MR^2-BENCH: GOING BEYOND MATCHING TO REASONING IN MULTIMODAL RETRIEVAL

Junjie Zhou[1,2*] **Ze Liu**[3,2*] **Lei Xiong**[4,2] **Jin-Ge Yao**[2] **Yueze Wang**[2]
Shitao Xiao[2] **Fenfen Lin**[2] **Miguel Hu Chen**[2] **Zhicheng Dou**[4] **Siqi Bao**[5]
Defu Lian[3] **Yongping Xiong**[1] **Zheng Liu**[2†]

[1]Beijing University of Posts and Telecommunications
[2]Beijing Academy of Artificial Intelligence
[3]University of Science and Technology of China
[4]Renmin University of China [5]Baidu Inc., China
{junjiebupt, zhengliu1026}@gmail.com lz123@mail.ustc.edu.cn

ABSTRACT

Multimodal retrieval is becoming a crucial component of modern AI applications, yet its evaluation lags behind the demands of more realistic and challenging scenarios. Existing benchmarks primarily probe surface-level semantic correspondence (e.g., object–text matching) while failing to assess the deeper reasoning required to capture complex relationships between visual and textual information. To address this gap, we introduce MR^2-Bench, a reasoning-intensive benchmark for multimodal retrieval. MR^2-Bench presents the following critical values: 1) all tasks are reasoning-driven, going beyond shallow matching to effectively assess models' capacity for logical, spatial, and causal inference; 2) it features diverse multimodal data, such as natural images, diagrams, and visual puzzles, enabling comprehensive evaluation across content types; 3) it supports complex queries and documents containing multiple images and covers diverse retrieval scenarios, more accurately reflecting real-world applications. Our benchmark contains 1,309 curated queries, derived either from manual collection and annotation or from selective consolidation of public datasets. Despite achieving strong results on existing benchmarks, current state-of-the-art models still struggle on MR^2-Bench: for example, the leading Seed1.6-Embedding model attains a Recall@1 of 77.78 on MMEB, but only 9.91 on MR^2-Bench. This substantial performance gap highlights both the increased challenge posed by our benchmark and the pressing need for further advances in reasoning-intensive multimodal retrieval. The dataset and evaluation code will be made publicly available at https://github.com/VectorSpaceLab/MR2-Bench.

1 INTRODUCTION

Multimodal retrieval is a crucial capability in contemporary AI applications, supporting tasks such as image search (Young et al., 2014; Zhang et al., 2024), retrieval-augmented generation (RAG) (Chen et al., 2022; Yu et al., 2024), and multimodal agentic systems (Geng et al., 2025; Wu et al., 2025). The field has evolved from traditional cross-modal matching (e.g., text-to-image retrieval (Chen et al., 2015)) to more advanced multimodal retrieval that accommodates compositional queries over interleaved image-text content (e.g., composed image retrieval (Baldrati et al., 2023) and multimodal knowledge retrieval (Chang et al., 2022; Luo et al., 2023)). Consequently, modern multimodal retrievers (Zhou et al., 2024; Zhang et al., 2025a; Meng et al., 2025) can process queries expressed in text, images, or combinations thereof, efficiently extracting relevant information from diverse data sources and bridging the gap between complex datasets and real-world user needs.

*Co-first authors.
†Corresponding author.

Benchmarks	#Queries	#Tasks	Multi-Modality	Reasoning-Intensive	Vision-Centric Reasoning	Multi-Domain	Free-Form
MS MARCO (Bajaj et al., 2016)	5,193	1	✗	✗	✗	✗	✗
BEIR (Muennighoff et al., 2022)	54,262	18	✗	✗	✗	✓	✗
RAR-b (Xiao et al., 2024a)	45,745	17	✗	✓	✗	✓	✗
BRIGHT (Hongjin et al., 2025)	1,384	12	✗	✓	✗	✓	✗
CIRR (Liu et al., 2021)	4,148	1	✓	✗	✗	✗	✗
WebQA (Chang et al., 2022)	7,540	1	✓	✗	✗	✗	✗
M-BEIR (Wei et al., 2024)	190,000	10	✓	✗	✗	✓	✗
ViDoRe (Faysse et al., 2025)	3,810	2	✓	✗	✗	✗	✗
MMEB (Jiang et al., 2025)	36,000	36	✓	✗	✗	✓	✗
MR^2-Bench (Ours)	1,309	12	✓	✓	✓	✓	✓

Table 1: Comparison of MR^2-Bench with existing benchmarks. Columns report the number of test queries (**#Queries**); the number of tasks (**#Tasks**); inclusion of image–text data (**Multi-Modality**); whether the benchmark is explicitly reasoning-focused (**Reasoning-Intensive**); whether it contains tasks solvable purely from images without textual cues (**Vision-Centric Reasoning**); domain coverage (**Multi-Domain**); and support for arbitrary text–image organization—interleaved ordering and multi-image on the query and document sides (**Free-Form**). The first block represents textual retrieval benchmarks, and the second block represents multimodal retrieval benchmarks.

Despite these advances, current evaluation methods remain misaligned with practical requirements. First, existing benchmarks primarily assess surface-level semantic correspondence, offering limited coverage of knowledge reasoning, spatial perception, and vision-centric challenges critical for diverse agentic applications. Second, these benchmarks predominantly feature natural images, with insufficient representation of visual puzzles, diagrams, and mathematical figures common in technical and educational contexts. Third, real-world documents often exhibit free-form, interleaved image-text layouts with multiple images positioned arbitrarily within the text. However, current benchmarks frequently limit each example to a single image (Chang et al., 2022; Baldrati et al., 2023; Hu et al., 2023; Jiang et al., 2025), failing to reflect the complex document structures prevalent in practice. These limitations hinder rigorous evaluation of multimodal retrieval systems in reasoning-intensive, real-world scenarios.

In this paper, we introduce **MR^2-Bench** (Multimodal Reasoning-intensive Retrieval Benchmark). We summarize the key features of MR^2-Bench compared to existing benchmarks in Table 1. In summary, MR^2-Bench presents the following critical advantages:

- **It is the first benchmark for multimodal reasoning-intensive retrieval.** MR^2-Bench is pioneering in its requirement for reasoning to capture relevance rather than relying on shallow semantic matching, thereby filling a significant gap in current multimodal retrieval benchmarks. While existing text-only reasoning-intensive retrieval benchmarks (Xiao et al., 2024a; Hongjin et al., 2025) have been developed, MR^2-Bench emphasizes multimodal capabilities with a variety of visually related reasoning-intensive retrieval tasks.
- **It introduces a broad range of multimodal data domains.** Beyond typical natural images, MR^2-Bench incorporates diverse image types such as mathematical visual proofs, visual puzzles, and economic charts, etc. These images have widespread applications and inherently require visual reasoning capabilities. However, previous multimodal retrieval tasks have largely overlooked these data types.
- **It offers diverse evaluation scenarios.** MR^2-Bench encompasses three meta-tasks: multimodal knowledge retrieval, visual illustration search, and visual relation reasoning, totaling 12 sub-tasks. These tasks provide a wide array of retrieval scenarios, including text-to-image, image-to-image, and mixed image-text queries, among others. Moreover, unlike previous multimodal benchmarks where queries or documents typically contain at most a single image (Wei et al., 2024; Jiang et al., 2025), both queries and documents in MR^2-Bench may include multiple images, more accurately reflecting real-world scenarios.

We conduct comprehensive evaluation experiments on existing methods and derive the following key conclusions. Firstly, *multimodal reasoning-intensive retrieval remains challenging for current retrievers*. Despite Seed1.6-Embedding (Seed, 2025) achieves the best performance on MR^2-Bench, it only reaches 30.68 nDCG@10. In contrast, it attains 77.78 Recall@1 on the MMEB dataset (Jiang

et al., 2025), while its MR^2-Bench Recall@1 is just 9.91. Consistent failures are observed across all methods, particularly in mathematical visual proofs and visual relation reasoning. Secondly, *the capability of visual understanding plays an important role in solving our benchmark*. On the one hand, augmenting text-only retrievers with image captions yields substantial gains compared to ignoring images. On the other hand, despite current multimodal retrievers not being optimized for reasoning-intensive retrieval, the two strongest methods in our evaluation are native multimodal retrievers. Finally, *reasoning capacity holds significant potential for enhancing performance on MR²-Bench*. We implement reasoning-enhanced strategies including query rewriting and reranking, which have demonstrated substantial improvements on MR²-Bench. These insights highlight the challenges and opportunities in multimodal retrieval. By exposing current strengths and weaknesses, we anticipate that MR^2-Bench will guide the development of more capable multimodal retrievers.

2 Related Work

Reasoning-intensive Retrieval. Information retrieval (IR) has advanced from lexical matching (Robertson et al., 2009) to capturing deep semantic relevance (Karpukhin et al., 2020; Xiao et al., 2024b; Zhang et al., 2025b). Recently, the rise of applications like retrieval-augmented generation and agentic systems (Li et al., 2025b; Jin et al., 2025; Qian & Liu, 2025) has spurred the need for a more advanced capability: reasoning-intensive retrieval. This paradigm challenges IR systems to address complex information needs where relevance cannot be determined by direct semantic overlap, but must be inferred through deep reasoning. Although there has been significant progress in text-only domains with pioneering benchmarks such as BRIGHT (Hongjin et al., 2025) and the development of specialized retrievers (Shao et al., 2025; Long et al., 2025), its application to multimodal scenarios remains largely unexplored. Our work addresses this gap for the first time. Beyond knowledge-oriented tasks, we introduce novel, vision-centric challenges, including visual illustration search and visual relational reasoning, requiring models to perform complex inference over integrated visual and textual data.

