ARK: A Dual-Axis Multimodal Retrieval Benchmark along Reasoning and Knowledge

Yijie Lin[1,4]* **Guofeng Ding**[2]* **Haochen Zhou**[2]* **Haobin Li**[2] **Mouxing Yang**[2] **Xi Peng**[1,3]†

[1]School of Artificial Intelligence, Sichuan University, Chengdu, China
[2]College of Computer Science, Sichuan University, Chengdu, China
[3]National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, Sichuan University, Chengdu, China
[4]The Hong Kong University of Science and Technology, Hong Kong SAR, China

{linyijie.gm,guofengding.gm,haochenzhou.xl,haobinli.gm,yangmouxing,pengx.gm}@gmail.com

https://huggingface.co/datasets/XLearning-SCU/ARK-Bench

Abstract

Existing multimodal retrieval benchmarks largely emphasize semantic matching on daily-life images and offer limited diagnostics of professional knowledge and complex reasoning. To address this gap, we introduce ARK, a benchmark designed to analyze multimodal retrieval from two complementary perspectives: (i) knowledge domains (five domains with 17 subtypes), which characterize the content and expertise retrieval relies on, and (ii) reasoning skills (six categories), which characterize the type of inference over multimodal evidence required to identify the correct candidate. Specifically, ARK evaluates retrieval with both unimodal and multimodal queries and candidates, covering 16 heterogeneous visual data types. To avoid shortcut matching during evaluation, most queries are paired with targeted hard negatives that require multi-step reasoning. We evaluate 25 representative text-based and multimodal retrievers and observe a pronounced gap between knowledge- and reasoning-intensive retrieval, with fine-grained visual and spatial reasoning as persistent bottlenecks. We further show that enhancements such as re-ranking, rewriting, and agentic retrieval yield consistent gains, but substantial headroom remains.

1 Introduction

Multimodal retrieval is a fundamental capability for modern AI systems, supporting applications such as recommender systems [1, 2], retrieval-augmented generation (RAG) [3, 4, 5], and Deep Research [6]. To assess progress, the community has developed a variety of benchmarks. In most settings, relevance between a query and candidate documents can be captured by relatively simple matching patterns, *e.g.*, shared object categories or surface-level semantic similarity [7, 8] as shown in Fig. 1(a).

However, in real-world scenarios, relevance is often established through nuanced and complex connections that go far beyond direct semantic overlap, requiring substantial domain knowledge and multi-step reasoning to retrieve the correct images or documents. As illustrated in Fig. 1(c), a historian may search for comics that implicitly encode a specific historical event, where relevance hinges on interpreting visual evidence (*e.g.*, characters and symbols) together with query-specific context (*e.g.*,

*Equal contribution
†Corresponding author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

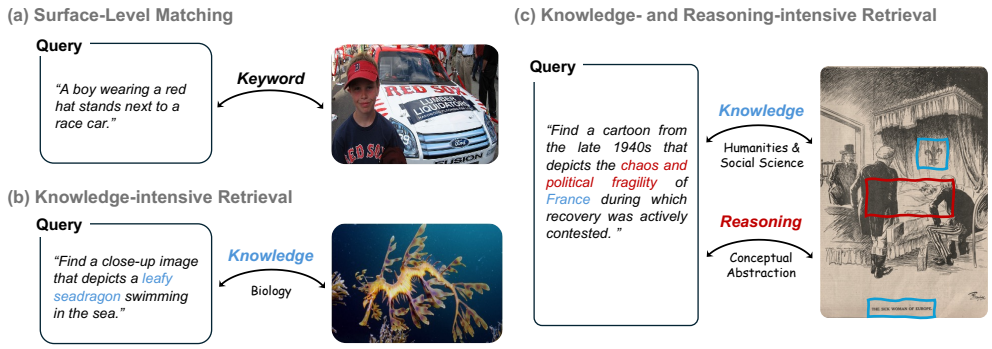


Figure 1: **Evolution of multimodal retrieval.** (a) Traditional benchmarks emphasize object-level or surface-level semantic similarity matching on daily imagery. (b) Recent benchmarks shift toward *knowledge-intensive* retrieval, where relevance hinges on domain knowledge beyond surface matching. (c) ARK explicitly separates *knowledge* and *reasoning* as two evaluation axes, enabling diagnosis of retrieval that requires both domain knowledge and structured reasoning over multimodal evidence.

temporal cues). In such cases, neither text-only nor vision-only cues suffice, and accurate retrieval depends on combining domain knowledge with reasoning over multimodal evidence.

