residual-only recovery becomes more frequent as DSG increases, indicating that the residual branch is particularly useful when the textual descriptions omit query-relevant information.

B.4 DOWNSTREAM EFFECT OF RESIDUAL-ONLY RECOVERY

We next examine whether supporting evidence recovered exclusively through the residual route is associated with improved downstream answers. For each query with annotated supporting evidence, we define residual-only recovery to occur when at least one supporting item is retrieved through the residual route but not through description retrieval, *i.e.*,

$$\mathcal{S}_q \cap (\mathcal{K}_{\text{res}} \setminus \mathcal{K}_{\text{desc}}) \neq \emptyset. \tag{17}$$

We then compare QCR^2 with description-only retrieval on the same queries and classify each outcome according to whether the benchmark-specific answer score improves, remains unchanged, or decreases.

When the residual route does not uniquely recover supporting evidence, QCR^2 changes the downstream result only infrequently, improving 9.4% of queries while leaving 84.8% unchanged. In contrast, when the residual route recovers supporting evidence missed by description retrieval, the improvement rate increases to 30.8%, while the fraction of degraded answers remains similar (6.2% versus 5.8%). This result connects the retrieval mechanism to downstream performance: the additional candidates introduced by the residual route are most useful when they contain supporting evidence unavailable to description retrieval.

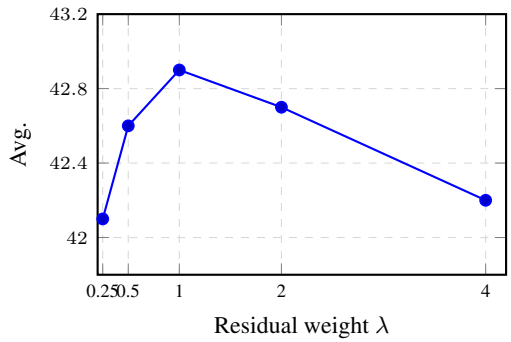


Figure C1: Sensitivity of QCR^2 to the residual weight λ. Performance is stable around $\lambda = 1$.

Table C5: Sensitivity to description and residual candidate budgets. The final budget is fixed to $K = 5$.

K_s	K_r	Recall@5	Avg.
8	2	46.1	42.1
5	5	48.3	42.9
2	8	46.9	42.3
2	2	44.2	41.6
10	10	48.9	43.0
20	20	49.2	43.0

Table C6: Effect of query access during description generation. Query-conditioned descriptions are used only as a diagnostic setting.

Description construction	Description only	QCR^2
Query-independent	38.6	42.9
Query-conditioned	42.4	42.8

C Additional Ablations

C.1 Number and type of textual descriptions.

The default $M = 4$ balances description coverage and indexing cost. We vary the number and construction of textual descriptions to examine how QCR^2 behaves as the text representation becomes more comprehensive. As shown in Table C4, increasing M improves description-based retrieval, but the improvement saturates as additional descriptions become increasingly redundant. The gain from QCR^2 correspondingly decreases but remains positive, suggesting that multiple query-independent descriptions still fail to preserve all information that may become relevant to an unseen query. As a diagnostic upper bound, we additionally generate a description with access to the downstream query. Under this setting, the difference between description retrieval and QCR^2 becomes substantially smaller. This provides a useful boundary case: when the description is constructed specifically for the query, considerably less query-relevant information remains for residual retrieval to recover.

C.2 Sensitivity to the residual weight

In Figure C1, we study the sensitivity of QCR^2 to the residual weight λ while keeping all other retrieval settings fixed. We evaluate $\lambda \in \{0.25, 0.5, 1, 2, 4\}$ and use $\lambda = 1$ as the default setting. Performance remains stable over a broad range around the default value, indicating that QCR^2 does not require precise calibration of the description and residual relevance scores.

C.3 Sensitivity to candidate budgets

We examine the effect of the candidate budgets K_s and K_r for the description and residual retrieval routes, respectively, while keeping the final evidence budget fixed at $K = 5$. We first vary the allocation between the two retrieval routes under a fixed total candidate budget. We further vary the total candidate pool using balanced description and residual budgets. The results show that the benefit of residual retrieval is robust to moderate changes in candidate allocation, with the default $K_s = K_r = 5$ providing a simple balanced configuration.

Table C7: Effect of residual normalization.

Residual score	Recall@5 ↑	Avg. ↑
$z_q^\top v_i^\perp$	47.2	42.4
$z_q^\top r_i$	**48.3**	**42.9**

Table D8: Retrieval efficiency. Index construction is performed once offline from precomputed embeddings. Latency is averaged per query using the same HNSW configuration.