Multimodal Retrieval. As real-world information is increasingly presented in multimodal formats, multimodal retrieval has become essential for effectively searching corpora that integrate text and visual data. Initially, the focus was on cross-modal retrieval, such as text-to-image searches (Chen et al., 2015). The field has since evolved to tackle more complex tasks, including image searches guided by textual instructions (Wu et al., 2021; Zhang et al., 2024), multimodal document retrieval (Chang et al., 2022), and knowledge retrieval using multimodal queries (Luo et al., 2023). With the advent of powerful pre-trained vision-language models (VLMs), researchers have been able to develop unified embedding models that effectively handle queries and documents in various formats (Lin et al., 2024; Zhou et al., 2025). Despite these advances, existing benchmarks and methods have largely concentrated on shallow semantic alignment or instance-level matching, neglecting the complex reasoning required to address many real-world information needs (Wei et al., 2024; Jiang et al., 2025). Moreover, these benchmarks often emphasize natural images, overlooking visually complex and abstract domains that demand visual-centric reasoning abilities, such as visual puzzles, mathematical diagrams, and multi-image relational scenarios. Consequently, there is a pressing need for a benchmark designed to evaluate deeper reasoning capabilities in multimodal retrieval.

3 MR^2-Bench: Multimodal Reasoning-intensive Retrieval Benchmark

We propose MR^2-Bench, the first multimodal reasoning-intensive retrieval benchmark. A brief overview of MR^2-Bench's statistics is presented in Table 2, and visual examples for each task type are shown in Figure 1. MR^2-Bench comprises 3 meta-tasks and 12 sub-tasks, encompassing a total of 1,309 queries. Detailed modalities of queries and documents, along with the instructions for each sub-task, are provided in Appendix B.

3.1 Multimodal Knowledge Retrieval

Traditional knowledge retrieval has focused primarily on text-only queries and corpora (Chen et al., 2017; Kwiatkowski et al., 2019). However, images play a crucial role in realistic knowledge retrieval scenarios. For instance, when users wish to explore an intriguing scientific phenomenon in

Meta-task	Multimodal Knowledge Retrieval						Visual Illustration Search			Visual Relation Reasoning		
Sub-task	**Biology**	**Cooking**	**Gardening**	**Physics**	**Chemistry**	**EarthScience**	**Economics**	**Mathematics**	**Nature**	**Spatial**	**VisualPuzzle**	**Analogy**
#Queries	79	76	129	76	124	99	84	86	100	149	160	147
#Corpus	4,455	2,786	5,636	6,656	4,317	3,014	7,572	944	2,017	1,000	5,375	3,970

Table 2: Data statistics of queries and corpus for each sub-task in MR^2-Bench

their daily lives, capturing an image for querying is often more intuitive and detailed than using text alone. Similarly, knowledge bases frequently integrate text and images, with images providing essential explanatory and knowledge representation functions. Although some benchmarks have been developed for multimodal knowledge search (Chang et al., 2022; Luo et al., 2023; Hu et al., 2023; Chen et al., 2023), they are predominantly based on annotations from sources like Wikidata, with questions that are often straightforward (e.g., *What is this mountain called?*[1]). These tasks typically rely on keyword matching, image instance matching, or simple shallow semantic alignment. However, real-world user queries can be highly complex, requiring intensive reasoning to identify relevant documents.

BRIGHT (Hongjin et al., 2025) introduced the first benchmark for evaluating reasoning-intensive knowledge retrieval by constructing retrieval pairs between real user queries from Stack Exchange[2] and relevant documents. The relevant documents are identified from external links referenced in high-scoring answers, establishing retrieval relationships that require reasoning over critical concepts or theories to bridge the query and the document. As a result, retrieval models evaluated on this benchmark must possess capabilities that go beyond simple lexical or semantic matching. However, BRIGHT is a text-only benchmark, leaving a gap in multimodal queries and documents.

Inspired by BRIGHT's task construction approach, we have developed a set of reasoning-intensive multimodal knowledge retrieval tasks in our MR^2-Bench. In contrast to BRIGHT, our approach rigorously ensures that images are essential components of the questions, rendering these inquiries invalid without the accompanying visual data. We also retain images from relevant documents if they are crucial for conveying knowledge. The annotation process is detailed in the Appendix C. Our benchmark covers six domains: **Biology**, **Cooking**, **Gardening**, **Physics**, **Chemistry**, and **Earth Science**. Examples of these tasks are illustrated in Figure 1(a)-(c). For instance, in Figure 1(a), the positive document does not mention *apple* or *grow together*. The key to connecting the document and the question lies in the accompanying image, which demonstrates a similar biological phenomenon in other species.

3.2 Visual Illustration Search

Text-to-image retrieval (e.g., Flickr30K (Young et al., 2014), MSCOCO (Chen et al., 2015)) is a canonical multimodal retrieval task, where the system need to retrieve the image that best matches a textual query. Classic benchmarks are largely limited to direct and surface-level semantic alignment, such as identifying a specific animal or a person performing a certain sport. However, real-world use cases often require domain knowledge and multi-step reasoning to retrieve the target image (e.g., professional charts and scientific illustrations). To address this gap, we introduce the **Visual Illustration Search** (VIS) task. In this task, the model is required to retrieve an image that functions as a visual illustration, intuitively explaining or solving a problem posed in a challenging, domain-specific textual query. Comprising three sub-tasks: **Economics**, **Mathematics**, and **Nature**, VIS evaluates a model's ability to perform cross-modal reasoning and knowledge-grounded understanding in complex multimodal scenarios.

Economics. Charts serve as intuitive illustrations across various disciplines. However, existing chart-related tasks (e.g., ChartVQA (Masry et al., 2022), ViDoRe (Faysse et al., 2025)) primarily test surface-level abilities solvable with basic OCR and arithmetic. To assess a model's ability to capture the deeper semantics and domain knowledge embedded in chart, we manually collected reports from the World Bank[3], extracted charts related to economics, and asked human experts to create questions grounded in these charts. The core annotation principle is that each question must demand sufficient reasoning to identify the positive chart. For instance, as shown in Figure 1(d), the positive chart does

[1]Query example curated from the OVEN benchmark (Hu et al., 2023)

[2]https://stackexchange.com

[3]https://data.worldbank.org

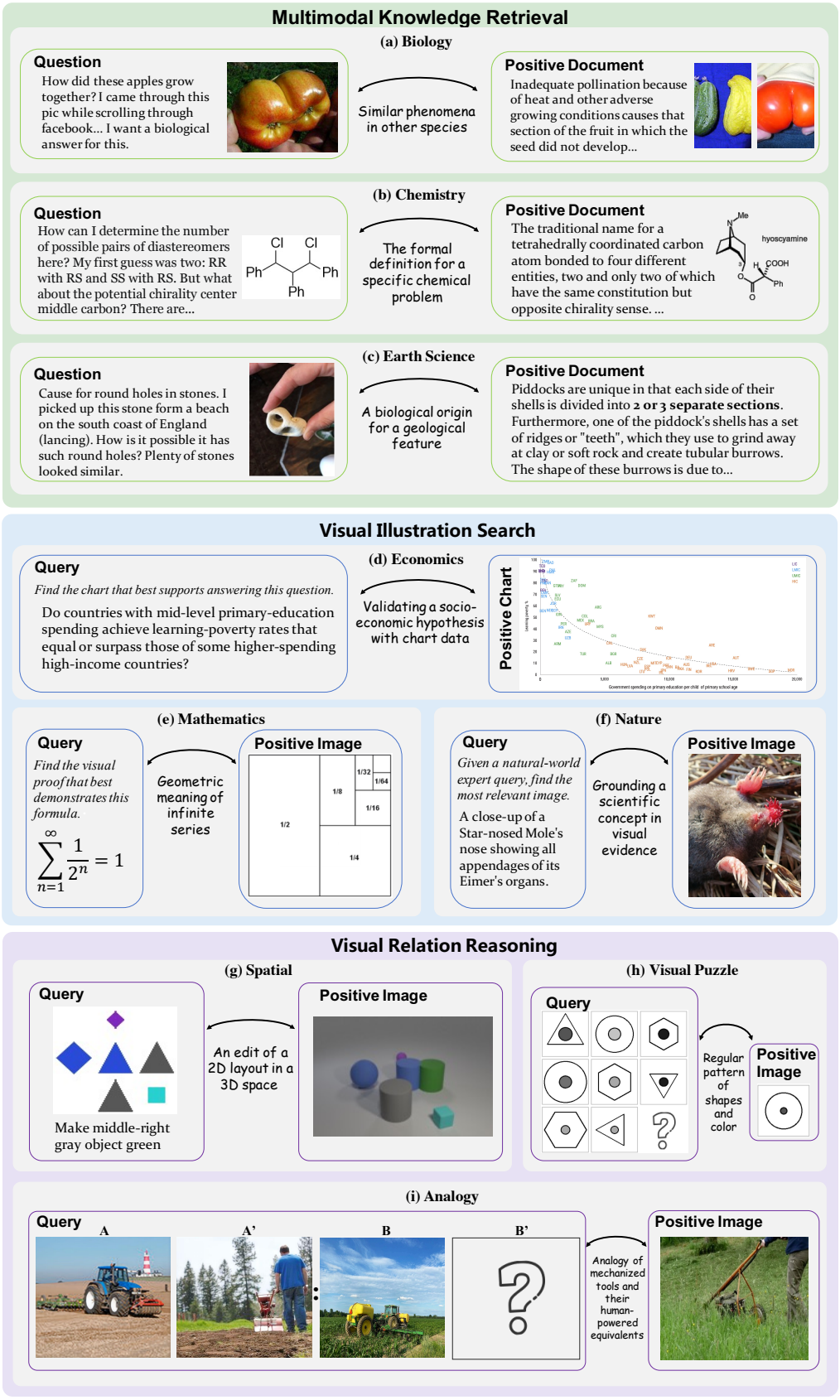

Figure 1: Visualized Examples of MR^2-Bench: Sub-task illustrations from three meta-tasks, with 3 out of 6 shown for the multimodal knowledge retrieval task.

not explicitly state the conclusion; only by comparing the relative positions of different countries in the chart and associating *spending quantiles* with *learning poverty rates* can one validate the hypothesis posed in the question. Following these principles, we constructed a reasoning-oriented retrieval subset centered on economic charts, comprising 84 high-quality questions.

