Zero-shot Multimodal Document Retrieval via Cross-modal Question Generation

Anonymous ACL submission

Abstract

Rapid advances in Multimodal Large Language Models (MLLMs) have expanded information retrieval beyond purely textual inputs, enabling retrieval from complex real-world documents that combine text and visuals. However, most documents are private—either owned by individuals or confined within corporate silos—and current retrievers struggle when faced with unseen domains or languages. To address this gap, we introduce PREMIR, a simple yet effective framework that leverages the broad knowledge of an MLLM to generate cross-modal pre-questions (preQs) before retrieval. Unlike earlier multimodal retrievers that compare embeddings in a single vector space, PREMIR leverages preQs from multiple complementary modalities to expand the scope of matching to the token level. Experiments show that PREMIR achieves state-of-the-art performance on out-of-distribution benchmarks, including closed-domain and multilingual settings, outperforming strong baselines across all retrieval metrics. We confirm the contribution of each component through in-depth ablation studies, and qualitative analyses of the generated preQs further highlight the framework's robustness in real-world settings. The code will be publicly available.

1 Introduction

Advances in language models (Reimers and Gurevych, 2019) have enabled the creation of powerful retrievers that perform semantic search across documents, returning results closely aligned with user query (Karpukhin et al., 2020; Khattab and Zaharia, 2020). These retrievers are now widely deployed in real-world Retrieval-Augmented Generation (RAG) systems (Lewis et al., 2020), where they assist Multimodal Large Language Models (MLLMs) (Xu et al., 2025; Liu et al., 2023) by reducing hallucinations (Ayala and Bechard, 2024) and by supplying relevant context for evidence-guided answer generation (Jeong et al., 2024).

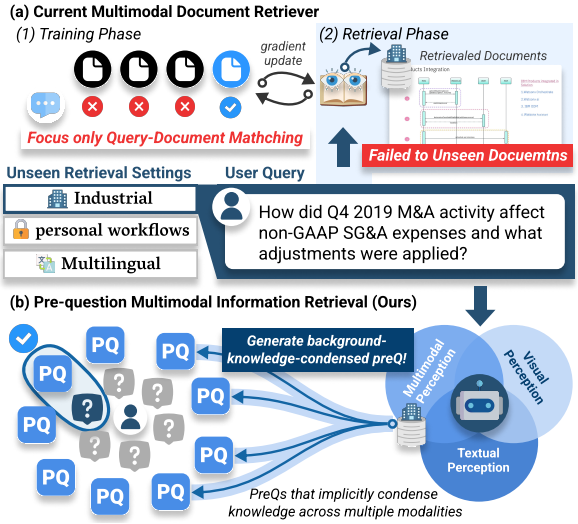

Figure 1: In unseen documents retrieval settings, (a) conventional latent-level contrastive learning approaches in multimodal retrievers struggle to generalize. In contrast, (b) PREMIR leverages token-level cross-modal complementary preQs to effectively handle such cases.

In conventional RAG systems, chunk-based text retriever is widely adopted (Liu et al., 2024b). However, this approach often overlooks crucial information such as images, tables, and document layout. Recent multimodal retrievers aim to address this limitation by extending retrieval capabilities to the visual domain—either by embedding both textual and visual elements using joint text–image encoders (Cao et al., 2019), or by leveraging MLLMs to compute page-level embeddings directly (Faysse et al., 2024; Yu et al., 2024). Despite these advancements, current multimodal retrieval systems still encounter challenges in real-world scenarios, such as in personal workflows or corporate settings.

Multimodal retrievers exhibit significant performance degradation in *out-of-distribution* settings, where documents often involve passages or images outside the scope of the training data. Existing multimodal models rely on directly comparing query

embeddings with page image embeddings, typically encoding the entire images into vectors. This training objective allows distinguishment of relevant from irrelevant images through maximizing the similarity with positives and minimizing with negatives (Mnih and Kavukcuoglu, 2013; Khattab and Zaharia, 2020). However, this contrastive training paradigm, focused on query-image alignment, often fails to learn a transferable latent space, resulting in poor generalization to *out-of-distribution* documents.

In addition, most existing methods treat MLLMs as static feature extractors, relying on fixed representations that overlook fine-grained cross-modal nuances. While some approaches (Nogueira et al., 2019a,b; Gospodinov et al., 2023) enrich document representations through query generation, they are limited to unimodal settings and remain highly dependent on the training distribution—further constraining their applicability in real-world scenarios.

To address these challenges, we propose the Pre-question Multimodal Information Retrieval (**PREMIR**) framework, which generates cross-modal complementary pre-questions (preQs) from documents by leveraging the extensive knowledge embedded in MLLM. These cross-modal preQs inherently capture comprehensive background knowledge and diverse contextual information, ensuring robust and effective performance even in challenging *out-of-distribution* scenarios, including multilingual and specialized closed-domain tasks. Furthermore, by representing queries at the token-level rather than as fixed feature vectors, the retriever is capable of capturing richer and more nuanced contextual details.

Experimental results demonstrate that PREMIR consistently outperforms strong baselines on multimodal document-retrieval tasks across both closed-domain and multilingual settings, establishing new state-of-the-art performance. Comprehensive ablation studies on each core module quantitatively confirm their individual contributions, and qualitative analyses offer intuitive insights into how our cross-modal preQs operate within the embedding space.

In summary, our contributions are three-fold:

1. We propose PREMIR, a multimodal retrieval method that mitigates domain shift without training by generating cross-modal preQs.
2. PREMIR achieves state-of-the-art performance on both multilingual and closed-domain benchmarks, demonstrating strong real-world applicability.
3. Comprehensive ablation studies and analysis demonstrate how cross-modal preQs significantly improve retrieval quality, offering insights into the mechanisms behind PREMIR's effectiveness.