Method	# Indices	Storage (GB)	Indexing time (min)	Latency (ms)
Description only	1	14.8	48	15.8
Direct multimodal	1	12.6	41	14.1
Description + multimodal	2	27.4	89	29.9
QCR^2	2	27.4	90	30.0

C.4 Query-conditioned description diagnostic

The default textual descriptions are generated offline without access to the downstream query. As a diagnostic upper-bound setting, we additionally generate descriptions conditioned on the query, allowing the description generator to explicitly emphasize query-relevant information. As shown in Table C6, the gap between description-only retrieval and QCR^2 becomes smaller, as substantially more query-relevant information is directly exposed by the textual representation. This experiment provides a boundary case for our hypothesis rather than a practical retrieval setting, since query-conditioned descriptions cannot be precomputed during offline indexing.

C.5 Residual magnitude and normalization

QCR^2 normalizes the description-orthogonal component $v^\perp$ before retrieval such that residual similarity primarily reflects directional alignment with the query rather than residual magnitude. We compare the normalized residual score, $z_q^\top r_i$, with the score without normalization, $z_q^\top v_i^\perp$. For $\|v_i^\perp\|_2 \geq \epsilon$, the score without normalization can equivalently be written as

$$z_q^\top v_i^\perp = \|v_i^\perp\|_2 \cdot z_q^\top r_i, \tag{18}$$

and therefore additionally weights query alignment by residual magnitude.

The normalized residual improves Recall@5 from 47.2 to 48.3 and average downstream performance from 42.4 to 42.9. Although residual magnitude provides a useful indication of how much of the multimodal representation lies outside the description subspace, directly incorporating this magnitude into the retrieval score is less effective than using query alignment with the normalized residual direction. This result supports the design of QCR^2, which separates the existence of description-orthogonal variation from its relevance to the current query.

D Retrieval Efficiency

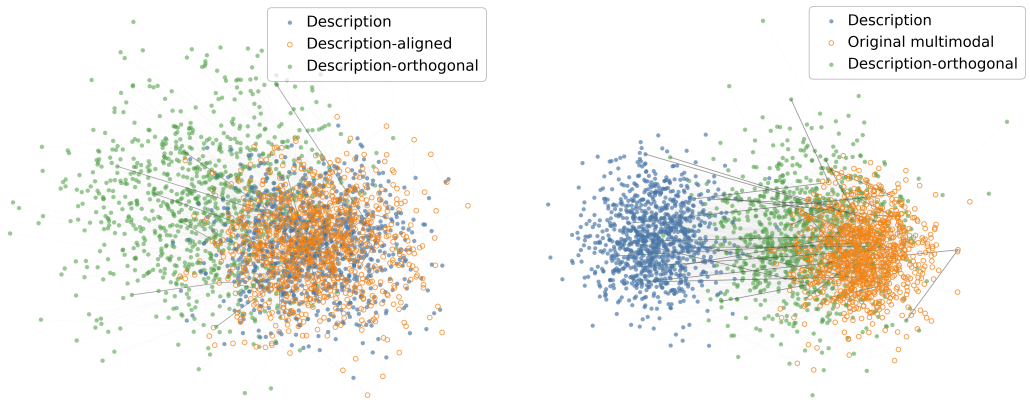
We evaluate the practical retrieval overhead of QCR^2 using the same FAISS HNSW configuration described in Section A. All representations are precomputed before evaluation, and all methods use the same embedding dimensionality and numerical precision. Experiments are conducted with four NVIDIA RTX A6000 GPUs for representation preprocessing.

Indexing and storage. Description-multimodal fusion and QCR^2 both maintain two indices. The former stores the description and original multimodal embeddings, whereas QCR^2 replaces the latter with the description-orthogonal residuals. Because the multimodal and residual representations have the same dimensionality, their storage requirements are effectively identical. The additional offline cost of QCR^2 therefore arises only from computing the description-conditioned decomposition before constructing the residual index.

Retrieval latency. As shown in Table D8, QCR^2 has nearly the same retrieval latency as description-multimodal fusion. Both methods perform two nearest-neighbor searches over equally sized repre-

Table E9: Residual-only supporting-evidence recovery across retrieval modalities. Recovery denotes the fraction of annotated supporting items retrieved by the residual index but not by the description index. The residual branch is not constructed for text and tables.