Mathematics. Images can effectively reinforce human comprehension of abstract knowledge. This holds especially in mathematics, where *visual proofs* are conical examples that use geometric relations to demonstrate abstract theorems intuitively. As shown in Figure 1(e), the recursive partition of the unit square gives a clear proof of the infinite series $\sum_{n=1}^{\infty} \frac{1}{2^n} = 1$. Although structurally simple, such proofs embody rigorous logic and require strong reasoning to connect visual patterns with abstract mathematical principles, providing an effective evaluation of model's reasoning ability. However, visual proofs are largely absent from existing multimodal retrieval benchmarks. Therefore, we curate 86 mathematical formulas from *Proofs Without Words* (Nelsen, 2015) and Wikimedia Commons[4], using each formula as a query and its corresponding visual proof as the positive image.

Nature. Natural-world images are more than depictions; they are visual reference for species identification, ecosystem monitoring, and science education (Van Horn et al., 2015; 2018), which require images that capture specific traits or morphology, rather than the generic picture of the organism. For example, as shown in Figure 1(f), the query seeks for *a close-up of star-nosed mole's distinctive organs*, which demands both expert biological knowledge and fine-grained visual recognition. Satisfying such knowledge-intensive visual requests is a challenging yet essential capability for models. To evaluate this, we carefully selected 100 queries from the publicly available INQUIRE-Rerank dataset (Vendrow et al., 2024) to construct the expert-level natural-world image retrieval task.

3.3 Visual Relation Reasoning

In prevailing multimodal retrieval benchmarks, textual queries are the primary driver of user intent. However, this paradigm often overlooks the rich, self-contained semantics inherent in purely visual structures and relationships that are independent of natural language. To address this gap, we introduce **Visual Relation Reasoning**, a suite of tasks for assessing high-level vision-centric reasoning through three distinct sub-tasks: **Spatial**, **Visual Puzzle**, and **Analogy**.

Spatial. The capacity for spatial perception, transformation, and reasoning is essential for models. To evaluate these capabilities, we incorporate tasks from the CSS dataset (Vo et al., 2019), a controlled synthetic dataset where each sample consists of a reference image, a textual modification instruction, and a corresponding target image, with scenes rendered as both 2D layouts and photorealistic 3D images. As illustrated in Figure 1(g), the query requires jointly parsing descriptions that combine relative position and attributes (i.e., *middle-right gray object*) and projecting the 2D layout into the corresponding 3D scene, yielding a comprehensive test of spatial ability. From CSS, we curated 149 queries to constitute the spatial-reasoning subtask of MR^2-Bench.

Visual Puzzle. Inspired by Raven's Progressive Matrices[5], this task is designed to evaluate pattern recognition and structural reasoning. As shown in Figure 1(h), for a given 3×3 matrix with the final cell missing, the model need to retrieve the positive image that logically completes the matrix's underlying pattern. This task is distinguished by its near-complete absence of linguistic signals, which compels the model to directly infer abstract patterns to perform higher-order reasoning from vision alone. We reorganized the RAVEN dataset (Zhang et al., 2019): for each rule-governed visual attribute, we selected a set of queries, pooled the corresponding candidate images and removed duplicates to build the corpus. In total, we curated 160 queries for this task.

Analogy. Derived from the VASR dataset (Bitton et al., 2023), this task tests a model's capability for visual analogical reasoning. As shown in Figure 1(i), the query comprises three images (A, A', B), where the pair (A, A') exemplifies a visual semantic transformation (e.g., *replacing a machine with human labor in a comparable scene*) that is expected to hold between B and B'. The model must infer the transformation from A to A', apply it to B, and retrieve the image B' that completes the analogy. It requires the model abstracts an implicit transformation rule from one image pair and generalizes it to another, which effectively tests its capacity for high-order visual reasoning. We

[4]https://commons.wikimedia.org/wiki/Category%3AProof_without_words

[5]https://en.wikipedia.org/wiki/Raven%27s_Progressive_Matrices

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Text Embedding Models													
BGE-M3	18.79	12.97	12.04	14.52	6.05	16.35	-	-	-	-	-	-	-
+ *Captions*	34.19	24.28	17.88	21.24	9.67	25.19	45.46	9.97	23.66	9.48	0.00	3.46	18.71
Qwen3	23.77	20.44	12.61	17.13	8.61	19.79	-	-	-	-	-	-	-
+ *Captions*	29.97	29.29	18.32	21.46	9.52	23.19	49.44	21.14	26.30	9.11	0.00	4.30	20.17
Diver-Emb.	27.32	16.94	15.17	18.05	10.06	22.57	-	-	-	-	-	-	-
+ *Captions*	38.46	30.87	22.84	23.62	14.46	31.40	54.67	25.91	24.88	8.52	0.00	7.47	23.59
BGE-Rea.	29.01	15.37	16.31	21.00	10.62	26.20	-	-	-	-	-	-	-
+ *Captions*	42.60	34.40	24.94	25.61	14.31	34.57	54.31	17.16	29.86	5.52	0.00	5.88	25.35
ReasonIR	29.85	19.72	16.22	21.56	9.83	23.56	-	-	-	-	-	-	-
+ *Captions*	44.75	41.91	18.79	27.33	17.45	41.22	**64.04**	**34.49**	30.70	11.65	0.00	10.89	25.72
Multimodal Embedding Models													
CLIP	32.85	30.57	14.06	14.86	3.50	33.23	12.97	5.64	49.34	20.89	0.19	5.09	18.59
BGE-VL	29.41	18.36	10.50	19.51	7.12	19.73	50.80	14.31	47.97	6.46	0.00	0.75	19.53
GME	34.34	39.50	19.04	19.29	7.73	28.59	36.95	7.19	39.35	15.70	0.22	**11.11**	21.59
VLM2Vec	39.37	39.38	19.87	20.28	9.03	35.71	51.44	14.16	35.06	13.94	0.62	5.85	23.72
MM-Emb.	**49.68**	**52.19**	23.67	**30.36**	17.44	**47.51**	42.99	21.58	48.41	**22.79**	0.21	5.93	30.23
Seed-1.6	40.64	38.12	**31.77**	27.91	**17.80**	37.17	56.13	26.10	**65.16**	17.29	**0.93**	9.21	**30.68**

Table 3: **The overall performance of embedding models on MR^2-Bench.** We report nDCG@10 for all sub-tasks. Avg. denotes the average score across 12 datasets. The best score on each dataset is shown in bold and the second best is underlined.

instantiate this task by converting VASR analogy triplets into a retrieval setting and curated 147 challenging queries.

4 EXPERIMENTS

4.1 SETTINGS

We evaluated 11 popular embedding models using our MR^2-Bench, categorizing them into two main types: text-only embedding models and multimodal embedding models. We employed nDCG@10 as the primary metric, with additional metric results provided in Appendix E.

For *text embedding models*, we assessed two categories: traditional models such as BGE-M3 (Chen et al., 2024) and Qwen3-Embedding (Zhang et al., 2025b), and models optimized for reasoning-intensive retrieval, including ReasonIR (Shao et al., 2025), BGE-Reasoner-Embed[6], and Diver-Embed (Long et al., 2025). We adopted two evaluation approaches for text embedding models: (1) Using only text information from queries and documents, which is limited for tasks where queries or candidates are purely image-based; (2) Replacing images with textual descriptions (captions). For *multimodal embedding models*, we evaluated CLIP (Radford et al., 2021), VISTA (Zhou et al., 2024), BGE-VL (Zhou et al., 2025), MM-Embed (Lin et al., 2024), GME (Zhang et al., 2025a), VLM2VecV2 (Meng et al., 2025), and Seed1.6-Embedding (Seed, 2025). Detailed information on the models and evaluation procedures can be found in Appendix D.

4.2 MAIN RESULTS

We summarize the overall evaluation results for all investigated retrieval baselines in MR^2-Bench in Table 3. For each sub-task, we report nDCG@10, along with the macro-average (Avg.) across all tasks. From these results, we draw some primary conclusions:

1) Current state-of-the-art models underperform on MR^2-Bench. The leading Seed-1.6 Embedding model (Seed, 2025) achieves only 30.68 nDCG@10 on our benchmark. In contrast, it reports 77.78 overall Recall@1 on the popular MMEB leaderboard (Jiang et al., 2025), but its performance

[6]https://huggingface.co/BAAI/bge-reasoner-embed-qwen3-8b-0923

drops significantly to 9.91 Recall@1 on MR^2-Bench. Additionally, the SOTA reasoning-intensive text retriever, Diver-Retriever (Long et al., 2025), achieves 33.90 nDCG@10 on BRIGHT (Hongjin et al., 2025), yet only reaches 23.59 nDCG@10 on MR^2-Bench when evaluated with auxiliary captions. These results highlight the increased challenges posed by our MR^2-Bench.