2 Method

PREMIR framework aims to generate cross-modal preQs that comprehensively cover the documents' explicit and implicit knowledge from multimodal components, and retrieve the most appropriate semantically relevant preQs in response to a user query. In this section, we first outline the task definition in Section 2.1, and then describe the cross-modal preQs generation and retrieval process of the PREMIR framework in Section 2.2.

2.1 Task Definition

Problem Setting. In multimodal RAG scenarios, several key design choices must be made. First is the choice of input modalities in the system - text, images, or both. Second is the level of granularity used for retrieval such as entire documents, individual pages, chunks, or specific image regions. Since real-world data is inherently multimodal and often distributed across heterogeneous sources, we adopt a practical *out-of-distribution* configuration tailored for dynamic environments such as enterprise or personal workflows. In such settings, the corpus is typically domain-specific or multilingual, and the retrieval system must identify the most relevant passages (i.e., pages) from the entire document collection given an input text query.

Preliminary notations We denote the text query as q, and define the retriever as $\theta(\cdot)$, which searches for relevant passages $p_{i,j}$—the j-th passage (i.e., a page) from the i-th document—in the document corpus $\mathcal{C} = \{p_{1,1}, \ldots, p_{i,j}, \ldots\}$. Each passage may contain text and multimodal components such as tables, figures, or charts. The retriever operates over a passage pool $\mathcal{P}$, which typically corresponds to the entire corpus $\mathcal{C}$.

2.2 PREMIR Framework

Unlike approaches that treat a page as a single image (Faysse et al., 2024; Yu et al., 2024), our framework captures both the page and its fine-grained multimodal components—figures and their

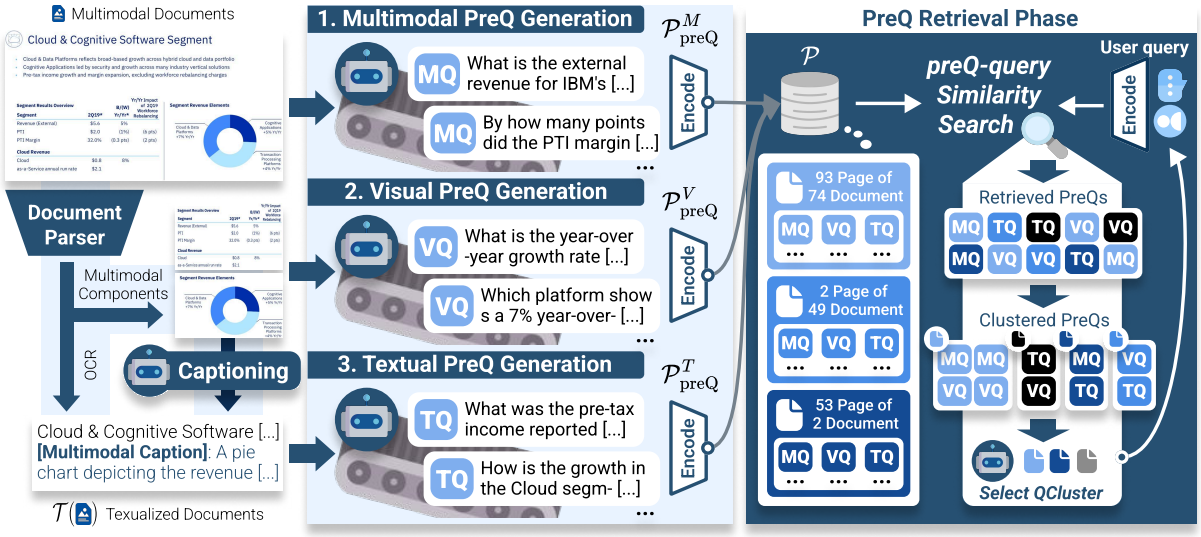

Figure 2: Overview of the PREMIR framework. PREMIR first parses multimodal content in a modality-aware manner and generates multimodal, visual and textual preQs, which are stored in a shared embedding space. During retrieval, the preQs most similar to the user query are retrieved. In here, Q-Cluster module then clusters these preQs by their source passages and returns the clusters whose passages are contextually aligned with the user query.

surrounding text—to extract richer cross-modal features. As illustrated in Figure 2, we first parse every document to extract both visual and textual components, and then generate cross-modal preQs from this enriched representation to ensure diversity and contextual relevance. We employ a powerful MLLM, GPT-4o (Hurst et al., 2024), and detailed prompts are provided in Appendix C.

Multimodal Document Parsing We employ a layout-aware document parser (Wang et al., 2024a) that fuses raw OCR output with grounded multimodal elements. For each page $p_{i,j}$, the parser returns the set of k detected multimodal components—tables, figures, charts, and so on—denoted by $p^{\text{mc}}_{i,j} = \{mc_1, \ldots, mc_k\}$, and the OCR text $p^{\text{ocr}}_{i,j}$.

Next, every component in $p^{\text{mc}}_{i,j}$ is captioned with MLLM, and these captions are merged with $p^{\text{ocr}}_{i,j}$ while preserving the original layout order. The result is a layout-aware textual surrogate $p^{\text{text}}_{i,j}$ that faithfully reflects the multimodal content of the page. Finally, the triplet $\langle p_{i,j}, p^{\text{mc}}_{i,j}, p^{\text{text}}_{i,j} \rangle$—comprising the raw page image, its component images, and the textual surrogate—is passed downstream for cross-modal preQs generation.