Retrieval modality	Residual-only recovery (%)
Text	–
Table	–
Image	12.8
Video clip	16.4
Video	18.1



(a) Description-aligned decomposition (b) Relation to original multimodal representation

Figure E2: **Representation-space visualization of the proposed decomposition.** We project 1,000 MRAG-Bench images and one randomly selected description per image using t-SNE. Light gray lines connect representations from the same item. (a) Description and description-aligned representations largely overlap, while description-orthogonal representations are broader and shifted. (b) Description and original multimodal representations are clearly separated, with description-orthogonal representations mainly occupying the intermediate region.

sentations followed by lightweight candidate-score fusion. Thus, the performance improvement of QCR^2 does not result from additional index capacity or a larger inference-time retrieval budget.

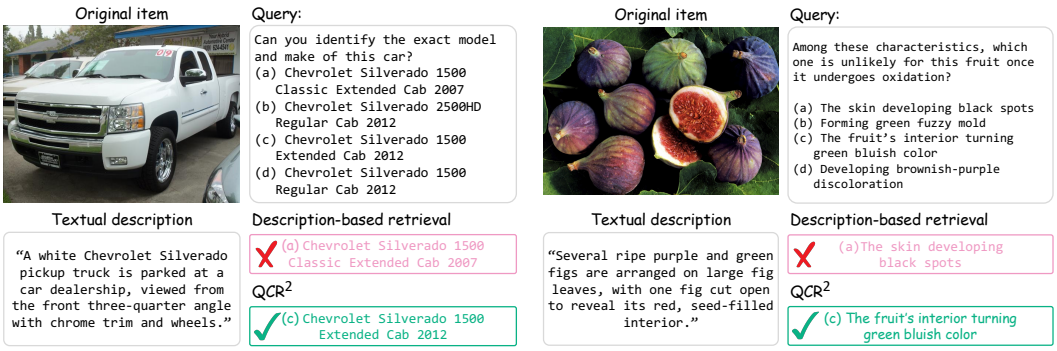
E Additional Analysis and Qualitative Results

E.1 Residual recovery across modalities

We further analyze residual-only supporting-evidence recovery separately across retrieval modalities. For each query with annotated supporting evidence, we measure the fraction of supporting items retrieved by $\mathcal{K}_{\text{res}} \setminus \mathcal{K}_{\text{desc}}$. As shown in Table E9, residual-only recovery occurs across all multimodal retrieval settings and is particularly frequent for video evidence, where query-independent descriptions can omit temporally localized details.

E.2 Representation visualization

In Figure E2, we further analyze how the proposed decomposition organizes the multimodal representation space. We randomly sample 1,000 images from the MRAG-Bench image corpus. For each image, we randomly select one of its four query-independent textual descriptions for visualization. We consider the textual-description representation, the original multimodal representation, and the two components obtained by our decomposition. All representations are jointly projected into two dimensions using t-SNE. Representations corresponding to the same image are connected by light gray lines. We emphasize that t-SNE is used only to visualize local neighborhood structure;



(a) Fine-grained vehicle identification (b) Fine-grained visual reasoning

Figure E3: **Successful examples of QCR^2.** Description-based retrieval misses query-relevant visual information, whereas the residual signal enables QCR^2 to recover the correct answer.

distances and angles in the projected space do not directly reflect those in the original embedding space, and the visualization should therefore not be interpreted as demonstrating orthogonality.

Description-aligned and description-orthogonal components. We depict the textual description, the description-aligned component $v^{\parallel}$, and the description-orthogonal component $v^{\perp}$ representations in Figure E2a. The description and description-aligned representations occupy largely overlapping regions, whereas the description-orthogonal representations show a substantially broader and shifted distribution. This behavior is consistent with the intended decomposition: the aligned component retains the part of the original multimodal representation represented within the description subspace, while the orthogonal component removes these directions. The resulting change in neighborhood structure indicates that the residual is not merely a perturbed copy of the description-aligned representation, but captures complementary variation that is suppressed when an item is represented only through its textual descriptions.

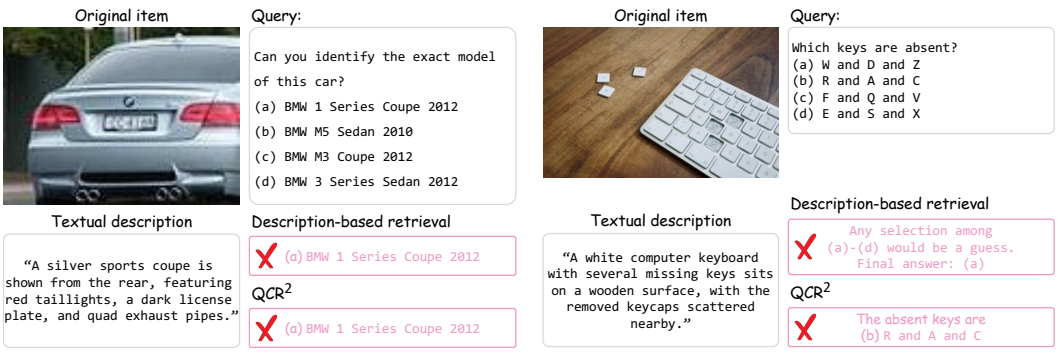
Relation to the original multimodal representation. Figure E2b compares the textual description, the original multimodal representation, and the description-orthogonal component. The textual and original multimodal representations form clearly distinguishable regions despite describing the same underlying images, revealing a substantial modality-dependent structure in the shared embedding space (Liang et al., 2022; Yeo et al., 2026). The description-orthogonal representations primarily occupy the region between two other representations and substantially overlap with the original multimodal representations. At the same time, the residual distribution extends toward the textual representation region rather than forming an isolated multimodal cluster.

