WHERE TO LOOK, WHAT TO REVEAL: PRIVACY PRESERVING MULTIMODAL RAG WITH MULTIPLE AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern large language model (LLM) systems increasingly rely on centrally incorporating task-relevant external context at inference time, in addition to scaling model capacity. Retrieval-augmented generation (RAG) provides a general mechanism for such context augmentation, yet existing cloud-centric RAG systems often require fine-grained user data to be centrally accessible, creating unnecessary privacy exposure when personalized information is distributed across heterogeneous local sources. We propose PPM-RAG, a model-frozen multi-agent multimodal RAG framework that enables collaborative retrieval over distributed private data while keeping raw documents local. At its core, PPM-RAG introduces a central–local Federated Schema Directory (FSD), where local agent specialists organize modality-specific documents into hierarchical semantic directories and expose only compact root-level summaries and metadata to the central agent. During inference, the central agent decomposes a user query into subqueries and traverses the central FSD to identify relevant specialists. Selected specialists then perform hierarchical retrieval within their local FSDs and return only query-relevant evidence for final evidence-grounded reasoning. This design separates the discovery of *where relevant knowledge resides* from the disclosure of *what information is required*, enabling efficient and privacy-aware collaboration across distributed multimodal sources. Extensive experiments on WebQA and MMDocRAG against six baselines show that PPM-RAG improves retrieval Recall by 12.4% on average and answer quality by up to 32.9%, while reducing retrieval time by 90.6% compared with OpenRAG. PPM-RAG further reduces IIR by 10.8% and improves UMBRELA by 28.1%, demonstrating effective, efficient, and privacy-aware multimodal retrieval.

1 INTRODUCTION

Scaling model capacity has been a major driver of the rapidly improving capabilities of large language models (LLMs) Tom et al. (2020); Berti et al. (2025). Alongside model scaling, a complementary trend is to enhance LLMs by providing richer task-relevant context at inference time. In-context learning Dong et al. (2024); Liu et al. (2025) and agent memory Luo et al. (2026) can be viewed as following the same retrieval-and-contextualization paradigm exemplified by RAG: relevant external information is first selected and then incorporated into the model context to support generation. RAG formalizes this paradigm by retrieving task-relevant knowledge from external data sources and augmenting the generation context with the retrieved evidence Lewis et al. (2020). Thus, increasingly capable LLM systems depend not only on parametric knowledge, but also on retrieving the right external context for each request.

A common realization of such context augmentation is a cloud-centric agent architecture, where a centrally hosted LLM processes the user query together with retrieved or user-provided context Cheng et al. (2025); Ye & Luo (2025). While convenient, this paradigm is problematic for personalized applications, where the required context may contain private information distributed across multiple user devices and services Shao et al. (2024). As illustrated in Figure 1(a), answering "Which restaurant did I visit with Alice last month, and what did I order?" may require searching personal photos and receipts through a globally accessible retrieval space, even though only a

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.

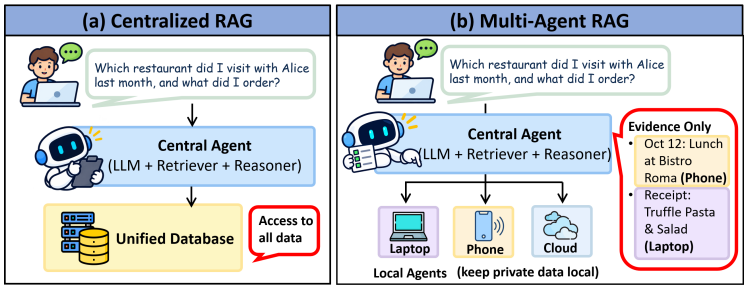


Figure 1: Comparison of (a) centralized RAG and (b) multi-agent RAG.

few pieces of information are actually needed for the answer. Making fine-grained personal data accessible to a central agent therefore introduces unnecessary data exposure and creates privacy concerns beyond the information required by the current query. Multi-agent RAG offer a promising alternative by distributing retrieval and reasoning across specialized agents Chang et al. (2025); He et al. (2025). However, existing approaches primarily focus on improving retrieval quality, agent collaboration, or federated model adaptation, rather than controlling what private information is revealed for a specific query Liang et al. (2026); He et al. (2025). This limitation becomes particularly important for personalized multimodal data: relevant knowledge may be distributed across images, documents, tables, and other modality-specific sources, while the central agent should acquire only the small amount of information necessary for answering the request. Figure 1(b) illustrates this desired behavior: local specialists retain their private data and disclose only query-relevant evidence, e.g., the restaurant identified from the phone and the ordered dishes identified from a receipt. The resulting challenge is therefore not only *which document to retrieve*, but first *where the relevant knowledge resides*, and then *what information should be revealed*.

To address these challenges, we propose PPM-RAG, a model-frozen multi-agent multimodal RAG framework based on a *Federated Schema Directory* (FSD). Each local agent specialist maintains one modality-specific data source and constructs a hierarchical local FSD over compact document schemas. Instead of sharing the underlying documents, only the summaries and metadata of local FSD roots are provided to the central agent and organized into a central FSD, yielding a coarse-to-fine directory of distributed knowledge. At inference time, the central agent decomposes the query and traverses the central FSD to identify relevant specialists; each selected specialist then retrieves within its local FSD and returns only query-relevant evidence for final reasoning. In this way, PPM-RAG enables efficient cross-agent multimodal retrieval while maintaining raw-document locality and avoiding unnecessary disclosure of unrelated local information.

Our contributions are summarized as follows:

- We propose PPM-RAG, a model-frozen multi-agent multimodal RAG framework that coordinates a central agent with modality-specific local agent specialists. PPM-RAG integrates offline Federated Schema Directory (FSD) construction with online multi-agent collaboration, enabling collaborative RAG over distributed personalized data without requiring the central agent to directly access the underlying local data.
- We design a central–local *Federated Schema Directory* (FSD) for efficient, locality-preserving retrieval. The FSD provides the central agent with a lightweight semantic directory of where relevant knowledge resides, accelerating retrieval while keeping fine-grained personalized data local.
- Building on the FSD, we develop an online multimodal RAG pipeline that integrates sub-query decomposition, hybrid FSD routing, hierarchical retrieval, and query-relevant evidence generation for efficient, privacy-aware multi-agent collaboration.
- Extensive experiments on WebQA and MMDocRAG against six baseline methods show that PPM-RAG improves retrieval Recall by 12.4% on average and answer quality by up to 32.9%, while reducing retrieval time by 90.6% compared with OpenRAG. It also reduces IIR by 10.8% and improves UMBRELA by 28.1%, demonstrating effective, efficient, and privacy-aware retrieval. We will release our code at https://github.com/papercode-DFL/PPMRAG to facilitate reproducibility.