Cross-modal PreQ Generation Given a triplet $\langle p_{i,j}, p^{\text{mc}}_{i,j}, p^{\text{text}}_{i,j} \rangle$ for each page (i, j) in the corpus $\mathcal{C}$, we construct three complementary preQ sets:

(i) *Multimodal preQs*, $\mathcal{P}^{M}_{\text{preQ}}$, generated directly from the raw page image $p_{i,j}$ to preserve the original layout and cross-modal context;

(ii) *Visual preQs*, $\mathcal{P}^{V}_{\text{preQ}}$, created from individual visual components $p^{\text{mc}}_{i,j}$—such as figures, tables, and charts—to expose modality-specific cues; and

(iii) *Textual preQs*, $\mathcal{P}^{T}_{\text{preQ}}$, derived from the layout-aware textual surrogate $p^{\text{text}}_{i,j}$.

In conventional settings, the passage pool $\mathcal{P}$ from which the retriever selects candidate passages corresponds to the corpus $\mathcal{C}$. In contrast, we define the retrieval pool as the union of the three complementary preQ sets:

$$\mathcal{P} = \mathcal{P}^{M}_{\text{preQ}} \cup \mathcal{P}^{V}_{\text{preQ}} \cup \mathcal{P}^{T}_{\text{preQ}}. \quad (1)$$

Each set $\mathcal{P}^{*}_{\text{preQ}}$ is generated by MLLM (Hurst et al., 2024), which produces up to n questions per passage in the corpus. We fix $n = 50$ to strike a balance between performance and cost. These questions are designed to address not only information explicitly stated in the passage but also knowledge implicitly conveyed, as they are generated based on the broad knowledge of an MLLM.

Q-Cluster Retrieval. After constructing the retrieval pool $\mathcal{P}$, we embed each preQ using the retriever's embedding function. Given a user query q, the retriever encodes it into the same vector space and retrieves the top-k preQs from the retrieval pool $\mathcal{P}$ based on the highest cosine similarity of their embeddings.

A single preQ, however, may not fully capture the user's intent, and enumerating every possible query would require an impractically large and di-

		VIDoSeek				REAL-MM-RAG			
	Model	Recall@1	Recall@3	Recall@5	MRR@5	Recall@1	Recall@3	Recall@5	MRR@5
Text	E5	0.488	0.715	0.802	0.611	0.176	0.280	0.328	0.221
	GTE	0.416	0.617	0.715	0.529	0.175	0.276	0.320	0.229
	BGE-M3	0.473	0.712	0.790	0.596	0.168	0.267	0.317	0.232
	ColBERT	0.559	0.747	0.827	0.660	0.316	0.448	0.502	0.387
Image	VisRAG-Ret	0.644	0.841	0.912	0.752	0.281	0.439	0.502	0.365
	ColPali	0.670	0.852	0.907	0.764	0.398	0.571	0.639	0.490
	ColQwen2.0	0.743	0.912	0.944	0.827	0.452	0.622	0.688	0.543
PREMIR (Ours)		**0.801**	**0.919**	**0.954**	**0.863**	**0.501**	**0.673**	**0.724**	**0.590**

Table 1: Experimental results for the zero-shot closed-domain, multimodal document retrieval task on VidoSeek (Wang et al., 2025) and REAL-MM-RAG (Wasserman et al., 2025). The best results are **boldfaced**, and the second-best results are underlined.

verse set. To address this, we cluster preQs that were derived from the same source passage, thereby increasing the likelihood of retrieving a passage that directly answers the user's query.

We first cluster the top-k preQs originating from the same passage into a group $\mathcal{G}$. If the retrieval pool $\mathcal{P}$ contains more than $100k$ entries, we set $k = 100$; otherwise, we set $k = 150$. Collecting all such groups yields $\mathcal{S} = \{\mathcal{G}_1, \ldots, \mathcal{G}_m\}$. The LLM (Hurst et al., 2024) then evaluates each cluster in $\mathcal{S}$ according to how well its associated passage answers the query and selects the most relevant candidates. By leveraging these preQ clusters, we alleviate the need to generate preQs that exhaustively cover all possible variations of user intent.

3 Experiments

We evaluate PREMIR under realistic multimodal retrieval conditions encompassing (i) multimodal inputs, (ii) multi-document collections, and (iii) closed-domain or multilingual scenarios—settings commonly encountered in both personal and industrial applications. We first describe the experimental setup in section 3.1. We then assess PREMIR in closed-domain and multilingual environments, presented in section 3.2 and section 3.3, respectively. Additional details on the experimental setup and generated preQs are provided in Appendix A.3.

3.1 Evaluation Settings

Baselines. Within the multimodal retrieval task defined in section 2.1, we compare two categories of retrievers based on their input modality:

(1) Text-based. These models process passages only in textual form. To ensure a fair comparison, we provide the same parsed pages and VLM-generated captions introduced in section 2.2. The embedding-based retrievers included in our evaluation are E5 (Wang et al., 2022), GTE (Li et al., 2023), BGE-M3 (Chen et al., 2024), and the late-interaction model ColBERT (Khattab and Zaharia, 2020), which compute query and document token embeddings independently and match them during inference.

(2) Image-based. In contrast to text-based methods, these models embed each document page as an image. VisRAG-Ret (Yu et al., 2024) leverages a MiniCPM-V 2.0 (Yao et al., 2024) + SigLIP (Zhai et al., 2023) MLLM backbone, whereas ColPaLI and ColQwen2 (Faysse et al., 2024) adopt PaLI-3B (Chen et al., 2022) and Qwen2-VL-2B (Wang et al., 2024c) backbones, respectively; both use the ColBERT late-interaction scheme to match query–document pairs.