To systematically evaluate such capabilities, we introduce a dual-Axis multimodal retrieval benchmark for Reasoning and Knowledge (ARK). Specifically, ARK is designed to assess retrieval systems along two orthogonal and complementary dimensions: i) knowledge-intensive scenarios, where correct retrieval mainly depends on recognizing and recalling domain concepts, and ii) reasoning-intensive scenarios, where retrieval requires multi-step inference, relational understanding, or abstraction. Compared with existing multimodal retrieval benchmarks, ARK provides several distinctive contributions:

- **The first multimodal retrieval benchmark that explicitly separates knowledge and reasoning as two evaluation axes**. Existing benchmarks [9, 10] conflate these two factors, obscuring whether failures stem from missing domain knowledge or insufficient reasoning. As illustrated in Fig. 1(b), retrieving images of "leafy seadragon" is largely knowledge-intensive but reasoning-light. Once the visual concept is known, the task can be solved at a glance, which may overestimate models that primarily rely on memorizing visual concepts. In contrast, ARK annotates each instance along both axes, enabling systematic analysis across knowledge domains and reasoning skills.
- **Broad coverage of knowledge domains and data types**. Most existing datasets focus on daily-life scenarios with natural images [11, 12, 13, 14], providing limited evaluation of professional or technical knowledge. Along the knowledge axis, ARK spans five domains: *visual cognition, natural science, formal science, humanities and social science, and engineering and technology* (Fig. 2(a)), and includes 16 heterogeneous visual types, such as *tables, line charts, chemical structures, artworks, comics, and cognitive maps* (Fig. 2(b)). This diversity enables evaluation of retrieval systems that require specialized knowledge beyond commonsense visual understanding.
- **Systematic evaluation of diverse reasoning skills**. ARK evaluates six reasoning categories: *knowledge reasoning, spatial reasoning, logical reasoning, symbolic reasoning, fine-grained visual reasoning, and conceptual abstraction* (Fig. 2(d)). These categories cover a broader and more fine-grained spectrum of reasoning phenomena than prior benchmarks [10]. To discourage shortcut matching, we carefully design queries and targeted hard negatives so that correct retrieval relies on genuine reasoning rather than superficial semantic cues.

Based on ARK, we evaluate 25 representative retriever configurations, including embedding-based retrievers, re-rankers, and agentic retrievers, spanning diverse model sizes and architectures, and obtain three key findings: (i) While current systems handle many knowledge-intensive queries reasonably well, they fall sharply behind on reasoning-intensive tasks. Performance also varies substantially across disciplines, revealing limited generalization and pronounced domain dependence. Together, these results indicate that closing the gap in multimodal retrieval requires not only broader knowledge coverage but also stronger and more transferable reasoning capabilities. (ii) Vision-centric reasoning is a key bottleneck on ARK. Breaking down performance by reasoning skill, we observe a consistent trend across models: performance is relatively stronger on knowledge reasoning and conceptual abstraction, but drops markedly on vision-centric skills such as fine-grained visual

reasoning and spatial reasoning. These results suggest that current retrievers struggle to localize subtle evidence in high-resolution images and to reason over geometric or topological relations. In short, visual grounding and spatial inference remain key obstacles for reasoning-intensive multimodal retrieval. (iii) While larger models generally achieve stronger retrieval performance, scaling alone does not eliminate the core reasoning bottlenecks on ARK. In contrast, explicitly injecting reasoning at inference time through lightweight interventions such as query rewriting and re-ranking yields consistent gains, highlighting a practical path to improve reasoning-intensive multimodal retrieval.

We hope that ARK and our empirical findings will help illuminate current limitations and guide the development of more robust and reasoning-aware multimodal retrieval systems.