2 Related Work

Existing methods for retrieval-augmented QA over large multimodal collections fall into two categories. *Fine-tuned methods* (Islam et al., 2024; Guo et al., 2025; Wang et al., 2022) update model parameters on task data, employing strategies such as sparse mixture-of-experts retrieval, graph-enhanced dual-level indexing, and unified vision–language sequence modelling. *Model-frozen methods* (Hu et al., 2022; 2025; Bu et al., 2025) require no task-specific training, instead coordinating retrieval and reasoning through in-context learning, adaptive planning graphs, or query-specific knowledge graphs. However, none accounts for pages distributed across private edge nodes rather than held in a single accessible collection.

Recent work extends retrieval-augmented question answering to settings where pages are distributed across edge nodes. FedMosaic (Liang et al., 2026) encodes client pages into parametric adapters and aggregates them selectively to prevent destructive merging, while pFedRAG (He et al., 2025) federates the embedding model via a shared encoder paired with a personalised per-client layer. Yet both operate within a single tier and offer training protocols rather than a hierarchical page index. Prior privacy-preserving RAG methods mainly protect the query, sanitize sensitive content, or secure the retrieval protocol. They do not explicitly optimize the amount of task-relevant information disclosed by heterogeneous private sources during iterative agent–source interaction.

3 Preliminaries and Problem Formulation

Multimodal RAG Multimodal RAG extends conventional RAG beyond text-only knowledge by retrieving relevant evidence from an external multimodal knowledge source, such as text and images, and incorporating the retrieved information into a MLLM for grounded response generation (Chen et al., 2022; Tian et al., 2025). Mathematically, given a query q and a multimodal document collection $\mathcal{D} = \{d_i\}_{i=1}^{|\mathcal{D}|}$, where $|\mathcal{D}|$ denotes the number of document, a retriever $\mathcal{R}$ identifies a set of relevant evidence $\mathcal{D}_q \subseteq \mathcal{D}$. A multimodal generator $\mathcal{G}$ then produces the answer $\hat{a}$ conditioned on both the query and the retrieved evidence:

$$\mathcal{D}_q = \mathcal{R}(q, \mathcal{D}), \qquad \hat{a} = \mathcal{G}(q, \mathcal{D}_q). \tag{1}$$

Problem Formulation We consider a multi-agent multimodal RAG system in which the multimodal database comprises N data sources, $\mathcal{D} = \{\mathcal{D}_i^m\}_{i=1}^N$, each containing documents from a single modality $m \in \mathcal{M}$, where $\mathcal{M}$ denotes the set of modalities, such as text, image, and table. In contrast to centralized agentic RAG systems, each data source is associated with a local agent specialist, denoted by $\mathcal{E} = \{E_i\}_{i=1}^N$, alongside the MLLM-based central agent G. Each agent specialist is built upon a multimodal SLM and retrieves relevant evidence from its local data source to facilitate the central agent in answering the query. Given a set of question–answer (QA) pairs $\mathcal{B} = \{(q, a)\}$ and a set of data source $\mathcal{D}$, our objective is to design a multi-agent collaboration strategy $\pi(\mathcal{E}, G)$ that improves downstream answer quality:

$$\arg \min_{\pi(\mathcal{E},G;\mathcal{D},\mathcal{B})} \sum_{(q,a)\in\mathcal{B}} U(a, \hat{a}_\pi) \tag{2}$$

$U(\cdot, \cdot)$ measures the discrepancy between the ground-truth answer a and the generated answer $\hat{a}$.

4 Methodology

Figure 2 illustrates PPM-RAG, a multimodal multi-agent RAG framework that coordinates agent collaboration through a novel central–local *Federated Schema Directory* (FSD) for multimodal retrieval and reasoning. We first introduce the offline FSD construction process, which consists of document-level schema extraction (DSE) over modality-specific local data sources, followed by schema aggregation to construct the local and central FSDs. We then present the online FSD-guided RAG process, which comprises four sequential stages. First, the central agent decomposes the user query into focused sub-queries to enable fine-grained and parallel retrieval. Second, it performs FSD-guided specialist routing by traversing the FSD to identify the specialists most relevant to each sub-query and dispatch the corresponding sub-queries. Third, each selected specialist retrieved its

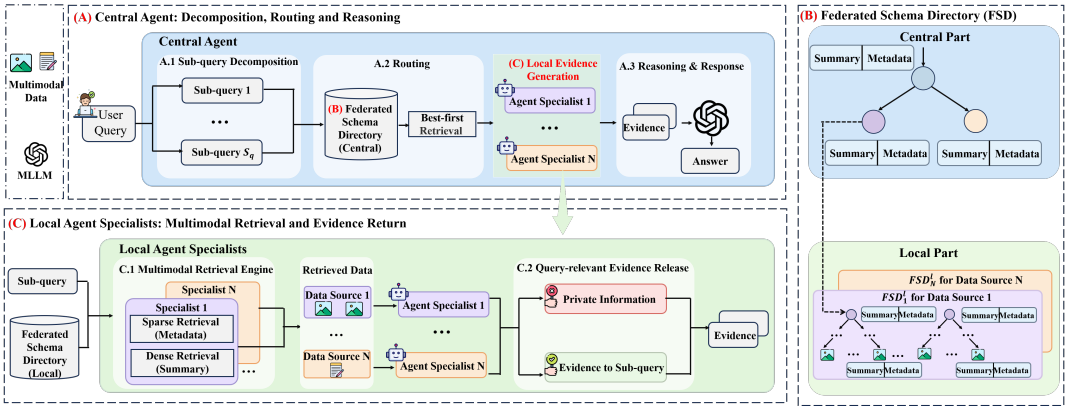


Figure 2: Overview of PPM-RAG.

modality-specific local data source and generates privacy-preserving evidence by retaining only information relevant to the assigned sub-query. Finally, the central agent aggregates the evidence returned by the specialists and performs evidence-grounded reasoning to generate the final answer.