Metrics. We evaluate retrieval performance with two complementary metrics. Recall@k measures coverage—the fraction of relevant passages that appear among the top-k results; we report values for $k \in \{1, 3, 5\}$. MRR@k captures how early the first relevant passage is retrieved, using $k = 5$. Together, Recall@k and MRR@5 reflect both breadth and ranking precision.

3.2 Closed-domain Experiments

Setup. We evaluate two closed-domain benchmarks characterized by multimodal inputs and multi-document collections: (i) ViDoSeek (Wang et al., 2025) spans 12 topics—including economics, technology, literature, and geography—and contains 292 document decks with 5,385 passages and 1,142 queries. PREMIR generates 328k preQs for this benchmark. (ii) REAL-MM-RAG (Wasserman

	CT^2C-QA (Chinese)				Allganize RAG (Korean)			
Model	Recall@1	Recall@3	Recall@5	MRR@5	Recall@1	Recall@3	Recall@5	MRR@5
ColBERT	0.050	0.107	0.148	0.084	0.054	0.090	0.108	0.074
ColQwen2.0	0.126	0.228	0.295	0.185	0.565	0.748	0.813	0.659
PREMIR (Ours)	**0.256**	**0.397**	**0.474**	**0.335**	**0.763**	**0.874**	**0.910**	**0.820**

Table 2: Experimental results on the zero-shot multilingual, multimodal document-retrieval task for the Chinese benchmark CT^2C-QA and the Korean benchmark Allganize RAG. Owing to its looser structure, CT^2C-QA is markedly more challenging for baseline models than Allganize RAG.

$\mathcal{P}^M_{\text{preQ}}$	$\mathcal{P}^V_{\text{preQ}}$	$\mathcal{P}^T_{\text{preQ}}$	Recall@1	Recall@5	MRR@5
✓	✓	✓	**0.678**	0.916	**0.770**
✓	✓		0.672	**0.919**	0.770
✓		✓	0.672	0.909	0.764
	✓	✓	0.590	0.869	0.701
✓			0.652	0.913	0.755
	✓		0.397	0.588	0.471
		✓	0.568	0.858	0.680

Table 3: Ablation study results of multimodal preQs $\mathcal{P}^M_{\text{preQ}}$, visual preQs $\mathcal{P}^V_{\text{preQ}}$, and textual preQs $\mathcal{P}^T_{\text{preQ}}$ on the VidoSeek without preQ clustering.

et al., 2025) targets industrial scenarios, providing 162 documents—a mixture of financial reports and technical manuals—yielding 8,604 passages, and 4,553 queries. PREMIR produces 528k preQs for this benchmark.

Results. Table 1 demonstrates that in closed-domain multimodal retrieval, PREMIR outperforms all baselines on every benchmark without any additional multimodal retrieval training. Text-based models often struggle to capture the distinctive features of multimodal inputs, leading to sub-optimal performance. In contrast, image-based models generally perform better by overcoming some of these limitations; however, they still face challenges when handling unseen data in out-of-distribution documents scenarios. In our case, we leverage cross-modal preQs that implicitly condense knowledge across multiple modalities, enabling PREMIR to generalize effectively to previously unseen data and achieve strong performance. These results demonstrate that PREMIR generalizes robustly across both diverse personal topics and industrial corpora.

3.3 Multilingual Experiments

Setup. Following the closed-domain experiments reported in section 3.2, we use as baselines ColBERT and ColQwen 2.0—the highest-performing models for each input modality. We evaluate them on two public benchmarks: (i) CT^2C-QA(Zhao et al., 2024) is a Chinese question-answering dataset compiled from the National Bureau of Statistics of China. Only a sampled subset is publicly available, consisting of 400 single-page passages and 20,480 queries. PREMIR generates 58k preQs for this benchmark. (ii) Allganize RAG[1] is a Korean benchmark designed to evaluate RAG performance across domains such as finance, the public sector, healthcare, legal, and commerce. The publicly available dataset consists of 62 documents, resulting in 1289 passages and 278 queries. PREMIR produces 56k preQs for this dataset.

Results. Table 2 shows that PREMIR consistently outperforms all baselines across every dataset and metric in the multilingual setting. On the Chinese benchmark, the documents are loosely curated, creating a more realistic retrieval scenario in which both text- and image-based models struggle. Even under these conditions, PREMIR surpasses the strong baseline ColQwen2.0 by more than a factor of two in Recall@1. For the Korean benchmark, whose passages are comparatively well organized, ColQwen2.0 attains higher scores than it does on the Chinese; however, its performance still drops in the closed, multilingual context, whereas PREMIR maintains a clear lead. These findings imply that PREMIR generalizes robustly in multilingual closed-domain retrieval, reinforcing its suitability for real-world applications.

4 Analysis of PREMIR

4.1 Ablation Study

To investigate the effectiveness of PREMIR's core modules and to justify our design choices, we conduct ablation experiments in this section.

[1] https://huggingface.co/datasets/allganize/RAG-Evaluation-Dataset-KO

Model	Recall@1	Recall@5	MRR@5
text-embedding-3-large	**0.678**	**0.916**	**0.770**
bge-large-en-v1.5	0.603	0.886	0.713
gte-Qwen2-7B-instruct	0.576	0.878	0.691

Table 4: Comparison of retrieval performance using different embedding backbones on VideoSeek without PreQ clustering.

Model	Recall@1	Recall@5	MRR@5
PREMIR	**0.801**	**0.954**	**0.863**
- Qcluster	0.678	0.916	0.770

Table 5: Ablation study results for the process of clustering the retrieved preQs and selecting the cluster that best satisfies the query over the VidoSeek.