2 Related Work

2.1 Benchmarks for Multimodal Retrieval

Early benchmarks mainly focus on text–image [7, 8] and text–video [15, 16, 17] matching based on semantic relevance, and have been further extended to composed image retrieval [18] and interactive retrieval [19, 20]. However, most existing benchmarks are dominated by everyday natural images and videos, offering limited coverage of professional knowledge and placing little emphasis on reasoning.

Very recently, MRMR [9] and MR^2-Bench [10] introduce more challenging retrieval settings with reasoning requirements. However, many of their queries are adapted from knowledge-heavy VQA sources, which tend to emphasize domain knowledge (*e.g.*, nature, cooking, medicine) rather than providing systematic coverage and diagnostics of reasoning skills. Moreover, the reasoning they probe is typically limited to relatively simple visual relations. In contrast, ARK targets a broader spectrum of reasoning demands, including richer spatial reasoning (*e.g.*, 3D projection/folding/rotation and cognitive-map understanding), fine-grained visual reasoning, and higher-level logical, symbolic, and abstract inference. Overall, ARK complements these efforts by explicitly separating knowledge and reasoning as two evaluation axes and enabling finer-grained evaluation across diverse domains and reasoning types. A detailed comparison with existing benchmarks is provided in Appendix A.

2.2 Multimodal Retrievers

Progress in multimodal retrieval [21] has closely tracked advances in model architectures and pretraining paradigms. Early retrievers typically adopt dual-stream designs that encode images and texts separately and align them in a shared embedding space for efficient similarity search [22, 23, 24, 25, 26]. With the rise of multimodal foundation models, an increasing number of methods move toward re-ranking architectures that model cross-modal interactions over concatenated inputs [27, 28, 29]. Recently, retrievers built on large language models (LLMs) have demonstrated strong retrieval capability [30, 31, 32, 33].

However, most of these approaches still frame retrieval primarily as a semantic similarity matching problem and therefore do not explicitly elicit multi-step reasoning during search. Inspired by recent progress on reasoning-augmented text retrievers [34, 35], emerging work has started to inject reasoning signals into multimodal retrieval, *e.g.*, via reasoning-enhanced fine-tuning [36] or by generating richer textual contexts [37] to improve embeddings. In this work, we systematically evaluate representative multimodal retrievers on ARK and analyze how model architecture, scale, and reasoning-oriented training affect both knowledge-intensive and reasoning-intensive retrieval performance.

3 ARK Benchmark

We introduce ARK, a knowledge- and reasoning-intensive benchmark for multimodal retrieval. In this section, we first present the task formulation in Section 3.1, then describe the dual-axis taxonomy of knowledge and reasoning that structures ARK in Section 3.2, and finally detail the data curation pipeline in Section 3.3.

Table 1: Data statistics of queries and gallery from the knowledge perspective in ARK benchmark.

Knowledge Domain	Subtype	Query	Gallery	Knowledge Domain	Subtype	Query	Gallery
NATURAL SCIENCE	Biology	83	1834	VISUAL COGNITION	Daily-Life	236	5248
	Physics	80	712		Fine-Grained	92	2101
	Chemistry	105	1127		Spatial-Cogmap	81	1270
FORMAL SCIENCE	Mathematics	152	889		Spatial-3D	131	7500
	Code-Drawing	100	551	HUMANITY & SOCIAL SCIENCE	Economics	80	2614
ENGINEERING & TECHNOLOGY	Computer-Science	80	5231		Comic	87	3468
	Mechanical	82	917		Art	100	2101
	Electronics	101	619		Calendar	80	378
					Metro	80	298
5 Domains	**17 Subtypes**	**1,750 Queries**		**36,858 Gallery Candidates**			

3.1 Task Formulation

ARK evaluates retrieval with both unimodal and multimodal queries and candidates. Given a query q and the gallery $\mathcal{D} = \{d_i\}_{i=1}^{N}$, the goal is to rank all candidates such that the ground-truth relevant item d^+ is ranked as high as possible.