4.1 Offline FSD Construction

Inspired by the principle that effective multi-agent collaboration requires sharing *where relevant knowledge resides*, rather than exposing *what each local data source contains* Shokouhi & Si (2011); Gravano & Garcia-Molina (1995), the central–local FSD is designed to enable efficient cross-agent retrieval without introducing privacy concerns associated with constructing a centralized index over the original multimodal documents. We first map heterogeneous documents into a unified retrieval-oriented schema space, hierarchically organize these schemas within each local specialist to construct local FSDs, and finally build the central FSD using only the roots of the local FSDs.

Document-level Schema Extraction (DSE) Consider the i-th modality-specific data source $\mathcal{D}_i^m$, which is exclusively accessible to local agent specialist E_i. To enable retrieval in a unified semantic space, we map each document $d_{i,j}^m \in \mathcal{D}_i^m$ into a compact textual schema that provides a semantically informative retrieval representation across modalities. Prior studies show that sparse lexical representations preserve precise term-level signals, whereas fixed-length dense representations capture broader semantic information but may lose fine-grained evidence, particularly for long documents, with their relative effectiveness further varying across domains Luan et al. (2021); Thakur et al. (2021). Motivated by these complementary properties, DSE represents each document through three retrieval-oriented views—*summary*, *keywords*, and *document-to-query (D2Q) generation*—capturing its global semantics, salient lexical cues, and potential information needs, respectively. Specifically, the multimodal SLM associated with E_i extracts

$$\sigma_{i,j} = (\mathcal{S}_{i,j}, \mathcal{K}_{i,j}, \mathcal{Q}_{i,j}), \tag{3}$$

where $\mathcal{S}_{i,j}$ is a semantic summary, $\mathcal{K}_{i,j}$ is a set of salient keywords, and $\mathcal{Q}_{i,j}$ is a set of synthetic queries for which the document may provide useful evidence. The three components of this compact textual schema triplet are generated separately by the multimodal local agent specialist. As shown at the bottom of Figure 2(B), the two lexical components are combined as document metadata, denoted by $\mathcal{T}_{i,j}$, serving as its sparse representation, while the semantic summary is embedded to obtain its dense representation, formulated as:

$$\underbrace{\mathcal{T}_{i,j} = \mathcal{K}_{i,j} \cup \mathcal{Q}_{i,j}}_{\text{sparse representation}}, \qquad \underbrace{e_{i,j} = f_i(\mathcal{S}_{i,j}) \in \mathbb{R}^h}_{\text{dense representation}}, \tag{4}$$

where $f_i(\cdot)$ is a text embedding model deployed at E_i and h is the embedding dimension. We instantiate f_i with a lightweight text embedding model, e.g., Qwen3-Embedding-0.6B (Zhang et al., 2025), deployed on-premise alongside the SLM of E_i.

Local FSD Construction To enable efficient retrieval while avoiding costly query-to-document comparisons Sarthi et al. (2024), each local FSD is organized as a collection of complete binary tree

hierarchies, where local documents are represented as leaf nodes. For specialist E_i, the local FSD is organized as a collection of complete binary trees:

$$\mathrm{FSD}_i^l = \{\mathcal{F}_i^1, \ldots, \mathcal{F}_i^{L_i}\}. \tag{5}$$

L_i denotes the number of trees maintained by specialist E_i, with each tree capturing a semantically coherent region of $\mathcal{D}_i^m$. We first partition the local document collection based on their dense representations $\{e_{i,j}\}_{j=1}^{|\mathcal{D}_i^m|}$. We apply an off-the-shelf unsupervised clustering algorithm, such as k-means (Faber et al., 1994) or a Gaussian mixture model (Reynolds et al., 2009), to divide the documents into L_i disjoint semantic groups $\mathcal{C}_i = \{\mathcal{C}_i^1, \ldots, \mathcal{C}_i^{L_i}\}$.

After semantic partitioning, we construct one balanced full binary FSD tree for each group, with all original documents placed at the leaf nodes. Starting from the leaf nodes, we build the tree bottom-up by repeatedly pairing the two nodes whose dense representations are most similar in cosine space. Each matched pair is then replaced by a new parent node, and the same procedure is applied recursively to the remaining set of nodes until a single root is obtained. For each parent node, the local SLM of specialist E_i generates a summary from the summaries of its two children. The parent keywords and D2Q metadata are obtained by merging and deduplicating the corresponding metadata from the children, and its dense representation is recomputed from the generated summary. Only the metadata and summary of the root nodes, $\{r_{i,j}\}_{j=1}^{L_i}$, where $r_{i,j} = \{\sigma_{i,j}^r\}$, are uploaded to the central agent for central FSD construction, while all intermediate nodes and original documents remain local to the specialist agent.

Central FSD Construction The central FSD provides the central agent with a coarse-grained view of the knowledge across local data sources without directly exposing their original documents. Specifically, we collect the uploaded root-node information from all local FSDs as

$$\mathcal{R} = \bigcup_{i=1}^{N} \{r_{i,j}\}_{j=1}^{L_i}, \qquad r_{i,j} = \{\sigma_{i,j}^r\}, \tag{6}$$

These root nodes are then used as the leaf nodes of a single central FSD, denoted by FSD^c, as illustrated by the dashed arrows in Figure 2(B).

We construct FSD^c directly from the uploaded local FSDroots, each of which contains a summary and its associated metadata. These root nodes serve as the leaf nodes of the central FSD. The central agent is equipped with the same lightweight embedding model used by the local specialists and independently encodes each root summary into the shared semantic space. Starting from the local FSD roots, we construct the central FSD bottom-up by repeatedly pairing the two nodes whose dense representations are most similar in cosine space. For each matched pair, the central agent's MLLM generates a new parent summary from the two child summaries. The parent metadata are obtained by merging and deduplicating the keywords and D2Q metadata of the children, after which the parent summary is encoded by the lightweight embedding model to obtain its dense representation. This process is repeated until all local roots are organized into a single balanced full binary tree.