Model	Recall@1	Recall@5	MRR@5
GPT-4o	**0.801**	**0.954**	**0.863**
DeepSeek-V3	0.758	0.943	0.837
Qwen2.5$_{72B}$	0.762	0.933	0.834
Llama-3.3$_{72B}$	0.751	0.941	0.828
Qwen2.5$_{7B}$	0.736	0.928	0.813

Table 6: Impact of different LLMs in the Q-Cluster module on retrieval performance over the VidoSeek.

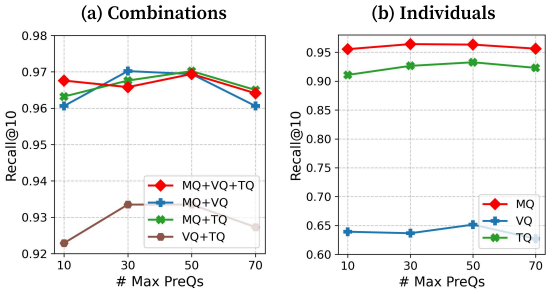

Figure 3: Ablation study on varying the number of generated preQs in ViDoSeek. (a) The left side shows results using combined multimodal, visual, and textual PreQs, while the (b) right side shows results using each modality individually.

Cross-modal PreQ Ablation. The results in Table 3 show that combining all three PreQ types achieves the best performance across all metrics. In particular, using the full set, yields the highest Recall@1 and MRR@5 scores, indicating that the three types complement each other. When used individually, $\mathcal{P}^{M}_{\text{preQ}}$ substantially outperforms both $\mathcal{P}^{V}_{\text{preQ}}$ and $\mathcal{P}^{T}_{\text{preQ}}$, underscoring the importance of preserving the original layout and cross-modal context in document understanding tasks.

Q-Cluster Ablation. Table 5 highlights the substantial performance gains obtained by introducing our Q-Cluster mechanism. Specifically, Q-Cluster increases Recall@1 from 0.6778 to 0.801, Recall@5 from 0.916 to 0.954, and MRR@5 from 0.770 to 0.863. This simple yet effective module helps the system prioritize passages that comprehensively address the user query, validating the value of this lightweight addition to the retrieval pipeline.

To assess the practical applicability of Q-Cluster, we replaced its backbone LLM with several alternatives and report the results in Table 6. PREMIR delivers consistent performance across all language models. While GPT-4o (Hurst et al., 2024) achieves the best scores, open-weight models such as DeepSeek-V3(Liu et al., 2024a), Qwen2.5-72B (Bai et al., 2025), Llama3.3-72B (Grattafiori et al., 2024), and even the compact Qwen2.5-7B suffer only minor degradation. These findings indicate that PREMIR can effectively leverage readily available open-weight models to achieve state-of-the-art multimodal document retrieval.

Embedding Model Ablation. Table 4 shows the results obtained with two open-weight embedding models—BGE (Chen et al., 2024) and the Qwen2-based GTE (Li et al., 2023). PREMIR delivers competitive retrieval quality even with these fully open embeddings, eliminating the need for proprietary solutions. Although the closed-weight baseline attains the highest overall score, the open-weight BGE and GTE variants remain close—especially in Recall@5, where they reach 0.886 and 0.878, respectively, versus 0.916. This narrow gap demonstrates that PREMIR maintains robust retrieval capability with widely accessible embeddings, making it practical for diverse deployment scenarios.

Hyperparameter Ablation. Figure 3 analyzes how the maximum number of generated preQs (n) influences retrieval performance. Recall@10 reaches its peak at n= 50 for most preQ variants, yet the drop is marginal when n is reduced to 10. In fact, Recall@10 remains 0.96 at n = 10, only slightly below the 0.97 achieved at the peak. These results show that generating fewer preQs can substantially cut computational cost with minimal loss in accuracy, which is especially valuable in resource-constrained settings.

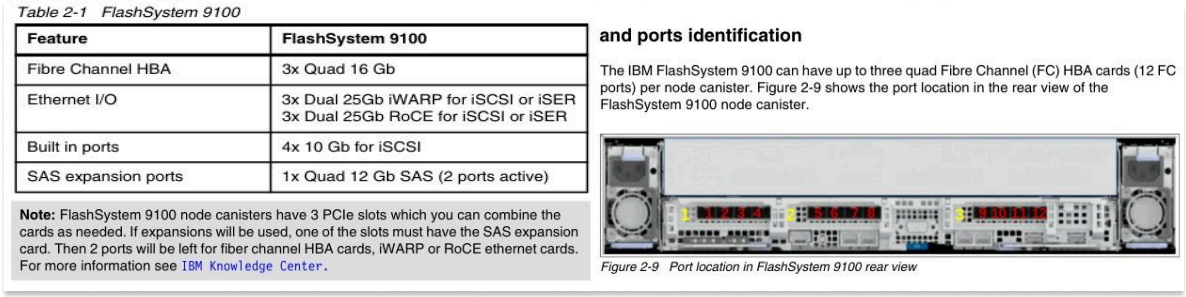

Table 2-1 FlashSystem 9100

Feature	FlashSystem 9100
Fibre Channel HBA	3x Quad 16 Gb
Ethernet I/O	3x Dual 25Gb iWARP for iSCSI or iSER 3x Dual 25Gb RoCE for iSCSI or iSER
Built in ports	4x 10 Gb for iSCSI
SAS expansion ports	1x Quad 12 Gb SAS (2 ports active)

Note: FlashSystem 9100 node canisters have 3 PCIe slots which you can combine the cards as needed. If expansions will be used, one of the slots must have the SAS expansion card. Then 2 ports will be left for fiber channel HBA cards, iWARP or RoCE ethernet cards. For more information see IBM Knowledge Center.

and ports identification

The IBM FlashSystem 9100 can have up to three quad Fibre Channel (FC) HBA cards (12 FC ports) per node canister. Figure 2-9 shows the port location in the rear view of the FlashSystem 9100 node canister.