Each query in ARK is associated with exactly one positive candidate and, in most cases, a set of carefully designed hard negatives that are semantically related but incorrect. These negatives are intentionally constructed to discourage shortcut matching where they may partially satisfy the query while violating the required relations or reasoning steps. As a result, correct retrieval cannot be achieved by relying on partial cues and instead requires following the full query and performing the intended reasoning, together with the necessary domain knowledge.

3.2 Knowledge and Reasoning Axes

ARK is structured along two orthogonal axes, *i.e.*, *knowledge domains* and *reasoning skills*. The knowledge axis captures the types of domain knowledge that retrieval instances rely on, while the reasoning axis reflects the nature and amount of inference required to resolve a query during retrieval. This dual-axis design facilitates fine-grained analysis of model strengths and weaknesses from complementary perspectives and helps avoid conflating domain knowledge with reasoning capability.

Knowledge axis. Table 1 summarizes ARK from the knowledge perspective, comprising 17 subtypes across five knowledge domains, with 1,750 queries and 36,858 gallery candidates in total. Along this axis, ARK is organized into five meta-domains: (i) *Visual Cognition* serves as a perceptual baseline. It evaluates everyday recognition and fine-grained perception of subtle details in high-resolution images, aligning with recent interest in "thinking with images" [38, 39]. It also covers spatial perception over structured visual objects; (ii) *Natural Science*, which requires scientific knowledge of concepts and phenomena (*e.g.*, biology, physics, and chemistry); (iii) *Formal Science*, which emphasizes symbolic and formal representations, including math problems and code-specified graphics; (iv) *Humanities & Social Science*, which involves culturally and socially grounded content, including history-, art-, and society-related materials; (v) *Engineering & Technology*, which covers technical domains such as computer science and engineering schematics, emphasizing understanding of technical diagrams and designs. For example, it includes extracting high-level conclusions from experimental results by identifying empirical trends such as *scaling law*. Across domains, ARK spans 16 heterogeneous visual types including natural images, tables, line/bar charts, chemical structure diagrams, artworks, comics, metro maps, calendars, 3D renderings, cognitive maps, code, design drawings, circuit schematics, symbolic mathematical plots, and physics phenomenon illustrations, enabling evaluation beyond common-sense visual understanding (Fig. 2(b)). Although modest compared with large-scale retrieval datasets, ARK is manually curated for reasoning-oriented evaluation: each instance requires query construction, positive-target verification, hard-negative validation, and reasoning-label annotation. Its scale is therefore comparable to recent reasoning-focused retrieval benchmarks [9, 40, 41, 42], where careful per-instance curation is more labor-intensive than collecting