4.2 Online FSD-Guided RAG

As shown in Figure 2(A) and (C), PPM-RAG performs online FSD-guided RAG through four stages. The central agent first decomposes the query into focused sub-queries and traverses the central FSD to route each sub-query only to relevant specialists, avoiding exhaustive retrieval across local data sources. Each selected specialist then retrieves evidence within its local FSD and returns only query-relevant information, keeping the original multimodal documents local. Finally, the central agent aggregates the returned evidence to generate an evidence-grounded answer.

Sub-query Decomposition Multimodal RAG tasks involve complementary evidence across heterogeneous modalities and data sources, which challenges a single query to capture all underlying information needs. Abootorabi et al. (2025); Jiang et al.; Ammann et al. (2025). To facilitate fine-grained specialist routing, the central agent decomposes the original query q into a set of focused sub-queries,

$$\mathcal{Q}_q = \{q_s\}_{s=1}^{S_q} = G^{\mathrm{dec}}(q), \tag{7}$$

where each q_s captures a specific information need that can be independently routed and retrieved. This decomposition reduces cross-modal retrieval ambiguity and enables complementary evidence from different specialists to be acquired in parallel.

FSD-guided Specialist Routing Given a sub-query q_s, the central agent traverses the central FSD to identify relevant local knowledge regions. We employ cosine similarity for dense retrieval (Reimers & Gurevych, 2019) and BM25 for sparse retrieval (Robertson & Zaragoza, 2009). For an FSD node v with dense representation $\mathbf{e}_v$ and metadata $\mathcal{T}_v$, we compute

$$\begin{aligned} S_d(q_s, v) &= \cos(f(q_s), \mathbf{e}_v)\,, \\ S_s(q_s, v) &= \mathrm{BM25}(q_s, \mathcal{T}_v)\,, \\ S(q_s, v) &= \lambda \frac{S_d(q_s, v) + 1}{2} + (1-\lambda)\frac{S_s(q_s, v)}{S_s(q_s, v) + \tau}, \end{aligned} \tag{8}$$

where λ balances dense and sparse relevance and τ calibrates the scale of BM25 scores. To avoid exhaustively evaluating the entire FSD, we adopt a best-first traversal strategy, following the classical principle of prioritizing the most promising node in a hierarchical index Hjaltason & Samet (1999). Specifically, the root is first inserted into a priority queue, and the node with the highest relevance score is iteratively selected for expansion. If the selected node is internal, its children are scored and inserted into the queue; if it is a leaf, its associated local FSD is added to the routing candidate set. The search terminates after collecting K_G candidate leaves or reaching an expansion budget B_G. We denote the resulting routing set as

$$\Omega_s = \mathrm{BestFirst}\left(q_s, \mathrm{FSD}^c; K_G, B_G\right), \tag{9}$$

where each $(i, l) \in \Omega_s$ identifies specialist E_i and its corresponding local FSD tree $\mathcal{F}_i^l$. The sub-query q_s is dispatched only to the specialists selected by Ω_s.

Privacy-preserving Evidence Generation After receiving a routed sub-query q_s, specialist E_i applies the same hybrid scoring and best-first traversal used in the central FSD to search the selected local FSD trees:

$$\mathcal{D}_{i,s}^* = \bigcup_{l:(i,l)\in\Omega_s} \mathrm{BestFirst}\left(q_s, \mathcal{F}_i^l; K_L, B_L\right), \tag{10}$$

where the retrieved leaves correspond to the original modality-specific documents. These documents remain local and are processed by the multimodal SLM of E_i to generate query-specific evidence:

$$e_{i,s} = E_i^{\mathrm{evi}}\left(q_s, \mathcal{D}_{i,s}^*\right). \tag{11}$$

Only $e_{i,s}$ is returned to the central agent. As the generated evidence is explicitly conditioned on q_s, it preserves query-relevant information while avoiding transmission of the original documents and unrelated local content.

Evidence-grounded Reasoning The central agent aggregates the query-relevant evidence returned by the selected specialists:

$$\mathcal{Z}_q = \{(q_s, e_{i,s}) \mid (i, l) \in \Omega_s,\ s = 1, \ldots, S_q\}. \tag{12}$$

Conditioned on the original query and the evidence collected from agent specialists, the central agent generates the final answer:

$$\hat{a}_\pi = G^{\mathrm{ans}}\left(q, \mathcal{Z}_q\right). \tag{13}$$

5 EXPERIMENT

We conduct comprehensive experiments on two multimodal datasets to evaluate the effectiveness of PPM-RAG and address the following research questions: RQ1: How does PPM-RAG compare with six baseline methods in terms of answer quality, retrieval effectiveness and retrieval efficiency? RQ2: How effectively does PPM-RAG prevent privacy leakage in privacy-sensitive scenarios? RQ3: How does PPM-RAG scale with increasing numbers of documents and local data sources? RQ4: How do the individual components, including FSD, SubQ, and BFS, contribute to the overall performance of PPM-RAG?

Table 1: Performance comparison of different methods on WebQA and MMDocRAG.

Method	WebQA					MMDocRAG				
	QA-Acc↑	QA-Flu↑	QA-All↑	Recall↑	Avg. Ret. Time↓	BLEU↑	RougeL↑	LLM-Score↑	Recall↑	Avg. Ret. Time↓
Qwen-3.5-2B										
Fine-tuned										
OpenRAG	0.464	0.525	0.311	0.738	0.082	0.091	0.186	0.475	0.484	0.087
LightRAG	0.486	0.529	0.336	0.854	0.184	0.097	0.203	0.497	0.554	0.182
OFA	0.573	0.581	0.384	0.888	0.752	0.109	0.231	0.550	0.599	0.778
Model-frozen										
PromptCAP	0.473	0.449	0.210	0.829	0.917	0.088	0.181	0.467	0.472	0.918
MMAPG	0.462	0.520	0.317	0.764	0.481	0.096	0.199	0.498	0.446	0.591
QDMG	0.538	0.513	0.311	0.782	0.209	0.101	0.212	0.502	0.489	0.214
Qwen3.5-9B										
Fine-tuned										
OpenRAG	0.535	0.609	0.407	0.856	0.107	0.103	0.223	0.580	0.581	0.108
LightRAG	0.501	0.612	0.396	0.884	0.186	0.106	0.229	0.584	0.579	0.193
OFA	0.654	**0.678**	0.512	0.890	0.756	0.142	0.259	0.647	0.601	0.781
Model-frozen										
PromptCap	0.495	0.543	0.266	0.843	0.934	0.098	0.205	0.499	0.485	0.926
MMAPG	0.529	0.600	0.407	0.833	0.601	0.110	0.230	0.592	0.509	0.728
QDMG	0.613	0.608	0.410	0.912	0.232	0.127	0.241	0.614	0.609	0.246
Ours	**0.663**	0.677	**0.539**	**0.932**	**0.006**	**0.150**	**0.272**	**0.656**	**0.612**	**0.010**

5.1 Experimental Setup

We consider PPM-RAG with $N = 8$ data sources, each storing 500 single-modality documents. We deploy Qwen3.5-2B Team (2026) as the on-device SLM backbones, while the cloud server uses Qwen3.5-9B Team (2026) for server-based the LLM backbone.