Figure 2-9 Port location in FlashSystem 9100 rear view

PreQs	$\mathcal{P}^{M}_{\text{preQ}}$	What are the benefits of keeping the port count equal on each fabric as mentioned in the guidelines? What is the configuration of Ethernet I/O in the IBM FlashSystem 9100 as shown in Table 2-1?
	$\mathcal{P}^{V}_{\text{preQ}}$	What do the numbers labeled in red on the hardware represent? How many slots are visible in the hardware's front panel?
	$\mathcal{P}^{T}_{\text{preQ}}$	What is the total maximum count of Ethernet I/O connections available in the FlashSystem 9100? What are the benefits of keeping the port count equal on each fabric as mentioned in the guidelines?

Figure 4: Qualitative examples of multimodal, visual, and textual preQs generated from the passage above. The multimodal preQs capture the overall context of the document, while the visual and textual preQ focus on specific visual and linguistic details, respectively.

4.2 PreQ Triplets Analysis

Complementary Roles of Cross-modal PreQs. As illustrated in Figures 4 and 5, the three cross-modal preQ modalities—multimodal, visual, and textual—form a mutually reinforcing triad that expands both document-level coverage and embedding-space reach.

At the document level, multimodal preQs ($\mathcal{P}^{M}_{\text{preQ}}$) analyze the document holistically, integrating content, tables, and visual elements to generate questions focused on overall narrative flow and high-level semantics. Visual preQs ($\mathcal{P}^{V}_{\text{preQ}}$) specifically process image inputs, generating targeted questions tailored to visual content without encompassing the document's broader context. Textual preQs ($\mathcal{P}^{T}_{\text{preQ}}$) delve deeply into fine-grained linguistic aspects, such as entity mentions and definitions, providing detailed linguistic context.

In the embedding space, these complementary modalities enhance retrieval accuracy across diverse query types by occupying distinct regions. For instance, as demonstrated with `user query2`, which emphasizes specific visual elements, visual PreQs ($\mathcal{P}^{V}_{\text{preQ}}$) effectively address such queries by leveraging modality-specific features embedded within figures, and other visual components. This partitioning strategy ensures comprehensive document understanding and consistently improves retrieval performance across a wide range of query scenarios.

Improved Passage Discrimination. Figure 6 compares conventional query–passage retrieval with the PREMIR framework by examining the embedding-space distances between a user query and candidate passages. In conventional retrieval, embeddings of incorrect passages often lie close to the query as well as to the correct target pas-

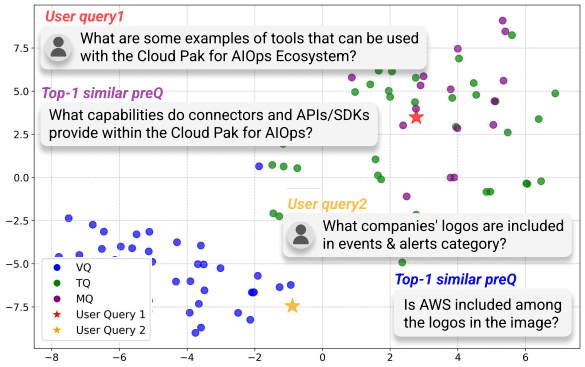

Figure 5: User query and cross-modal preQs in the embedding space visualized with t-SNE (van der Maaten and Hinton, 2008). The top-1 multimodal and visual PreQs are well aligned with the user's intent.

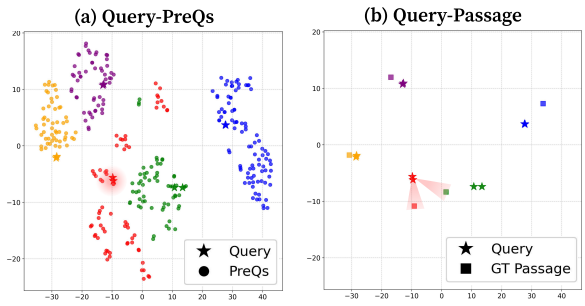

Figure 6: Comparison of query to preQ retrieval and query to passage retrieval. Objects of the same color represent the ground truth retrieval targets.

sage, making misretrievals more likely—an especially critical issue when the pool of relevant passages is small. By contrast, PREMIR alleviates this problem, its use of cross-modal preQs generates intermediate representations that carve out clearer semantic boundaries between passage clusters. As a result, embeddings of correct targets are more cleanly separated from those of confusable passages, leading to more reliable discrimination during retrieval.

5 Related Work

5.1 Multimodal Document Retrieval

Recent efforts to bridge the semantic gap between queries and documents have explored a range of approaches. Early dense retrieval methods such as DPR (Karpukhin et al., 2020) and ColBERT (Khattab and Zaharia, 2020) improved semantic matching by learning better text embeddings. In the domain of multimodal document understanding, cross-modal embedding techniques have driven significant progress. For example, LayoutLM (Xu et al., 2020) introduced a joint embedding of textual, visual, and layout information. This was further advanced by models like DocFormer (Appalaraju et al., 2021) and UDOP (Tang et al., 2023), which used transformer-based architectures to enhance multimodal integration. However, these early models suffered from the limited contextual understanding of complex real-world document layouts. Contemporary multimodal retrievers, such as DSE (Ma et al., 2024), ColPaLI (Faysse et al., 2024), and VisRAG-Ret (Yu et al., 2024), introduced unified input formats that preserve full document information without the need for content extraction. While these approaches offer improved generality, they still face challenges when handling out-of-distribution (OOD) documents. To address this, ColQwen2 leverage powerful vision-language models such as Qwen2-VL (Wang et al., 2024b) to directly encode documents as images, bypassing the need for separate modality handling. Building on such development, PREMIR has demonstrated stronger performance in both OOD scenarios—including closed-domain and multilingual benchmarks—and in handling complex, real-world multimodal documents without the need for training.