2) Text retrievers augmented with image captions provide a strong and practical baseline on MR^2-Bench. Since text retrievers cannot directly process images, we replace each image in queries and candidate documents with detailed natural-language descriptions. This augmentation leads to notable improvements. For instance, ReasonIR+*Captions* surpasses popular open-source multimodal retrievers like VLM2Vec-V2 (Meng et al., 2025). On the Stack Exchange subset, adding captions consistently boosts performance across most tasks. These findings confirm that MR^2-Bench is fundamentally multimodal, with retrieval performance significantly enhanced by the visual information provided through captions.

3) Reasoning-oriented text retrievers significantly outperform traditional matching-based retrievers. Models optimized for reasoning-intensive retrieval, such as ReasonIR and Diver-Retriever, consistently achieve higher nDCG@10 scores on MR^2-Bench compared to matching-centric retrievers like BGE-M3 and Qwen3-Embedding. This advantage is evident across various meta-tasks and persists whether visual content is absent or represented as detailed captions. Collectively, these findings suggest that reasoning-oriented capabilities learned in text retrieval effectively transfer to multimodal retrieval tasks requiring complex reasoning.

4) Multimodal retrievers show potential on MR^2-Bench. Although not specifically designed for reasoning-intensive tasks, multimodal embedding models like MM-Embed and Seed1.6-Embedding lead performance on MR^2-Bench. These models notably outperform caption-augmented text retrievers, including those optimized for reasoning. This gap suggests a promising direction for future research in developing reasoning-intensive multimodal retrievers.

5) Existing methods struggle with capturing complex visual relationships and abstract concepts. Current models face challenges in effectively perceiving multi-image relationships (Analogy), spatial configurations (Spatial), and abstract graphics (Mathematics, Visual Puzzle). We hypothesize that these difficulties stem from the inherently visual-centric nature of these tasks, which existing embedding models struggle to comprehend fully. Nonetheless, these images are crucial for real-world applications, as their information is difficult to convey through language alone. This indicates substantial potential for future research to enhance multimodal embedding models.

4.3 More Analysis

4.3.1 The Effectiveness of Query Rewriting

6) Query rewriting enhances both text and multimodal baselines on MR^2-Bench. This generation-augmented retrieval technique clarifies complex user intent and highlights latent constraints, thus facilitating reasoning-intensive retrieval. Although extensively studied in text-only contexts (Gao et al., 2023; Li et al., 2025a), its application to multimodal retrieval remains underexplored. We evaluated a simple, model-agnostic query rewriting pipeline on MR^2-Bench. For each query, GPT-5 (OpenAI, 2025) generates step-by-step reasoning, which is then utilized by each retriever (details in Appendix F). As shown in Table 4, both text and multimodal retrievers show notable average improvements. These results indicate that query rewriting is a practical method for enhancing multimodal reasoning-intensive retrieval tasks, consistently improving performance without the need for fine-tuning existing retrievers.

Methods	Stack Exchange						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
BGE-M3	34.19	24.28	17.88	21.24	9.67	25.19	45.46	9.97	23.66	9.48	0.00	3.46	18.71
+ *Rewrite*	40.41	32.94	25.66	23.12	11.98	33.63	50.88	20.09	23.38	7.13	0.00	7.91	**23.09**
Seed-1.6	40.64	38.12	31.77	27.91	17.80	37.17	56.13	26.10	65.16	17.29	0.93	9.21	30.68
+ *Rewrite*	41.13	41.47	37.68	29.47	20.70	42.02	50.08	30.37	65.84	31.87	1.24	14.62	**33.87**

Table 4: Performance comparison of BGE-M3 and Seed-1.6 Embedding on MR^2-Bench before and after query rewriting, showing significant improvements across most tasks.

4.3.2 THE EFFECTIVENESS OF ADVANCED RERANKING

A common approach to improve retrieval performance is to employ rerankers that jointly process both the query and its retrieved candidates. Existing studies have shown that incorporating an intermediate reasoning step before final scoring can lead to more accurate rankings (Weller et al., 2025; Zhuang et al., 2025; Liu et al., 2025). We also investigate this by incorporating a reranking stage after the initial retrieval on MR^2-Bench. Specifically, we test a wide range of rerankers to rerank the top-$k = 20$ candidates retrieved by three base retrievers: Qwen3-Embedding, GME, and Seed-1.6-Embedding. Their retrieved candidates are reanked by: 1) *textual rerankers*: RankLLaMA-7B and RankLLaMA-14B (Ma et al., 2024); 2) *reasoning-enhanced textual rerankers*: Rank1-7B (Weller et al., 2025), RankR1-14B (Zhuang et al., 2025), ReasonRank-32B (Liu et al., 2025), and BGE-Reasoner-Reranker-32B [7]; 3) *multimodal rerankers*: MonoQwen2-VL-v0.1 (Chaffin & Lac, 2024) and Jina-Reranker-m0 (JinaAI, 2025); and 4) *reasoning-enhanced multimodal rerankers*: Gemma3-27B (Team, 2025), Qwen2.5-VL-72B (Bai et al., 2025), GLM-4.5V (Team et al., 2025), and GPT-5 (OpenAI, 2025). Since there are no off-the-shelf multimodal rerankers that natively support reasoning, we prompt these MLLMs to first perform reasoning and then output a relevance score. Full implementation details are available in Appendix G.1. Average performance based on Seed-1.6-Embedding is shown in Figure 2, and detailed results for all three base retrievers are provided in the Appendix G.2.

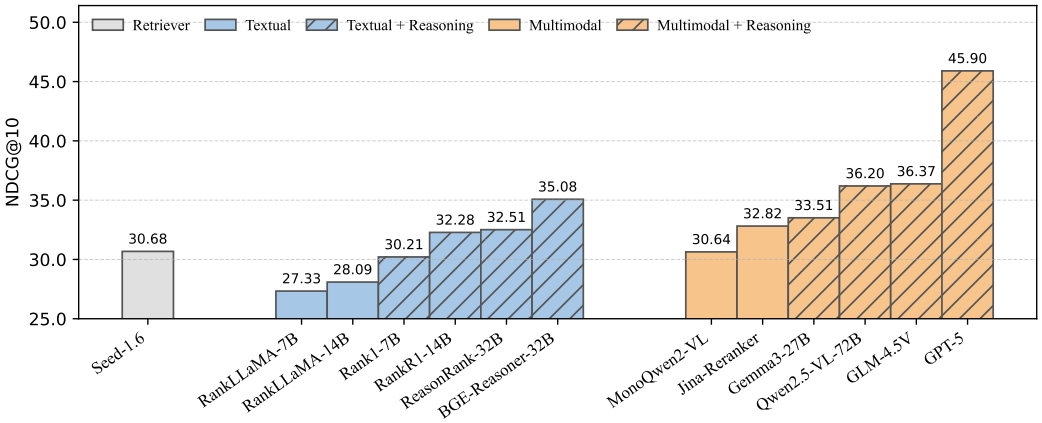

Figure 2: Reranking performance on MR^2-Bench with Seed-1.6-Embedding as the base retriever.

From the results presented in Figure 2, we have following findings:

7) Rerankers deliver substantial gains on MR^2-Bench. Most rerankers significantly outperform the strong Seed-1.6-Embedding baseline, demonstrating the benefit of joint modeling of queries and candidates. Notably, GPT-5 achieves an nDCG@10 of 45.90, an absolute gain of 15.22 over the baseline, indicating the substantial headroom for improvement unlocked by reranking.

8) An explicit reasoning step before scoring proves to be beneficial. Across text-only rerankers, those incorporating reasoning consistently outperform their non-reasoning, size-matched counterparts (e.g., Rank1-7B vs. RankLLaMA-7B; RankR1-14B vs. RankLLaMA-14B). This is further substantiated by BGE-Reasoner-Reranker-32B: using only textual input, it achieves an nDCG@10 of 35.08, outperforming the strong base retriever by 4.2 points. Moreover, for multimodal rerankers, models prompted to reason and then rank outperform those trained non-reasoning rerankers. These results confirm that explicit reasoning drives the gains on MR^2-Bench.

9) Multimodal information plays a significant role in enhancing performance. Despite being built on the lightweight Qwen2-VL-2B backbone, Jina-Reranker-m0 surpasses several larger text-only rerankers, demonstrating clear gains from multimodal information. Furthermore, multimodal models prompted to first reason and then rank (e.g., Qwen2.5-VL-72B, GLM-4.5V, and GPT-5) surpass BGE-Reasoner-Reranker-32B, the best-performing textual reranker specifically trained with reasoning capabilities. GPT-5 achieves the highest overall score, underscoring the importance of

[7] https://github.com/FlagOpen/FlagEmbedding/tree/master/research/BGE_Reasoner

utilizing multimodal information with reasoning in tackling the complex retrieval demands posed by MR^2-Bench.