Together, the two visualizations provide a complementary view of the role of the residual representation. The description-aligned component largely preserves the representation structure already exposed by textual descriptions, whereas the description-orthogonal component changes the neighborhood structure while remaining closely related to the original multimodal representation. This supports the use of the residual as a complementary retrieval signal rather than as a replacement for description-based retrieval.

E.3 Qualitative retrieval examples

Figure E3 presents two representative cases in which QCR^2 recovers information that is insufficiently represented by the textual description.

In Figure E3a, the textual description correctly identifies the item as a Chevrolet Silverado and captures coarse appearance such as its color, viewpoint, and chrome trim, but does not specify the exact vehicle variant. Description-based retrieval consequently selects an incorrect Silverado model, whereas QCR^2 identifies the Chevrolet Silverado 1500 Extended Cab 2012. This example illustrates the central motivation of QCR^2: a description can be semantically correct while remaining insufficient for a query requiring fine-grained visual distinctions.



(a) Query-irrelevant residual variation.

(b) Information not recoverable from the representation.

Figure E4: **Failure cases of QCR^2. (a)** Residual variation does not necessarily correspond to the information required by the query. **(b)** Information absent from both the textual description and the underlying multimodal representation cannot be recovered by residual retrieval.

A similar behavior appears in Figure E3b. The description captures the dominant visual content, *e.g.*, purple and green figs, their leaves, and the exposed red interior, but does not explicitly encode the visual characteristics needed to answer the oxidation question. Description-based retrieval selects an incorrect option, while QCR^2 recovers the correct answer through information preserved outside the description-aligned representation. These examples show that residual retrieval is particularly useful when the textual description preserves the overall semantics of an item but omits query-specific visual details. We provide an additional qualitative example with a text-only query in Figure E5.

E.4 Failure cases

Despite the complementary information provided by the residual representation, QCR^2 remains limited by the information encoded in the underlying multimodal representation and by the query relevance of the residual signal. Figure E4 illustrates these two failure modes.

Query-irrelevant residual variation. The residual is defined geometrically rather than as an explicit representation of omitted query-relevant semantics. Consequently, removing description-aligned directions can retain variations such as viewpoint, local appearance, background, or other fine-grained visual properties that are unrelated to the current question. In Figure E4a, the query requires discriminating among closely related BMW models. Both description-based retrieval and QCR^2 select the same incorrect model, showing that the additional residual signal does not necessarily provide the model-specific evidence required by the query. More generally, an irrelevant item can obtain high residual similarity when its description-orthogonal variation happens to align with the query.

Information absent from the underlying representation. Residual retrieval cannot reconstruct information that is not preserved by the multimodal encoder itself. The example in Figure E4b requires identifying the exact set of missing keys, a highly localized visual detail. The textual description provides no useful information about these keys, and QCR^2 also produces an incorrect answer. Since the residual is computed directly from the original multimodal representation, information discarded by that representation cannot be recovered through orthogonal decomposition. This failure therefore reflects an inherent limitation of the underlying representation rather than the retrieval rule itself.

An additional failure case with a text-only query is shown in Figure E6.

F Properties of the Residual Decomposition

We provide properties of the decomposition used by QCR^2.



Figure E5: **Successful retrieval example of QCR^2.** The query asks whether the National Museum of the American Indian and the Xanadu House have the same exterior color. Description-based retrieval identifies the relevant buildings but does not reliably preserve the queried color attribute. QCR^2 retrieves the required visual evidence through its complementary residual route, allowing the answer model to correctly determine that both buildings are beige.