5.1.1 Datasets

We conduct experiments on two multimodal datasets, WebQA Chang et al. (2022) and MMDocRAG Dong et al. (2026). To evaluate answer quality, we use QA-Acc, QA-Flu, and QA-All for WebQA, and BLEU, ROUGE-L, and LLM-Score for MMDocRAG. For both datasets, we evaluate retrieval effectiveness and efficiency using Recall and Average Retrieval Time (Avg. Rec. Time), respectively. To evaluate privacy-preserving capability, we introduce the Irrelevant Information Ratio (IIR) to quantify the irrelevance of retrieved information to the original query and employ the LLM-based UMBRELA metric Upadhyay et al. (2024) to assess retrieval relevance. Detailed descriptions of dataset statistics and all evaluation metrics are provided in the **Appendix A.1 and A.2**.

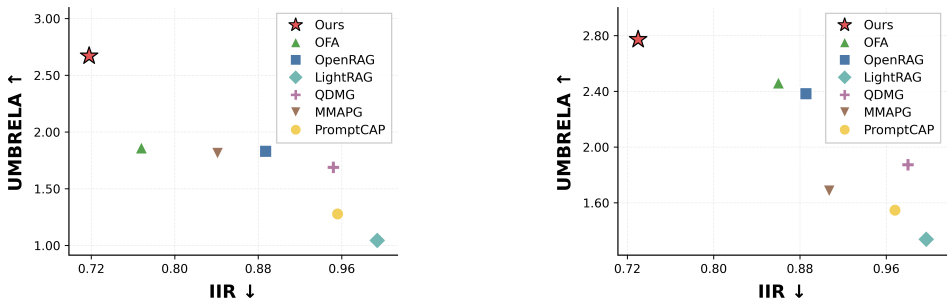
Dataset Statistic The main statistics of WebQA and MMDocRAG are summarized in Table 3. We sample 1,000 queries and 4,000 documents from each dataset, with equal numbers of text and image documents. WebQA contains only single-modal queries, whereas MMDocRAG includes 392 (39.2%) single-modal and 608 (60.8%) cross-modal queries requiring evidence from both text and image documents, making it a more challenging setting for cross-modal reasoning. On average, each query is associated with 1.75 and 2.71 gold documents in WebQA and MMDocRAG, respectively.

5.1.2 Baselines

We compare PPM-RAG with two categories of baseline methods: (1) **Fine-tuned methods**: OpenRAG Islam et al. (2024), LightRAG Guo et al. (2025), and OFA Wang et al. (2022); and (2) **Model-frozen methods**: PromptCap Hu et al. (2022), MMAPG Hu et al. (2025), and QDMG Bu et al. (2025). Detailed descriptions of the baseline methods and their experimental configurations are provided in **Appendix A.3**.

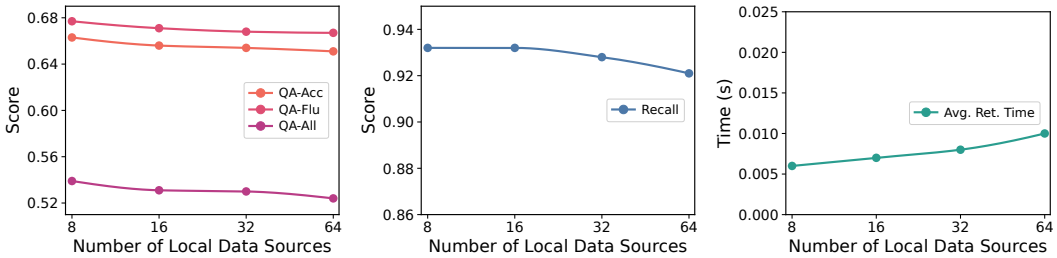
5.2 Overall Performance

To evaluate the effectiveness and efficiency of PPM-RAG, we compare it against six fine-tuned and training-free baseline methods using Qwen3.5-9B and Qwen3.5-2B, which serve within the PPM-RAG framework as the backbone models for simulating the central agent and local agent specialist, respectively. As shown in Table 1, PPM-RAG outperforms all baselines by 4.4%–32.9% in QA-All on WebQA, and by 1.3%–9.1% and 0.9%–18.9% in ROUGE-L and LLM-Score on MMDocRAG,



(a) WebQA (b) MMDocRAG

Figure 3: Performance comparison in terms of IIR and UMBRELA on two datasets. Lower IIR and higher UMBRELA indicate better performance.



(a) Answer quality. (b) Retrieval effectiveness. (c) Retrieval efficiency.

Figure 4: Scalability performance comparison of PPM-RAG

respectively. Benefiting from a best-first search strategy that jointly leverages dense and sparse retrieval signals, PPM-RAG improves retrieval recall by an average of 12.4% across the two datasets compared with the baselines. Moreover, the average retrieval time is reduced by 90.6% compared with OpenRAG, demonstrating the substantial efficiency gains of the complete-binary-tree-based PageIndex. Notably, reducing the base model from Qwen3.5-9B to Qwen3.5-2B consistently degrades baseline method performance, with QA-Flu on WebQA and BLEU on MMDocRAG decreasing by 14.4% and 14.5% on average, respectively, while yielding an average reduction of 9.6% in retrieval time.