5.2 Applications of Query Expansion

Query expansion techniques have been applied to address challenges across various domains. In dialogue systems, query expansion has been employed to improve response relevance and contextual understanding. Expanding conversational queries with contextual information (Ni et al., 2023) significantly enhances dialogue coherence and response appropriateness. For domain-specific search, Peikos et al. (2024) applied query expansion to medical information retrieval, where bridging the gap between layperson terminology and professional medical vocabulary proved crucial. In legal document retrieval, Nguyen et al. (2024) showed how expansion techniques help navigate complex legal terminology and precedent relationships. In addressing vocabulary mismatch issue in information retrieval, Doc2query (Nogueira et al., 2019b) pioneered predicting potential queries for documents, while DocT5query (Nogueira et al., 2019a) refined this using T5's pre-trained knowledge. InPars (Bonifacio et al., 2022) leveraged LLMs for synthetic query generation. However, these approaches are limited to textual content and struggle with multimodal documents, failing to capture cross-modal interactions necessary for comprehensive document understanding and effective retrieval.

6 Conclusion

We introduced PREMIR, a powerful multimodal retrieval framework utilizing the broad knowledge of a MLLM to generate cross-modal preQs prior to retrieval. Unlike traditional multimodal retrieval methods limited by distribution-dependent training, our proposed cross-modal preQs implicitly condense information across modalities, enabling strong out-of-distribution retrieval performance. Remarkably, PREMIR achieves state-of-the-art results across all metrics under challenging out-of-distribution scenarios, including closed-domain and multilingual settings, without requiring additional training. Comprehensive ablation studies and analysis further demonstrate the effectiveness of cross-modal preQs in significantly enhancing retrieval quality, providing insights into the underlying mechanisms and highlighting the strong potential of PREMIR for real-world applications.

7 Limitations

PREMIR shows a limitation in consistently generating specific cross-modal PreQs using an MLLM. Despite explicit instructions, the model occasionally produces generic questions due to the subjective nature of "specificity." Fortunately, these generic PreQs have minimal impact on retrieval performance, as they are less likely to match user queries and rank low. Future work should focus on enhancing specificity—either by suppressing generic questions during generation or applying filtering mechanisms. Additionally, adaptive PreQ generation based on document complexity may improve efficiency by reducing computational costs.

References

Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. 2021. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 993–1003.

Orlando Ayala and Patrice Bechard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2387–2392.

Wenming Cao, Qiubin Lin, Zhihai He, and Zhiquan He. 2019. Hybrid representation learning for cross-modal retrieval. *Neurocomputing*, 345:45–57.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. 2022. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*.

Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. 2023. Doc2query–: when less is more. In *European Conference on Information Retrieval*, pages 414–422. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 39–48.

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

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in*

neural information processing systems, 36:34892–34916.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505.

Andriy Mnih and Koray Kavukcuoglu. 2013. Learning word embeddings efficiently with noise-contrastive estimation. *Advances in neural information processing systems*, 26.

Hai-Long Nguyen, Duc-Minh Nguyen, Tan-Minh Nguyen, Ha-Thanh Nguyen, Thi-Hai-Yen Vuong, and Ken Satoh. 2024. Enhancing legal document retrieval: A multi-phase approach with large language models. *arXiv preprint arXiv:2403.18093*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

Rodrigo Nogueira, Jimmy Lin, and AI Epistemic. 2019a. From doc2query to doctttttquery. *Online preprint*, 6(2).

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019b. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Georgios Peikos, Pranav Kasela, and Gabriella Pasi. 2024. Leveraging large language models for medical information extraction and query generation. *arXiv preprint arXiv:2410.23851*.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19254–19264.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, and 1 others. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Qiuchen Wang, Ruixue Ding, Zehui Chen, Weiqi Wu, Shihang Wang, Pengjun Xie, and Feng Zhao. 2025. Vidorag: Visual document retrieval-augmented generation via dynamic iterative reasoning agents. *arXiv preprint arXiv:2502.18017*.

Navve Wasserman, Roi Pony, Oshri Naparstek, Adi Raz Goldfarb, Eli Schwartz, Udi Barzelay, and Leonid Karlinsky. 2025. Real-mm-rag: A real-world multi-modal retrieval benchmark. *arXiv preprint arXiv:2502.12342*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200.

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

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Bowen Zhao, Tianhao Cheng, Yuejie Zhang, Ying Cheng, Rui Feng, and Xiaobo Zhang. 2024. Ct2c-qa: Multimodal question answering over chinese text, table and chart. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3897–3906.

A Implementation details

A.1 Benchmark Details

Since the CT^{2}C-QA dataset was not officially available, we utilized only 400 samples. For the Allganize dataset, we constructed our dataset by selecting only the documents that are practically available.

The distribution of multimodal, visual, and textual pre-questions across the datasets used in this study is summarized in Table 7.

A.2 PREMIR details

For retrieval, we used OpenAI's text embedding model *text-embedding-3-large* for query embedding. The LLM used in Qcluster is *gpt-4o*. MQ and VQ generation was done using *gpt-4o*, while TQ generation was done using *gpt-4o-mini*. For captioning the parsed components, we used *gpt-4o-mini*.

A.3 Experiments details

For the VisRAG-Ret experiments with colqwen and colpali, we used the NVIDIA RTX 3090 GPU.