Orthogonality to the description subspace. By construction,

$$U_i^\top v_i^\perp = 0. \tag{19}$$

Since every description embedding $t_i^{(m)}$ lies in $\mathrm{span}(U_i)$, it follows that

$$(t_i^{(m)})^\top v_i^\perp = 0 \qquad \forall m. \tag{20}$$

Thus, the residual removes directions represented by the descriptions of the same item. Importantly, this is an item-specific property: r_i need not be orthogonal to descriptions associated with other items.

Relation to the original multimodal representation. The two components are mutually orthogonal, *i.e.*, $(v_i^\parallel)^\top v_i^\perp = 0$, and therefore satisfy

$$\|v_i\|_2^2 = \|v_i^\parallel\|_2^2 + \|v_i^\perp\|_2^2. \tag{21}$$

Query:

Are the satellites on the Soviet space control/monitoring ship Kosmonavt Yuriy Gagarin always oriented in the same direction?

Description-only retrieval (Top-5)

1. "The Soviet space-control ship Kosmonavt Yuriy Gagarin is underway at sea."
2. "Yuri Gagarin appears at a press conference in Finland."
3. "Yuri Gagarin meets NASA astronauts at an international air show."
4. "Yuri Gagarin boards an aircraft in Sweden."
5. "A statue of Yuri Gagarin is unveiled in London."

✗ Answer: Yes

Multimodal-only retrieval (Top-5)



✗ Answer: Yes

QCR^2 (Top-5)

From textual relevance

1. "Space control-monitoring ship "Kosmonaut Yuri Gagarin" in 1987 (2).png A port beam view of the Soviet space control-monitoring ship KOSMONAVT YURIY GAGARIN underway."
2. "Yuri Gagarin (1961) - Restoration Photograph of the Soviet cosmonaut Yuri Gagarin at a press conference during his visit to Finland.
3. "Gemini 4 Astronauts Meet Yuri Gagarin Soviet cosmonaut Yuri Gagarin shakes hand with NASA's Gemini 4 astronauts, Edward H. White II and James A. McDivitt at the Paris International Air Show in June 1965."

From residual relevance

4. "Space control-monitoring ship "Kosmonaut Yuri Gagarin" in 1987.png A port view of the Soviet space control-monitoring ship KOSMONAVT YURIY GAGARIN underway."
5. "Research ship Meteor The German research ship Meteor in Bremerhaven."



✗ Answer: Yes

Figure E6: **Failure case of QCR^2.** The query asks whether the satellite dishes on the Soviet space control ship *Kosmonavt Yuriy Gagarin* are always oriented in the same direction. Answering correctly requires comparing multiple views of the ship. QCR^2 fails to retrieve sufficient complementary views within the final evidence set, illustrating that residual retrieval can still fail when the required evidence is distributed across multiple items.

For normalized v_i and a nonzero residual,

$$v_i^\top r_i = \|v_i^\perp\|_2. \tag{22}$$

Hence, the residual remains directly related to the original multimodal representation, with their similarity determined by the magnitude of the description-orthogonal component. This also explains why the residual representations in Figure E2b remain close to the multimodal population despite removing description-aligned directions.

Geometric residual versus query relevance. A nonzero residual does not necessarily indicate information useful for a particular query. The decomposition is purely geometric and should not be interpreted as explicitly identifying omitted semantic content. Conversely, query sufficiency of the textual descriptions does not require $v_i^\perp = 0$: an item may contain substantial description-orthogonal variation that is irrelevant to the current query. QCR^2 therefore evaluates the residual through its query-conditioned similarity, *i.e.*, $s_{\text{res}}(q, i) = z_q^\top r_i$. The residual becomes useful for retrieval only when the query aligns with directions that are not represented by the textual descriptions.

Degenerate residuals. When $\|v_i^{\perp}\|_2$ is very small, the original multimodal representation is almost entirely contained in the description subspace. The denominator $\max(\|v_i^{\perp}\|_2, \epsilon)$ prevents numerical instability in this regime. Accordingly, QCR^2 does not assume that every multimodal item contains a substantial complementary residual; the residual branch is useful only when such information exists and is relevant to the query.

G LIMITATION

The residual used by QCR^2 is defined geometrically in the representation space of a frozen multimodal encoder and should not be interpreted as an explicit decomposition of semantic information. Consequently, information omitted by a textual description can only be recovered when it is preserved by the underlying multimodal representation. The effectiveness of QCR^2 may therefore depend on both the quality of the retrieval encoder and the coverage of the generated descriptions. In addition, maintaining separate description and residual indices introduces additional storage and retrieval cost relative to description-only retrieval, although this cost is comparable to other two-representation fusion methods. Our current formulation also focuses primarily on non-text modalities for which distinct original and textual representations are available; extending the same principle to structured modalities and learned description spaces remains an interesting direction for future work.