5.3 Privacy-Aware Retrieval Evaluation

To evaluate the privacy-preserving capability of PPM-RAG, we measure the exposure of query-irrelevant information during retrieval. On average across WebQA and MMDocRAG, PPM-RAG reduces IIR by 10.8% and improves UMBRELA by 28.1% over the strongest baseline, OFA. These results demonstrate that PPM-RAG retrieves highly relevant context from local data sources while effectively reducing the exposure of unrelated information to the central agent. A detailed case study is provided in **Appendix A.4** to further demonstrate the superior privacy-preserving performance of our method compared with the baseline OFA. The results of impact of different foundation model backbones are provided in **Appendix A.5**.

5.4 Scalability

We investigate the scalability of PPM-RAG on WebQA by increasing the number of documents from 4,000 to 8,000, 16,000, and 32,000, corresponding to 8, 16, 32, and 64 participating local data sources, respectively. As shown in Fig. 4, PPM-RAG maintains stable answer quality as the document scale increases, with QA-Acc, QA-Flu, and QA-All decreasing by only approximately 1.8%, 1.5%, and 2.8%, respectively, from 8 to 64 local data sources. Retrieval effectiveness remains similarly stable, with Recall decreasing from 0.932 to 0.920, a relative reduction of only 1.3%. Notably, the average retrieval time increases only from approximately 0.006 s to 0.010 s, remaining at the millisecond level owing to the complete-binary-tree-based PageIndex design.

Table 2: Performance of PPM-RAG for ablation study

Method	WebQA				
	QA-Acc↑	QA-Flu↑	QA-All↑	Recall↑	Avg. Ret. Time↓
Ours	0.663	0.677	0.539	0.932	0.006
-w/o FSD_{SY}	0.615	0.648	0.499	0.901	0.005
-w/o FSD_{D2Q}	0.473	0.510	0.328	0.872	0.005
-w/o FSD_{KW}	0.581	0.656	0.484	0.893	0.006
-w/o SubQ	0.596	0.658	0.498	0.906	0.004
-w/o BFS	0.621	0.655	0.508	0.913	0.006
Method	**MMDocRAG**				
	BLEU↑	RougeL↑	LLM-Score↑	Recall↑	Avg. Ret. Time↓
Ours	0.150	0.272	0.656	0.612	0.010
-w/o FSD_{SY}	0.124	0.248	0.619	0.585	0.008
-w/o FSD_{D2Q}	0.108	0.207	0.598	0.558	0.008
-w/o FSD_{KW}	0.119	0.241	0.610	0.567	0.008
-w/o SubQ	0.121	0.243	0.624	0.568	0.007
-w/o BFS	0.137	0.251	0.628	0.584	0.010

5.5 Ablation Study

To validate the contribution of each key component in PPM-RAG, we construct five ablated variants, namely PPM-RAG w/o FSD_{SY}, PPM-RAG w/o FSD_{D2Q}, PPM-RAG w/o FSD_{KW}, PPM-RAG w/o SubQ, and PPM-RAG w/o BFS. Specifically, the first three variants remove the summary, doc2query, and keyword signals from the retrieval process, respectively; PPM-RAG w/o SubQ removes the divide-and-conquer query decomposition step and directly retrieves using the original query; and PPM-RAG w/o BFS replaces the best-first-search strategy with a beam-search-like layer-wise traversal. As shown in Table 2, PPM-RAG consistently outperforms all ablations on both WebQA and MMDocRAG. Averaged over the answer-quality and retrieval metrics, PPM-RAG achieves gains of 4.58%–36.03% on WebQA and 6.8%–22.4% on MMDocRAG over the five variants, demonstrating that each component contributes to the overall effectiveness of the framework. Among all ablations, removing FSD_{D2Q} causes the most severe degradation, reducing QA-All and Recall on WebQA from 0.539 and 0.932 to 0.328 and 0.872, respectively, and reducing Rouge-L, LLM-Score, and Recall on MMDocRAG from 0.272, 0.656, and 0.612 to 0.207, 0.598, and 0.558, respectively. This indicates that doc2query provides highly informative retrieval cues for bridging questions and candidate documents. Meanwhile, removing the summary or keyword signals also consistently hurts performance across both datasets, suggesting that these retrieval cues are complementary. In addition, removing SubQ leads to clear drops on both WebQA (e.g., QA-All: $0.539 \rightarrow 0.498$) and MMDocRAG (LLM-Score: $0.656 \rightarrow 0.624$), confirming the importance of decomposing complex questions into more manageable sub-questions for evidence localization. Finally, replacing BFS with a layer-wise beam-search-style traversal also degrades the results, particularly in Recall ($0.932 \rightarrow 0.913$ on WebQA and $0.612 \rightarrow 0.584$ on MMDocRAG), showing that the proposed best-first-search strategy is more effective in navigating the hierarchical retrieval space. Although some simplified variants slightly reduce the average retrieval time, these efficiency gains come at the cost of noticeably worse answer quality and retrieval effectiveness, highlighting that the full design of PPM-RAG offers the best overall trade-off.

6 Conclusion

We introduced PPM-RAG, a model-frozen multi-agent multimodal RAG framework for distributed personalized data. PPM-RAG employs a central–local Federated Schema Directory (FSD) to locate relevant knowledge without exposing fine-grained local data, and integrates sub-query decomposition, specialist routing, hierarchical retrieval, and privacy-aware evidence generation. Experiments on WebQA and MMDocRAG show that PPM-RAG improves answer quality and retrieval effectiveness while reducing retrieval latency and irrelevant information disclosure, demonstrating the effectiveness of separating knowledge discovery from information disclosure for efficient and privacy-aware multi-agent RAG.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin

Asgari. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16776–16809, 2025.

Paul JL Ammann, Jonas Golde, and Alan Akbik. Question decomposition for retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 497–507, 2025.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*, 2025.

Chenyang Bu, Guojie Chang, Zihao Chen, CunYuan Dang, Zhize Wu, Yi He, and Xindong Wu. Query-driven multimodal graphrag: Dynamic local knowledge graph construction for online reasoning. In *Findings of the association for computational linguistics: ACL 2025*, pp. 21360–21380, 2025.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, et al. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2607–2622, 2025.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16495–16504, 2022.

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 *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 5558–5570, 2022.

Yihang Cheng, Lan Zhang, Junyang Wang, Mu Yuan, and Yunhao Yao. Remoterag: A privacy-preserving llm cloud rag service. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 3820–3837, 2025.