5 Conclusion

In this paper, we introduce MR^2-Bench, a novel benchmark for the assessment of multimodal reasoning-intensive retrieval. The comprehensive investigation of existing methods reveals that current retrievers perform poorly on MR^2-Bench, with the best models achieving only 30.68 nDCG@10. Our experimental results underscore the importance of multimodal information and reasoning capabilities for effectively addressing MR^2-Bench, highlighting significant potential for improvement in this research area. Additionally, we demonstrate that techniques such as query rewriting and reranking can enhance performance on MR^2-Bench. We anticipate that this benchmark will facilitate future research in multimodal retrieval, contributing to more realistic and challenging AI applications.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Yonatan Bitton, Ron Yosef, Eliyahu Strugo, Dafna Shahaf, Roy Schwartz, and Gabriel Stanovsky. Vasr: Visual analogies of situation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 241–249, 2023.

Antoine Chaffin and Aurélien Lac. Monoqwen: Visual document reranking, 2024. URL https://huggingface.co/lightonai/MonoQwen2-VL-v0.1.

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.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1870–1879, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1171. URL https://aclanthology.org/P17-1171/.

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

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pp. 5558–5570. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.EMNLP-MAIN.375. URL https://doi.org/10.18653/v1/2022.emnlp-main.375.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14948–14968, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.925. URL https://aclanthology.org/2023.emnlp-main.925/.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.

SU Hongjin, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Liu Haisu, Quan Shi, Zachary S Siegel, Michael Tang, et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. In *The Thirteenth International Conference on Learning Representations*, 2025.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 12031–12041. IEEE, 2023. doi: 10.1109/ICCV51070.2023.01108. URL https://doi.org/10.1109/ICCV51070.2023.01108.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*, abs/2503.09516, 2025. doi: 10.48550/ARXIV.2503.09516. URL https://doi.org/10.48550/arXiv.2503.09516.

JinaAI. jina-reranker-m0. https://jina.ai/news/jina-reranker-m0-multilingual-multimodal-document-reranker/, April 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 6769–6781. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.EMNLP-MAIN.550. URL https://doi.org/10.18653/v1/2020.emnlp-main.550.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Chaofan Li, Jianlyu Chen, Yingxia Shao, Chaozhuo Li, Quanqing Xu, Defu Lian, and Zheng Liu. Reinforced IR: A self-boosting framework for domain-adapted information retrieval. In

Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22061–22073, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1071. URL https://aclanthology.org/2025.acl-long.1071/.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *CoRR*, abs/2501.05366, 2025b. doi: 10.48550/ARXIV.2501.05366. URL https://doi.org/10.48550/arXiv.2501.05366.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*, 2024.

Wenhan Liu, Xinyu Ma, Weiwei Sun, Yutao Zhu, Yuchen Li, Dawei Yin, and Zhicheng Dou. Reasonrank: Empowering passage ranking with strong reasoning ability. *arXiv preprint arXiv:2508.07050*, 2025.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2125–2134, 2021.

Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yue Shen, Jian Wang, Peng Wei, Jinjie Gu, and Jiahai Wang. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*, 2025.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. End-to-end knowledge retrieval with multi-modal queries. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pp. 8573–8589. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-LONG.478. URL https://doi.org/10.18653/v1/2023.acl-long.478.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning llama for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2421–2425, 2024.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

Rui Meng, Ziyan Jiang, Ye Liu, Mingyi Su, Xinyi Yang, Yuepeng Fu, Can Qin, Zeyuan Chen, Ran Xu, Caiming Xiong, et al. Vlm2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. *arXiv preprint arXiv:2507.04590*, 2025.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022.

Roger B Nelsen. *Proofs without words III: Further exercises in visual thinking*, volume 52. American Mathematical Soc., 2015.

OpenAI. Gpt-5. https://openai.com/gpt-5/, August 2025.

Hongjin Qian and Zheng Liu. Scent of knowledge: Optimizing search-enhanced reasoning with information foraging. *arXiv preprint arXiv:2505.09316*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

ByteDance Seed. Seed1-6 embedding. https://seed1-6-embedding.github.io, June 2025.

Rulin Shao, Rui Qiao, Varsha Kishore, Niklas Muennighoff, Xi Victoria Lin, Daniela Rus, Bryan Kian Hsiang Low, Sewon Min, Wen-tau Yih, Pang Wei Koh, et al. Reasonir: Training retrievers for reasoning tasks. *arXiv preprint arXiv:2504.20595*, 2025.

Gemma Team. Gemma 3. 2025. URL https://goo.gle/Gemma3Report.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Grant Van Horn, Steve Branson, Ryan Farrell, Scott Haber, Jessie Barry, Panos Ipeirotis, Pietro Perona, and Serge Belongie. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 595–604, 2015.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8769–8778, 2018.

Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E Jones, Oisin Mac Aodha, Sara Beery, and Grant Van Horn. Inquire: A natural world text-to-image retrieval benchmark. *NeurIPS*, 2024.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and image for image retrieval-an empirical odyssey. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6439–6448, 2019.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVII*, volume 15145 of *Lecture Notes in Computer Science*, pp. 387–404. Springer, 2024. doi: 10.1007/978-3-031-73021-4_23. URL https://doi.org/10.1007/978-3-031-73021-4_23.

Orion Weller, Kathryn Ricci, Eugene Yang, Andrew Yates, Dawn Lawrie, and Benjamin Van Durme. Rank1: Test-time compute for reranking in information retrieval, 2025. URL https://arxiv.org/abs/2502.18418.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp. 11307–11317, 2021.

Jinming Wu, Zihao Deng, Wei Li, Yiding Liu, Bo You, Bo Li, Zejun Ma, and Ziwei Liu. Mmsearch-r1: Incentivizing lmms to search. *arXiv preprint arXiv:2506.20670*, 2025.

Chenghao Xiao, G Thomas Hudson, and Noura Al Moubayed. Rar-b: Reasoning as retrieval benchmark. *arXiv preprint arXiv:2404.06347*, 2024a.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

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

Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5317–5327, 2019.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Zc22RDtsvP.

Xin Zhang, Yanzhao Zhang, Wen Xie, Mingxin Li, Ziqi Dai, Dingkun Long, Pengjun Xie, Meishan Zhang, Wenjie Li, and Min Zhang. Bridging modalities: Improving universal multimodal retrieval by multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9274–9285, 2025a.

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*, 2025b.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized text embedding for universal multi-modal retrieval. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3185–3200, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.175. URL https://aclanthology.org/2024.acl-long.175.

Junjie Zhou, Yongping Xiong, Zheng Liu, Ze Liu, Shitao Xiao, Yueze Wang, Bo Zhao, Chen Jason Zhang, and Defu Lian. MegaPairs: Massive data synthesis for universal multimodal retrieval. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19076–19095, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.935. URL https://aclanthology.org/2025.acl-long.935/.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. Rank-r1: Enhancing reasoning in llm-based document rerankers via reinforcement learning. *arXiv preprint arXiv:2503.06034*, 2025.

APPENDIX

A USE OF LLMS

In preparing this manuscript, large language models (LLMs) were utilized solely for English grammar checking and polishing. All substantive content and analyses were developed independently by the authors. For dataset construction, GPT-5 (OpenAI, 2025) was employed only for preliminary filtering of candidate data and generating some challenging negative examples, with all final selections and included negative examples thoroughly reviewed and validated by human experts. The relevant procedures are detailed in the appropriate sections of the paper.

B DETAILED OVERVIEW OF MR^2-BENCH

We provide detailed modalities of queries and documents, along with the instructions for each subtask in Table 5.

Meta-Task	Sub-Task	Modality ($q \to c$)	#Queries	#Corpus	Instruction
MULTIMODAL KNOWLEDGE RETRIEVAL	Biology	$q_{i+t} \to c_{i/t/i+t}$	79	4,455	*Find paragraph(s) that could support answering this question.*
	Cooking	$q_{i+t} \to c_{i/t/i+t}$	76	2,786	*Find paragraph(s) that could support answering this question.*
	Gardening	$q_{i+t} \to c_{i/t/i+t}$	129	5,636	*Find paragraph(s) that could support answering this question.*
	Physics	$q_{i+t} \to c_{i/t/i+t}$	76	6,656	*Find paragraph(s) that could support answering this question.*
	Chemistry	$q_{i+t} \to c_{i/t/i+t}$	124	4,317	*Find paragraph(s) that could support answering this question.*
	EarthScience	$q_{i+t} \to c_{i/t/i+t}$	99	3,014	*Find paragraph(s) that could support answering this question.*
VISUAL ILLUSTRATION SEARCH	Economics	$q_t \to c_i$	84	7,572	*Find the chart that best supports answering this question.*
	Mathematics	$q_t \to c_i$	86	944	*Find the visual proof that best demonstrates this formula.*
	Nature	$q_t \to c_i$	100	2,017	*Given a natural-world expert query, find the most relevant image.*
VISUAL RELATION	Spatial	$q_{i+t} \to c_i$	149	1,000	*Given a reference image and a text modification, retrieve the image that best matches the modified reference.*
	Visual Puzzle	$q_i \to c_i$	160	5,375	*From a 3×3 grid with one missing cell, retrieve the best candidate image to complete the bottom-right cell based on patterns and relations.*
	Analogy	$q_i \to c_i$	147	3,970	*Given three images, complete the analogy by retrieving the candidate that applies to the third image the relation from the first to the second.*

Table 5: **The overview of MR^2-Bench.** MR^2-Bench consists of three meta-tasks and twelve subtasks, totaling 1,309 queries. Subscripts indicate the modalities of the query q and candidate c: i denotes image, t denotes text, and $i{+}t$ denotes interleaved image-text.