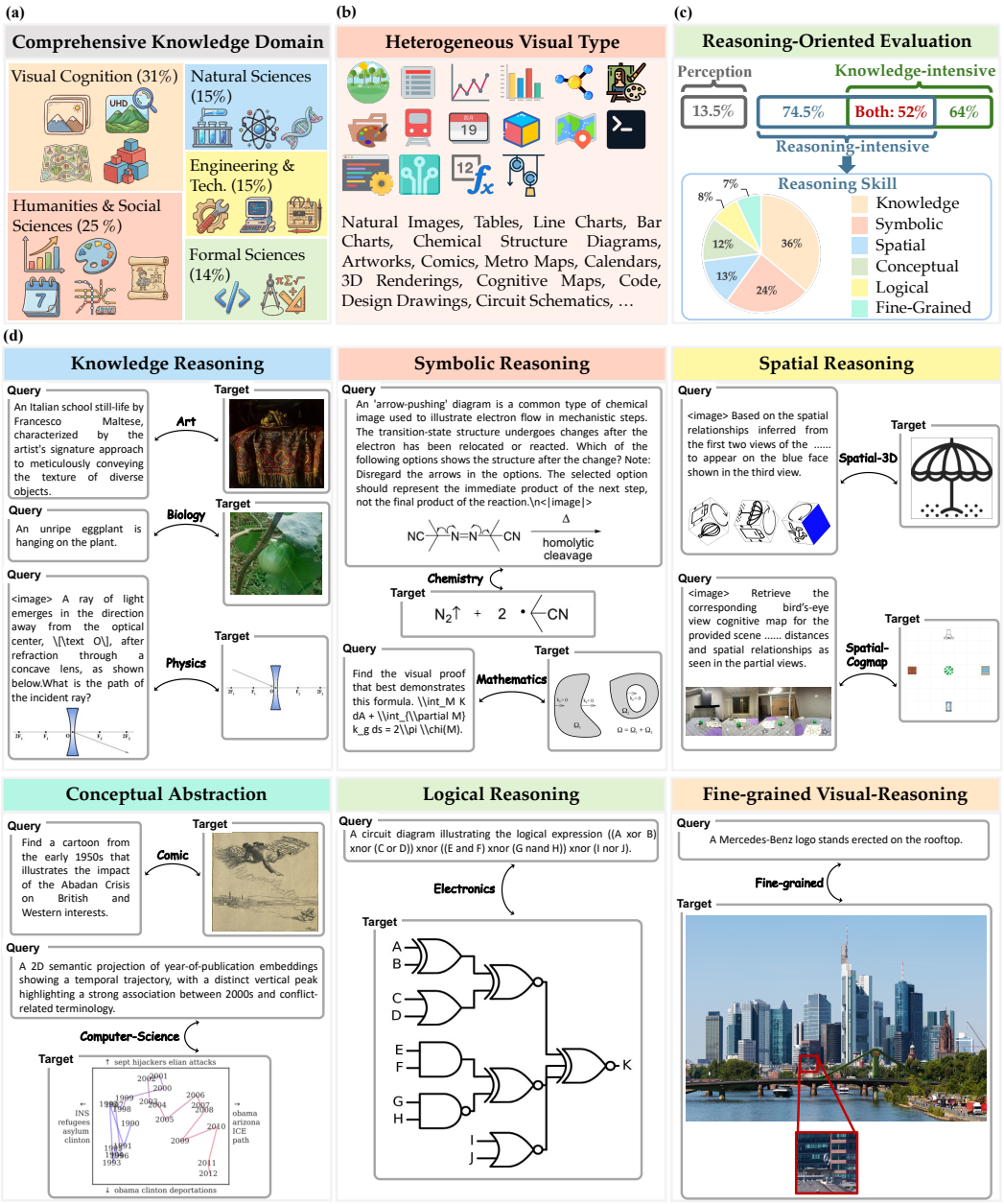


Figure 2: ARK is a dual-axis benchmark for multimodal retrieval. (a)–(c) summarize three key characteristics: (a) **Comprehensive knowledge domains**: it covers five broad knowledge domains and 17 fine-grained subtypes, (b) **Heterogeneous visual types**: queries and candidates span diverse visual formats beyond everyday imagery, and (c) **Reasoning-oriented evaluation**: each instance in ARK is first labeled by its overall cognitive demand as *perception-only*, *knowledge-intensive*, *reasoning-intensive*, or *both knowledge- and reasoning-intensive*. Reasoning-intensive instances are further annotated with six reasoning-skill categories, enabling fine-grained analysis of model behavior. Note that a single instance may involve multiple reasoning skills. (d) Examples: representative instances of each reasoning skill. For reference, the primary reasoning skills and visual data types for each knowledge subtype are summarized in Table 5.

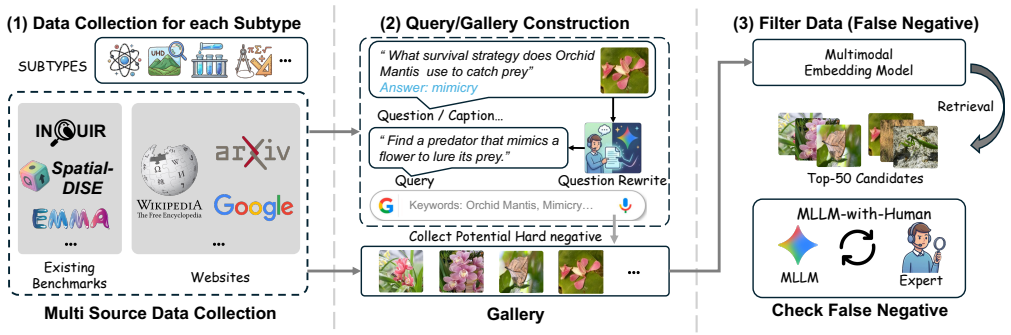


Figure 3: Illustration of the data curation process for ARK.

surface-level semantic matching pairs. More details and examples for every subtype are provided in Appendix B.