Kuicai Dong, CHANG YUJING, Shijie Huang, Yasheng Wang, Ruiming Tang, and Yong Liu. Benchmarking retrieval-augmented multimomal generation for document question answering. *Advances in Neural Information Processing Systems*, 38, 2026.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, et al. A survey on in-context learning. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 1107–1128, 2024.

Vance Faber et al. Clustering and the continuous k-means algorithm. 1994.

Luis Gravano and Hector Garcia-Molina. Generalizing gloss to vector-space databases and broker hierarchies. In *VLDB*, volume 95, pp. 78–89, 1995.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tian Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. In *EMNLP (Findings)*, pp. 10746–10761, 2025.

Hangyu He, Xin Yuan, Kai Wu, Ren Ping Liu, and Wei Ni. pfedrag: A personalized federated retrieval-augmented generation system with depth-adaptive tiered embedding tuning. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Gísli R Hjaltason and Hanan Samet. Distance browsing in spatial databases. *ACM Transactions on Database Systems (TODS)*, 24(2):265–318, 1999.

Yiheng Hu, Xiaoyang Wang, Qing Liu, Xiwei Xu, Qian Fu, Wenjie Zhang, and Liming Zhu. Mmapg: A training-free framework for multimodal multi-hop question answering via adaptive planning graphs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14195–14211, 2025.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. Promptcap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

Shayekh Bin Islam, Md Asib Rahman, KSM Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. Open-rag: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 14231–14244, 2024.

Yigeng Jiang, Tingjun Su, Tong Wu, Shumeng Zhang, Yanxu Zhao, Zhaohong Huang, Tianyu Xie, Yuhang Wu, Yisheng Lin, Yuze Liu, et al. Divide, verify, and conquer: Verification-guided dag reasoning for large language models.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Zhilin Liang, Yuxiang Wang, Zimu Zhou, Hainan Zhang, Boyi Liu, and Yongxin Tong. Fedmosaic: Federated retrieval-augmented generation via parametric adapters. *arXiv preprint arXiv:2602.05235*, 2026.

Yuze Liu, Tingjie Liu, Tiehua Zhang, Youhua Xia, Jinze Wang, Zhishu Shen, Jiong Jin, Zhijun Ding, and Fei Richard Yu. Grl-prompt: Towards prompts optimization via graph-empowered reinforcement learning using llms' feedback. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 426–438. Springer, 2025.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345, 2021.

Jinghao Luo, Yuchen Tian, Chuxue Cao, Ziyang Luo, Hongzhan Lin, Kaixin Li, Chuyi Kong, Ruichao Yang, and Jing Ma. From storage to experience: A survey on the evolution of llm agent memory mechanisms. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 41622–41652, 2026.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Douglas A Reynolds et al. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663):3, 2009.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and trends® in information retrieval*, 4(1-2):1–174, 2009.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations*, volume 2024, pp. 32628–32649, 2024.

Yijia Shao, Tianshi Li, Weiyan Shi, Yanchen Liu, and Diyi Yang. Privacylens: Evaluating privacy norm awareness of language models in action. *Advances in Neural Information Processing Systems*, 37:89373–89407, 2024.

Milad Shokouhi and Luo Si. Federated search. *Foundations and trends® in information retrieval*, 5 (1):1–102, 2011.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

MiniCPM Team, Chaojun Xiao, Yuxuan Li, Xu Han, Yuzhuo Bai, Jie Cai, Haotian Chen, Wentong Chen, Xin Cong, Ganqu Cui, et al. Minicpm4: Ultra-efficient llms on end devices. *arXiv preprint arXiv:2506.07900*, 2025.

Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

Table 3: Dataset Statistics for WebQA and MMDocRAG.

Statistic	WebQA	MMDocRAG
Questions		
# Query	1000	1000
# Single-modal Qs (%)	1000 (100%)	392 (39.2%)
# Cross-modal Qs (%)	0 (0%)	608 (60.8%)
Avg. # Gold Docs / Q	1.75	2.71
Document		
# Documents	4000	4000
# Modalities	Text (50%) Image (50%)	Text (50%) Image (50%)

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Yang Tian, Fan Liu, Jingyuan Zhang, Yupeng Hu, Liqiang Nie, et al. Core-mmrag: Cross-source knowledge reconciliation for multimodal rag. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32967–32982, 2025.

Brown Tom, Mann Benjamin, Ryder Nick, Subbiah Melanie, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33(1):1877–1901, 2020.

Shivani Upadhyay, Ronak Pradeep, Nandan Thakur, Nick Craswell, and Jimmy Lin. Umbrela: Umbrela is the (open-source reproduction of the) bing relevance assessor. *arXiv preprint arXiv:2406.06519*, 2024.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pp. 23318–23340. PMLR, 2022.

Zipeng Ye and Wenjian Luo. Llms are privacy erasable. In *EMNLP (Findings)*, pp. 5672–5692, 2025.

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

A Appendix

A.1 Dataset Statistics

A.2 Descriptions of Metrics

Evaluation Metrics

- **QA-Acc, QA-Fl, and QA**: Following WebQA Chang et al. (2022), QA-Acc measures answer accuracy based on category-aware lexical matching, while QA-Flu evaluates answer fluency and semantic consistency using BARTScore. QA-Acc is the overall score computed from QA-Acc and QA-Flu.
- **BLEU, ROUGEL and LLM-Score**: Following MMDocRAG Dong et al. (2026), we use BLEU and ROUGE-L to measure the lexical similarity between generated and reference answers. Furthermore, LLM-Score aggregates five LLM-as-Judge dimensions: fluency,

citation quality, text-image coherence, reasoning logic, and factuality. Each dimension is normalized to $[0, 1]$, and their mean is reported as the final LLM-Score.

- **Recall**: Recall measures the proportion of gold documents successfully retrieved, reflecting the effectiveness of the retrieval process.
- **Average Retrieval Time (Avg. Ret. Time)**: We report the Avg. Ret. Time (s) per query to evaluate retrieval efficiency.
- **Irrelevant Information Ratio (IIR)**: IIR measures the proportion of retrieved information that is irrelevant to the input query, formulated as $1 - \mathrm{BLEU}$, where a lower value indicates less irrelevant information.
- **UMBRELA** Upadhyay et al. (2024): UMBRELA employs an LLM-based relevance assessor to evaluate the relevance between the query and retrieved information, with higher scores indicating stronger retrieval relevance.