C MORE DETAILS OF DATA CONSTRUCTION FOR MULTIMODAL KNOWLEDGE RETRIEVAL TASKS

We collected real posts from the Stack Exchange platform to construct our multimodal knowledge retrieval sub-tasks. Queries are derived from actual user questions, while positive documents are sourced from external links in highly voted answers. We utilize BRIGHT's definition to identify a query's positive document: *A document is relevant only if cited in a highly voted answer and confirmed by annotators and domain experts as aiding in reasoning through the query with critical concepts or theories* (Hongjin et al., 2025). Given the multimodal nature of the task in MR^2-Bench, our annotation process diverges from BRIGHT's construction methodology. The specific steps of our process are summarized as follows:

Initial Posts Collection and Filtering. We initiated the process by gathering a substantial set of posts from Stack Exchange. To ensure data quality and relevance, we retained posts meeting specific criteria: (1) the question must contain image(s) essential for understanding the query; (2) the post must have received at least five community votes, indicating reliability; and (3) the answer must include at least one external link to facilitate further content acquisition.

Web Page Acquisition and Paragraph Annotation. For each qualifying post, annotators are required to visit the external links provided in the answers and copy the interleaved text-image content in the order it appears, excluding Wikipedia.[8] They then segment this content into paragraphs, preserving images to maintain multimodal information. This process generates a collection of candidate paragraphs for each query, including both text-only and image-containing segments. Initial identification of positive paragraphs is performed using GPT-5 (OpenAI, 2025), followed by expert validation to ensure accuracy and relevance. Only queries with at least one confirmed positive paragraph are included in the final dataset.

Incorporation of Challenging Negative Examples. To rigorously assess the reasoning capabilities of evaluation methods, we introduced challenging negative samples for each retained query using two strategies: (1) retrieving topic-related documents from an internal corpus using the query's keywords, with GPT-5 initially verifying they are not false negatives; and (2) using GPT-5 to generate documents that, while topically related, provide unhelpful information. All negative samples were subsequently reviewed by human experts to ensure the integrity of the benchmark.

D More Details of Baselines

In our evaluation, we classify the retriever baseline into two main categories: text embedding models and multimodal embedding models. We assess the Seed1.6-Embedding model (Seed, 2025) via its official API, whereas all other models are evaluated using their publicly available code and open-source checkpoints. Below, we provide a comprehensive overview of the implementation details for all baselines used in the evaluation process.

D.1 Text Embedding Models

The evaluated text retrievers include: BGE-M3 (Chen et al., 2024), Qwen3-Embedding (Zhang et al., 2025b), ReasonIR (Shao et al., 2025), BGR-Reasoner-Embed[9], and Diver-Embed (Long et al., 2025). Notably, the last three models have been fine-tuned specifically for reasoning-intensive retrieval tasks, as detailed in their technical reports or repository descriptions.

We consider two input configurations for all text-only retrievers. The first configuration ignores images, utilizing only the textual content from queries and documents; this setup is not applicable to some sub-tasks where either the query or candidates are purely visual. The second configuration employs a caption-augmented approach, where every image in both queries and documents is replaced with a textual description. Specifically, we use the Qwen2.5-VL-7B model (Bai et al., 2025) to generate captions for the images with the prompt: *Write a detailed English caption for this image, covering the main objects, their attributes, relationships, actions, layout, and background elements.* Each image in the original input is then substituted with a caption prefixed by its identifier, formatted as `[IMAGE_id]: image_caption`.

D.2 Multimodal Embedding Models

The evaluated multimodal retrievers include CLIP (Radford et al., 2021), BGE-VL (Zhou et al., 2025), GME (Zhang et al., 2025a), VLM2Vec-V2 (Meng et al., 2025), MM-Embed (Lin et al., 2024), and Seed1.6-Embedding (Seed, 2025). All these models can process individual images and texts directly. However, for interleaved image-text data with multiple images, different models require specific handling approaches:

[8] Wikipedia content was automatically extracted using Playwright to minimize manual effort.

[9] https://huggingface.co/BAAI/bge-reasoner-embed-qwen3-8b-0923

For the CLIP model, we employ a score fusion strategy, following previous work (Wei et al., 2024). This involves separately embedding the image and text data and then combining these embeddings through element-wise addition to achieve the final image-text representation.

For models that can only input a single image in image-text data, specifically BGE-VL (Zhou et al., 2025) and MM-Embed (Lin et al., 2024), we create a composite image by tiling multiple images together, which is then processed jointly with the text.

For other models capable of handling interleaved image-text data with multiple images, we preserve the sequence of images and text, allowing their processors to generate interleaved image-text tokens, which are then used to derive the final embeddings. s and diagnostic analyses are provided in the Appendix.

E Detailed Evaluation Metrics of MR^2-Bench

In this section, we provide more detailed evaluation results of the embedding models on MR²-Bench. Table 6, Table 7, Table 8, Table 9, and Table 10 present the performance of the embedding models in terms of Recall@1, Recall@5, Recall@10, nDCG@5 and nDCG@20.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Text Embedding Models													
BGE-M3	3.61	2.25	3.26	2.70	1.04	3.77	-	-	-	-	-	-	-
+ *Captions*	10.22	3.23	6.44	5.00	1.43	7.26	32.14	3.49	6.67	4.00	0.00	1.36	6.77
Qwen3	5.46	2.67	3.11	1.86	1.13	4.67	-	-	-	-	-	-	-
+ *Captions*	7.21	6.87	5.15	4.84	1.20	5.93	32.14	6.98	4.92	4.03	0.00	0.68	6.66
Diver-Emb.	5.73	2.55	4.74	1.60	0.38	3.71	-	-	-	-	-	-	-
+ *Captions*	12.37	7.06	9.87	4.60	2.39	6.60	36.90	8.14	3.00	3.33	0.00	0.68	7.91
BGE-Rea.	3.69	3.13	4.10	2.59	1.36	4.64	-	-	-	-	-	-	-
+ *Captions*	16.03	9.84	9.74	6.19	1.24	10.28	41.67	12.21	1.67	2.00	0.00	0.68	9.29
ReasonIR	7.68	3.13	3.75	4.35	0.91	4.21	-	-	-	-	-	-	-
+ *Captions*	16.87	13.81	7.13	5.32	3.50	11.59	39.29	7.56	6.58	2.00	0.00	0.68	9.53
Multimodal Embedding Models													
CLIP	12.49	8.28	4.37	2.58	1.42	11.72	3.57	1.16	10.92	12.67	0.00	0.00	5.77
BGE-VL	8.96	2.30	2.93	4.35	0.32	4.81	34.52	6.98	10.83	2.01	0.00	1.36	6.62
GME	10.07	14.48	7.84	3.97	1.43	8.39	21.43	2.33	8.08	8.00	0.00	3.40	7.45
VLM2Vec	13.58	13.73	5.41	3.73	1.44	14.54	38.10	3.49	9.53	4.00	0.62	0.68	9.07
MM-Emb.	17.18	20.81	7.10	7.05	4.35	17.54	34.52	9.59	11.25	11.33	0.00	0.00	11.73
Seed-1.6	13.65	9.02	9.85	5.20	3.69	9.81	33.33	6.98	19.33	8.00	0.00	0.00	9.91

Table 6: The overall performance of embedding models on MR^2-Bench in terms of the recall@1.

F More Details of Implementation for Query Rewriting

Given the strong reasoning capabilities of Multimodal Large Language Models (MLLMs), we take advantage of their ability to produce explicit step-by-step chain-of-thought reasoning in order to improve the effectiveness of query rewriting and thereby enhance retrieval performance. Instead of relying on a single direct reformulation, we design a prompting strategy that guides the MLLM through a structured reasoning process. Concretely, the model is first asked to (i) identify the most salient subquestions that are implicitly contained in the given instruction and query, ensuring that complex or multifaceted information needs are decomposed into clear components. Next, the model is prompted to (ii) reason step-by-step about what types of evidence, textual patterns, and document attributes would be necessary for relevant sources to contain, which encourages a more targeted and discriminative retrieval process. Finally, model (iii) produces both an explicit reasoning trace, which captures its internal deliberation, and a set of candidate rewritten queries or answers that can

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Text Embedding Models													
BGE-M3	14.09	10.43	11.66	10.12	7.27	12.78	-	-	-	-	-	-	-
+ *Captions*	28.85	23.21	17.32	13.60	8.76	20.66	53.57	10.85	20.25	11.33	0.00	3.40	17.65
Qwen3	17.36	12.48	12.17	11.93	6.38	15.12	-	-	-	-	-	-	-
+ *Captions*	24.76	27.10	16.30	13.73	6.61	17.95	60.71	30.33	24.42	11.41	0.00	5.44	19.90
Diver-Emb.	24.54	12.43	15.95	13.61	8.07	19.33	-	-	-	-	-	-	-
+ *Captions*	30.82	27.00	20.75	16.47	11.35	29.76	65.48	37.50	21.67	10.00	0.00	10.88	23.47
BGE-Rea.	23.36	8.80	15.83	13.79	8.05	19.01	-	-	-	-	-	-	-
+ *Captions*	33.49	32.50	25.09	17.00	12.56	26.61	70.24	46.71	25.42	8.67	0.00	6.80	25.42
ReasonIR	26.10	16.55	14.73	14.94	10.08	20.38	-	-	-	-	-	-	-
+ *Captions*	33.01	36.37	16.49	20.05	17.45	36.66	61.90	20.93	24.33	6.00	0.00	8.16	23.45
Multimodal Embedding Models													
CLIP	27.54	28.63	9.72	7.60	4.01	29.62	16.67	4.65	48.17	22.67	0.00	6.80	17.17
BGE-VL	22.17	12.55	10.96	15.18	6.30	15.22	63.10	17.64	47.33	10.07	0.00	5.44	18.83
GME	27.06	33.94	15.53	13.36	6.25	24.78	45.24	6.40	37.42	20.67	0.00	12.24	20.24
VLM2Vec	30.64	34.61	18.89	12.44	7.36	32.18	55.95	17.25	31.75	17.33	0.63	4.08	21.93
MM-Emb.	38.24	48.89	21.07	21.03	16.53	42.79	48.81	22.58	42.08	28.00	0.00	6.12	28.01
Seed-1.6	31.93	32.51	28.95	22.17	14.52	31.65	69.05	38.76	61.25	19.33	0.63	8.16	29.91