B Limitations details

MLLM occasionally generates generic questions alongside specific ones. While these generic PreQs could potentially lead to incorrect retrieval, Figure 7 demonstrates that they rarely appear in top-K results when users submit specific queries. The retrieval process naturally filters out generic PreQs as they lack the distinctive characteristics to match well with specific user information needs. Therefore, despite the challenge of generating consistently specific PreQs, generic questions have minimal impact on overall system performance.

C Prompt

This section presents the prompts used throughout the image parsing and question generation pipeline. For image captioning, we refer to the prompt in Figure 9. For pre-question generation, we show the prompts used for multimodal pre-questions, visual pre-questions (Figure 10), and textual pre-questions (Figure 9). Additionally, the prompt used for Q-Cluster is also provided in Figure 11.

D Icon attribution

Icons used in the figures are from Flaticon https://www.flaticon.com. They are attributed to the respective authors as required by Flaticon's license.

Dataset	MQ	VQ	TQ	Total
VidoSeek	49,523	46,659	232,003	328,185
REAL-MM-RAG	107,137	90,433	384,352	581,922
CT^{2}C-QA	8,976	31,523	17,418	57,917
Allganize	8,979	7,625	39,177	55,781

Table 7: Statistics of question types in each dataset: MQ (multimodal preQs), VQ (visual preQs), and TQ (textual preQs).

Query: How did IBM's financial records reflect income tax obligations during the initial six months of 2014?

Generic PreQs	Retrieved Top-K PreQs
• What does this financial data show?	• What was IBM's provision for income tax in the second quarter of 2014?
• How did IBM perform financially in 2014?	• What is the provision for income tax reported by IBM for the first quarter of 2014?
• What was IBM's gross profit in 2014?	• How does IBM account for income taxes according to financial report for 2014?
• What were IBM's earnings during this period?	

Figure 7: Generated PreQ Limitation

You are given an image that represents part of a document, such as a figure, table, chart, or diagram.

Your task is to generate a clear, informative, and self-contained caption that describes:
1. What kind of image this is (e.g., chart, table, photograph, infographic) — provide a high-level description.
2. The detailed content within the image, including specific values, trends, comparisons, categories, or key insights, if applicable.

If the image contains a data visualization (e.g., a chart or table), describe the type of data, major trends, significant differences, or any notable patterns.

Avoid referring to the image as "this image" or using phrases like "shown here." Just write the caption as if it were placed directly below the image.

Figure 8: A prompt for generating captioned images during document parsing. The inputs of the prompts are **boldfaced** and image.

You are a helpful assistant for generating pre-questions based on a document.

Your task is to create "pre-questions" that a user might naturally ask **before** reading the document.

Each pre-question must satisfy the following conditions:

1. The question must be **specific and clearly formulated**, since it is asked *before* reading the document.
   – Do **not** use vague expressions like "this model", "in this document", or "According to the table".
   – Instead, **explicitly mention** the target of the question.
   – For example: "What is the performance of model A on dataset B?"

2. The question must have a **clear and verifiable answer within the document itself**.
   – Do not generate questions that cannot be answered using the document's content.

3. Generate up to **{cfg.max_new_questions}** questions.
   – All questions must be **diverse and non-redundant**.
   – Avoid repeating the same type of question or asking the same thing in different ways.

Output format:
– Return the questions as a JSON array of objects.
– Each object must follow this format:

```
{{
   "question": "string"
}}
```

- - -
Document:
{document_text}
- - -
Output:

Figure 9: A prompt designed to create both visual and multimodal pre-questions. The inputs of the prompts are **boldfaced**.

You are a helpful assistant for generating pre-questions based on an image-based document.

Your task is to create "pre-questions" that a user might naturally ask **before** reading this image-based document.

Each pre-question must satisfy the following conditions:

1. The question must be **specific and clearly formulated**, since it is asked *before* reading the document.
   – Do **not** use vague expressions like "this model", "in this document", or "According to the table".
   – Instead, **explicitly mention** the target of the question.
   – For example: "What is the performance of model A on dataset B?"

2. The question must have a **clear and verifiable answer within the document itself**.
   – The answer should be grounded in the document's content, including **multimodal elements** such as:
     - Figures (e.g., line graphs, bar charts)
     - Tables with numerical or categorical data
     - Diagrams, labeled illustrations, or structured visual layouts
   – Do not generate questions that cannot be answered using these visual or textual components.

3. Generate up to **{cfg.max_new_questions}** questions.
   – All questions must be **diverse and non-redundant**.
   – Avoid repeating the same type of question or asking the same thing in different ways.

Output format:
– Return the questions as a JSON array of objects.
– If the document contains no visual elements, return an empty list: `[]`
– Otherwise, format your output as a JSON array, where each object has the following structure:

```
[
  {
    "question": "string"
  }
]
```

Output:

Figure 10: A prompt designed to create both visual and multimodal preQs. The inputs of the prompts are **boldfaced** and image.

User query: **{query}**
Retrieved questions (grouped by source):
{questions_text}

Each question belongs to a source group (e.g., same document or generator). Some questions may be semantically similar because they come from the same source.

Please rank the TOP 5 source groups by how relevant and helpful their associated questions are for answering the user's query. Within each group, consider the best representative question to assess relevance.

Your goal is to select and rank the top 5 most useful groups such that the most useful ones are listed first, based on semantic similarity to the user's query.

IMPORTANT: Only include the 5 MOST RELEVANT group numbers in your ranking. If there are fewer than 5 groups total, include all of them.

Output only the group numbers in ranked order, separated by commas.
Example output: 2,1,4,3,5

Figure 11: A prompt used for Q-Cluster. The inputs of the prompts are **boldfaced**.