A.3 Details of Baseline Methods

- OpenRAG Islam et al. (2024) transforms a dense open-source LLM into a parameter-efficient sparse MoE for enhanced reasoning over retrieved evidence, using 8 experts with Top-2 expert activation and 512-dimensional QLoRA adapters. It further introduces a hybrid adaptive retrieval strategy based on reflection tokens and model confidence, and ranks retrieved contexts using relevance, grounding, and utility scores with default weights of 1.0, 1.0, and 0.5, respectively.
- LightRAG Guo et al. (2025) constructs a graph-enhanced text index by extracting and deduplicating entities and relationships from document chunks, and introduces a dual-level retrieval paradigm that performs low-level retrieval over specific entities and high-level retrieval over broader relations and themes. It combines graph structures with vector-based keyword matching and one-hop neighborhood expansion to retrieve structurally relevant contexts, while supporting incremental knowledge-base updates without rebuilding the entire index. The chunk size is set to 1200, and the gleaning parameter is set to 1.
- OFA Wang et al. (2022) unifies vision, language, and cross-modal tasks within a task-agnostic and modality-agnostic sequence-to-sequence framework, representing images with ResNet-based patch features and text with BPE tokens in a shared multimodal representation space. It employs a Transformer encoder-decoder architecture without additional task-specific layers and jointly pretrains on publicly available image-text pairs across tasks including VQA, image captioning, visual grounding, image-text matching, and image infilling.
- PromptCAP Hu et al. (2022) introduces a query-aware image captioning framework that conditions caption generation on both the input image and the corresponding query, enabling the model to emphasize task-relevant visual details rather than generic image content. The method adopts 8-shot in-context learning without additional task-specific training.
- MMAPG Hu et al. (2025) introduces a training-free multimodal multi-hop QA framework that uses an Adaptive Planning Graph to dynamically coordinate planning, retrieval, and reasoning over multimodal evidence. The method utilizes 5-shot prompting for text retrieval and 6-shot prompting for targeted and descriptive image retrieval, without additional task-specific training.
- QDMG Bu et al. (2025) introduces an unsupervised framework that dynamically constructs query-specific local knowledge graphs and employs multi-path retrieval and iterative knowledge refinement for multimodal reasoning, with up to Top-25 candidates retained during retrieval.

A.4 Case Study for Privacy-Aware Retrieval Evaluation

A.5 Impact of Different Foundation Model Backbones

We further investigate the sensitivity of PPM-RAG to different foundation model backbones by evaluating three open-source multimodal model families under the same experimental setting:

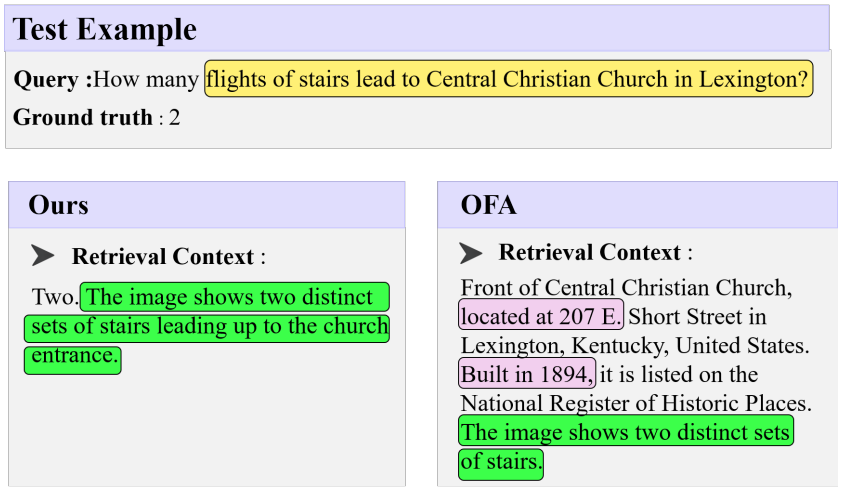
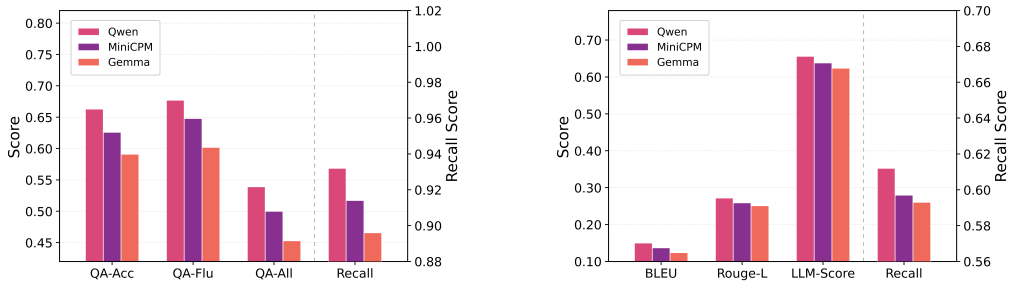


Figure 5: Example of retrieved contexts for a specific query. Our method retrieves contexts (green) that closely align with the query-relevant information (yellow), whereas the baseline OFA additionally retrieves irrelevant source-specific details, such as the exact address and year (pink), potentially leading to unintended leakage of edge-source information.



(a) Performance on WebQA

(b) Performance on MMDocRAG

Figure 6: Performance comparison of PPM-RAG under different base models on two datasets.

Qwen (Qwen3.5-9B for the central agent and Qwen3.5-2B for local agent specialist) Team (2026), MiniCPM (MiniCPM-o-4.5-9B and MiniCPM4-2B) Team et al. (2025), and Gemma (Gemma-4-E4B and Gemma-4-E2B) Team et al. (2026), respectively. As shown in Fig. 6, Qwen consistently achieves stronger performance than MiniCPM and Gemma on both WebQA and MMDocRAG. Specifically, Qwen improves the answer-quality metrics by an average of 9.47% over the other two backbones, whereas the average improvement in Recall is only 2.93% across the two datasets. The substantially smaller variation in retrieval performance suggests that answer quality is more sensitive to the choice of LLM backbone, while the retrieval capability of PPM-RAG remains comparatively stable. These results demonstrate that the retrieval mechanism of PPM-RAG is robust to backbone variations and can generalize effectively across different foundation model families.