Table 7: The overall performance of embedding models on MR^2-Bench in terms of the recall@5.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Text Embedding Models													
BGE-M3	25.92	18.48	17.29	17.00	9.11	22.31	-	-	-	-	-	-	-
+ *Captions*	39.67	35.42	20.81	21.23	14.02	32.55	61.90	19.57	33.83	16.67	0.00	7.48	25.26
Qwen3	32.83	30.81	17.08	20.64	13.94	27.05	-	-	-	-	-	-	-
+ *Captions*	33.91	38.52	24.39	21.03	15.41	31.27	67.86	38.08	39.67	16.11	0.00	8.84	27.92
Diver-Emb.	35.18	25.13	20.01	22.07	16.42	33.76	-	-	-	-	-	-	-
+ *Captions*	43.81	42.45	26.21	25.29	22.66	43.71	70.24	43.31	40.83	16.00	0.00	15.65	32.51
BGE-Rea.	41.29	24.18	23.80	23.82	17.24	37.44	-	-	-	-	-	-	-
+ *Captions*	46.35	42.84	29.78	24.51	22.96	43.02	79.76	58.53	40.42	12.00	0.00	11.56	34.31
ReasonIR	37.42	29.45	22.93	24.88	16.21	34.28	-	-	-	-	-	-	-
+ *Captions*	46.05	50.89	20.99	28.93	25.72	52.31	69.05	28.88	44.08	10.00	0.00	12.93	32.48
Multimodal Embedding Models													
CLIP	33.08	38.08	14.38	14.05	5.64	38.85	26.19	12.21	70.42	31.33	0.63	11.56	24.70
BGE-VL	38.03	27.68	16.66	22.85	11.62	28.46	66.67	23.45	67.83	12.08	0.00	12.93	27.35
GME	35.13	45.64	21.93	19.66	13.14	36.10	54.76	15.50	57.17	25.33	0.63	23.13	29.01
VLM2Vec	41.02	44.31	23.69	20.66	13.20	39.49	66.67	27.23	47.67	27.33	0.63	14.97	30.57
MM-Emb.	50.98	55.18	26.60	28.91	22.79	54.61	51.19	35.08	68.42	35.33	0.63	14.29	37.00
Seed-1.6	47.99	49.13	38.60	30.05	26.32	48.90	79.76	47.87	84.17	30.67	2.50	22.45	42.37

Table 8: The overall performance of embedding models on MR^2-Bench in terms of the recall@10.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Text Embedding Models													
BGE-M3	14.89	10.15	9.78	13.18	5.33	13.23	-	-	-	-	-	-	-
+ *Captions*	32.01	19.33	17.27	21.47	8.02	21.21	42.82	7.13	17.66	7.81	0.00	2.18	16.41
Qwen3	18.71	13.35	10.80	14.72	5.76	15.71	-	-	-	-	-	-	-
+ *Captions*	27.36	24.45	15.25	20.60	6.75	18.77	47.01	18.56	19.48	7.70	0.00	3.20	17.43
Diver-Emb.	24.49	11.88	13.45	16.24	6.70	17.47	-	-	-	-	-	-	-
+ *Captions*	36.03	24.59	20.72	21.95	10.66	26.57	53.13	23.94	16.49	6.60	0.00	5.97	20.56
BGE-Rea.	23.63	8.90	12.79	18.98	6.92	19.62	-	-	-	-	-	-	-
+ *Captions*	40.07	30.43	23.44	25.26	10.06	28.38	57.44	30.40	18.43	5.59	0.00	4.11	22.80
ReasonIR	26.90	14.77	13.12	19.15	7.68	17.96	-	-	-	-	-	-	-
+ *Captions*	42.83	36.92	17.98	26.60	14.74	36.36	52.01	14.34	21.23	4.30	0.00	4.36	22.64
Multimodal Embedding Models													
CLIP	33.19	27.83	10.42	15.34	4.90	30.50	9.91	3.13	39.38	18.14	0.00	3.57	16.36
BGE-VL	26.00	12.99	9.92	18.10	5.71	14.74	49.60	12.44	39.07	5.83	0.00	3.53	16.49
GME	33.91	35.58	17.03	18.41	5.46	25.61	33.89	4.18	30.64	14.21	0.00	7.64	18.88
VLM2Vec	38.31	36.87	18.75	19.66	7.46	34.05	47.87	10.86	28.12	10.74	0.63	2.46	21.31
MM-Emb.	48.80	50.58	22.22	30.84	15.50	44.52	42.15	17.22	37.07	20.34	0.00	3.29	27.71
Seed-1.6	36.14	32.45	28.34	27.69	13.46	31.52	52.63	22.95	55.12	13.67	0.31	4.49	26.56

Table 9: The overall performance of embedding models on MR^2-Bench in terms of the nDCG@5.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Text Embedding Models													
BGE-M3	22.66	16.33	13.67	17.72	7.89	19.84	-	-	-	-	-	-	-
+ *Captions*	37.22	28.43	19.70	24.36	11.39	30.06	47.59	12.65	27.18	10.95	0.00	5.14	21.22
Qwen3	30.34	24.25	15.39	20.39	12.16	24.16	-	-	-	-	-	-	-
+ *Captions*	36.76	33.24	21.41	24.91	13.21	29.03	49.75	24.00	30.48	10.96	0.00	6.70	23.37
Diver-Emb.	32.45	23.00	17.59	24.00	13.65	28.12	-	-	-	-	-	-	-
+ *Captions*	43.90	36.21	25.50	28.19	18.26	37.13	58.30	29.83	29.68	10.52	0.17	8.86	27.21
BGE-Rea.	33.43	21.87	18.97	25.40	14.43	30.65	-	-	-	-	-	-	-
+ *Captions*	47.04	39.94	28.24	29.94	19.20	40.59	63.25	35.94	33.36	8.87	0.00	7.89	29.52
ReasonIR	36.90	24.69	18.92	25.97	13.12	30.39	-	-	-	-	-	-	-
+ *Captions*	48.18	45.34	21.83	28.71	21.15	44.74	57.61	19.33	36.48	6.19	0.00	8.91	28.21
Multimodal Embedding Models													
CLIP	35.49	31.94	13.96	16.53	6.01	34.38	14.76	6.57	56.32	23.04	0.53	6.71	20.52
BGE-VL	36.96	26.11	17.09	23.25	9.47	26.61	52.91	16.12	53.97	7.71	0.17	8.46	23.24
GME	38.17	43.48	20.72	21.56	10.82	33.43	40.88	9.08	45.85	18.74	0.22	14.88	24.82
VLM2Vec	42.11	43.24	21.36	21.93	11.63	38.91	55.66	18.46	40.38	17.18	0.63	8.77	26.69
MM-Emb.	51.83	54.36	26.38	32.74	20.39	51.77	45.99	22.91	55.04	23.97	0.36	8.00	32.81
Seed-1.6	46.01	43.31	35.86	32.99	22.85	43.71	58.25	28.38	69.97	21.20	1.67	11.76	34.66

Table 10: The overall performance of embedding models on MR^2-Bench in terms of the nDCG@20.

be used to drive retrieval more effectively. We employ GPT-5 (OpenAI, 2025), the SOTA multimodal reasoning model, to perform query rewriting. The prompt is provided in Figure 3.

Task Description:
You are an AI assistant specializing in information retrieval and reasoning. Given an instruction and a question (consist of text and images), your task is to generate a "Chain-of-thought" reasoning process. This process must clearly outline the key information that needs to be found in relevant document to answer the question.

Execution Flow:
(1) Identify the Essential Problem: First, precisely extract the fundamental problem that needs to be solved.
(2) Reason on Required Information: Based on the essential problem, conduct step-by-step reasoning to specify the content that needs to be retrieved. This should include relevant terms, phenomena, causes, characteristics, risks, or solutions.
(3) Synthesize the Answer: Based on the reasoning, formulate a direct and concise answer to the problem.
(4) Combine for Output: Consolidate the "Essential Problem", the "Reasoning on Required Information", and the "Synthesized Answer" into a single, coherent text. This text must be simple, easy to understand, and kept within 100 words.

Input Content:
The provided instruction, question text and question images are as follows:
Original instruction: <instruction>
Original question text: <question text>
Original question images: <question images>

Figure 3: Prompt used by GPT-5 for query rewriting.

G More Details of Reranking

G.1 Implementation Details

For text-only rerankers, following the second input configuration described in Section D, we append image captions as auxiliary context. For multimodal rerankers, MLLMs are prompted in a *reason-then-rank* format; the full prompt is provided in Figure 4. We evaluate GPT-5 via its official API [10], and BGE-Reasoner-Reranker-32B with the authors' code and checkpoint obtained via email. For open-source MLLMs (Gemma-3-27B, Qwen2.5-VL-72B, GLM-4.5V), we run inference with SGLang [11] to accelerate the reasoning stage. All other models are evaluated using their released code and checkpoints.

G.2 Detailed Results

We report detailed reranking results for three retrievers (Qwen3-Embedding, GME, and Seed-1.6-Embedding) in Table 11, Table 12, and Table 13, respectively.

[10] gpt-5-2025-08-07
[11] https://docs.sglang.ai/

Task Description:
You are an objective, evidence-based multimodal judge. Given a Query and a Candidate, determine whether the Candidate appropriately corresponds to the Query (satisfies its requirements, answers its question, or retrieves the relevant information). Your task is to provide a discrete integer score from 0 to 100:
- 80-100 (Highly Relevant): The Candidate directly and comprehensively addresses the Query's intent.
- 60-80 (Relevant): The Candidate substantially addresses the Query's intent, providing most of the key information or details, but might miss some minor details.
- 40-60 (Moderately Relevant): The Candidate is relevant and addresses a part of the Query's intent, but it is not comprehensive.
- 20-40 (Slightly Relevant): The Candidate mentions some aspects about the Query, but its main intent is different. It offers very limited value or information.
- 0-20 (Irrelevant): The Candidate does not address the Query's intent at all and is off-topic or wrong.

Reasoning Process:
Before providing your answer, analyze the Query and the Candidate step by step and provide your analysis process:
1) **Query analysis**:
- If the Query contains image(s): analyze the concrete visual elements (objects, attributes, colors, materials, text-in-image/OCR, spatial relations, layout/scene, etc.).
- If the Query contains text(s): analyze the explicit intent and constraints (entities, attributes, quantities, relations, actions/edits, categories/styles, temporal/spatial cues, etc.).
- Accurately capture the Query's true intent, identifying the key challenges and core elements.
2) **Candidate analysis**:
- If the Candidate contains image(s): analyze the concrete visual elements (objects, attributes, colors, materials, text-in-image/OCR, spatial relations, layout/scene, etc.).
- If the Candidate contains text: analyze its explicit content (entities, attributes, quantities, relations, categories, etc.).
- Carefully analyze and discuss the Candidate against the Query's intent and constraints to determine whether it satisfies the Query's requirements and true intent. Avoid erroneous acceptance or rejection; base judgments strictly on observable details and reasonable reasoning.

After providing your detailed analysis and justification for all the steps above, conclude your entire response with the final score. The score must be enclosed within <score> </score> tags. Please output the score with the tag only, no other text.

Your output should follow the following format:
your analysis process
<score>XX</score>

Figure 4: Prompt used by MLLMs to score query-candidate pairs after reasoning.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Base Retriever													
Qwen3-Embedding	29.97	29.29	18.32	21.46	9.52	23.19	49.44	21.14	26.30	9.11	0.00	4.30	20.17
Textual Rerankers													
RankLLaMa-7B	30.58	28.35	14.86	23.71	10.44	26.29	48.66	13.48	34.31	11.03	0.00	9.39	20.92
RankLLaMa-14B	33.92	32.62	13.82	22.35	11.98	32.82	40.31	15.34	29.22	15.76	0.00	8.57	21.39
Rank1-7B	32.05	37.15	17.50	20.11	12.04	30.06	55.96	19.14	39.81	16.41	0.00	6.87	23.92
RankR1-14B	35.39	35.87	19.49	23.89	12.38	29.86	59.66	16.84	40.43	20.26	0.00	10.54	25.38
ReasonRank-32B	34.49	36.50	20.02	23.49	12.45	30.21	59.08	18.86	37.60	17.25	0.00	11.32	25.11
BGE-Reasoner-Reranker-32B	37.05	40.29	21.98	22.33	14.24	32.43	62.27	18.45	44.00	24.13	0.00	11.24	27.37
Multimodal Rerankers													
MonoQwen2-VL	31.61	35.58	17.13	19.82	9.49	23.25	60.35	14.81	55.25	13.87	0.24	11.42	24.40
Jina-Reranker	31.97	36.23	19.39	19.56	8.92	23.78	63.58	17.20	54.47	23.24	0.39	10.29	25.84
Gemma-3-27B	36.20	42.07	19.72	21.16	14.82	27.93	49.19	19.01	44.77	26.94	0.00	16.21	26.50
Qwen2.5-VL-72B	35.36	38.93	21.18	20.97	13.56	30.49	57.23	21.41	49.84	26.62	0.20	16.37	27.68
GLM-4.5V	35.60	41.26	18.95	21.10	14.53	28.70	55.39	20.21	52.55	30.73	0.62	17.83	28.12

Table 11: Detailed reranking performance (nDCG@10) on MR^2-Bench with Qwen3-Embedding as the base retriever.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Base Retriever													
GME	34.34	39.50	19.04	19.29	7.73	28.59	36.95	7.19	39.35	15.70	0.22	11.11	21.59
Textual Rerankers													
RankLLaMa-7B	30.58	28.35	14.86	23.71	10.44	26.29	48.66	13.48	34.31	11.03	0.00	9.39	20.92
RankLLaMa-14B	33.92	32.62	13.82	22.35	11.98	32.82	40.31	15.34	29.22	15.76	0.00	8.57	21.39
Rank1-7B	32.05	37.15	17.50	20.11	12.04	30.06	55.96	19.14	39.81	16.41	0.00	6.87	23.92
RankR1-14B	35.39	35.87	19.49	23.89	12.38	29.86	59.66	16.84	40.43	20.26	0.00	10.54	25.38
ReasonRank-32B	34.49	36.50	20.02	23.49	12.45	30.21	59.08	18.86	37.60	17.25	0.00	11.32	25.11
BGE-Reasoner-Reranker-32B	37.05	40.29	21.98	22.33	14.24	32.43	62.27	18.45	44.00	24.13	0.00	11.24	27.37
Multimodal Rerankers													
MonoQwen2-VL	31.61	35.58	17.13	19.82	9.49	23.25	60.35	14.81	55.25	13.87	0.24	11.42	24.40
Jina-Reranker	31.97	36.23	19.39	19.56	8.92	23.78	63.58	17.20	54.47	23.24	0.39	10.29	25.84
Gemma-3-27B	36.20	42.07	19.72	21.16	14.82	27.93	49.19	19.01	44.77	26.94	0.00	16.21	26.50
Qwen2.5-VL-72B	35.36	38.93	21.18	20.97	13.56	30.49	57.23	21.41	49.84	26.62	0.20	16.37	27.68
GLM-4.5V	35.60	41.26	18.95	21.10	14.53	28.70	55.39	20.21	52.55	30.73	0.62	17.83	28.12

Table 12: Detailed reranking performance (nDCG@10) on MR^2-Bench with GME as the base retriever.

Methods	Multimodal Knowledge Retrieval						Visual Illustration			Visual Relation			Avg.
	Bio.	Cook.	Gar.	Phy.	Chem.	Earth.	Econ.	Math.	Nat.	Spa.	Puzz.	Ana.	
Base Retriever													
Seed-1.6-Embedding	40.64	38.12	31.77	27.91	17.80	37.17	56.13	26.10	65.16	17.29	0.93	9.21	30.68
Textual Rerankers													
RankLLaMa-7B	37.92	34.53	30.40	27.74	16.31	30.35	54.62	32.71	40.76	15.21	1.04	6.42	27.33
RankLLaMa-14B	43.27	38.94	29.61	26.92	20.32	37.03	45.98	34.65	35.08	16.62	1.49	7.20	28.09
Rank1-7B	36.95	35.07	24.49	25.34	21.21	33.87	67.19	45.01	47.22	17.39	1.96	6.29	30.21
RankR1-14B	40.11	34.96	26.88	28.58	19.60	34.36	77.49	39.56	46.30	24.41	1.93	13.19	32.28
ReasonRank-32B	41.70	37.31	28.98	28.34	18.61	35.57	72.19	43.53	45.35	21.10	2.30	15.16	32.51
BGE-Reasoner-Reranker-32B	45.19	39.31	32.18	28.57	20.69	39.26	76.53	44.35	53.07	28.35	2.09	11.31	35.08
Multimodal Rerankers													
MonoQwen2-VL	35.33	36.41	24.71	23.96	15.31	27.96	70.60	38.93	64.83	18.23	1.14	12.28	30.64
Jina-Reranker	34.23	35.45	28.13	24.25	15.67	25.86	79.48	43.21	63.63	31.90	0.60	11.35	32.82
Gemma-3-27B	39.94	39.67	26.15	25.57	25.11	33.94	59.75	44.20	54.44	34.99	1.96	16.44	33.51
Qwen2.5-VL-72B	42.95	38.78	29.60	28.21	21.17	37.66	72.57	51.09	61.47	31.97	4.58	14.29	36.20
GLM-4.5V-thinking	42.43	41.28	26.37	29.78	24.34	34.15	70.52	50.24	60.24	36.06	3.78	17.29	36.37
GPT-5	52.35	55.41	37.46	37.10	31.96	51.12	83.83	55.63	79.16	41.41	3.94	21.48	45.90

Table 13: Detailed reranking performance (nDCG@10) on MR^2-Bench with Seed-1.6-Embedding as the base retriever.