Reasoning axis. Fig. 2(c) reports the distribution of instances by required reasoning skills, and Fig. 2(d) provides representative examples for each category. Along this axis, ARK organizes instances into six reasoning categories: (i) *Knowledge Reasoning* covers queries that require inference grounded in background knowledge, beyond mere recognition or direct factual retrieval. (ii) *Spatial Reasoning* requires understanding geometric and topological relations, such as spatial layouts in cognitive maps and metro networks, as well as 3D object projection, folding, and rotation. (iii) *Fine-grained Visual Reasoning* requires detail-sensitive visual inference in high-resolution images, where the target can only be identified by subtle, localized cues [43]. This capability closely aligns with the recent emphasis on "thinking with images" for careful visual inspection and verification [38]. (iv) *Logical Reasoning* involves structured deduction over conditions and relations (*e.g.*, conjunction/disjunction/negation, implication, and Boolean expressions such as AND/OR/NOT), where correct retrieval depends on satisfying explicit logical constraints rather than surface-level semantic cues. (v) *Symbolic Reasoning* focuses on interpreting structured symbolic representations and grounding them into the corresponding visual structures, such as mapping mathematical expressions to plots, code-like specifications to diagrams, or reasoning over chemical structure diagrams to infer the resulting molecular structure. (vi) *Conceptual Abstraction* evaluates higher-level inference over patterns, analogies, and inductive generalization, such as interpreting historical cartoons or summarizing high-level conclusions from experimental results in scientific papers.

Importantly, each retrieval instance in ARK is annotated along both knowledge and reasoning axes, enabling fine-grained analysis across knowledge, reasoning, and their interactions. In addition, each instance is labeled by its overall cognitive demand as *perception-only*, *knowledge-intensive*, *reasoning-intensive*, or *both knowledge- and reasoning-intensive*. Fig. 2(c) summarizes the distribution and Table 5 reports the primary reasoning skills and visual data types associated with each knowledge subtype.

3.3 Data Curation Process

We curate ARK through a multi-stage pipeline that combines taxonomy-driven collection, automated refinement, and human verification, as illustrated in Fig. 3.

Taxonomy-driven collection. We first define a taxonomy spanning five knowledge domains and six reasoning dimensions. Guided by this framework, we curate the gallery by selecting 17 subtypes across the five domains and collecting high-quality multimodal query–target and question–answer pairs from authoritative sources and expert-level benchmarks, forming the initial pool of positive pairs and gallery items.

Reasoning-aligned refinement and quality control. After collecting the initial pairs, we use multimodal LLMs [44] to rewrite the original queries and questions, ensuring better alignment with the intended knowledge constraints and reasoning requirements. For query–target pairs, we rewrite queries to remove trivial cues (*e.g.*, overly explicit keywords) while preserving the event background and the core reasoning intent. For question–answer pairs, we convert the original questions into

Table 2: Domain-wise Recall@1 on ARK (knowledge axis). Columns correspond to ARK's five knowledge domains and corresponding subtypes. "Avg." denotes the macro-average across all subtypes. Models within each group are ordered by parameter size. The **bold** and underlined numbers indicate the best and second-best results within each subtype, respectively.

Model	Visual Cognition				Natural Science			Formal Science		Humanities & Social Science					Engineering & Technology			Avg.
	Daily.	Fine.	SCog.	S3D.	Bio.	Phy.	Che.	Mat.	Cod.	Eco.	Comic.	Art	Cal.	Met.	Compu.	Mecha.	Ele.	
Text Embedding Models																		
BGE-M3-0.6B	37.29	2.17	0.00	0.00	32.53	15.00	0.95	3.29	2.00	36.25	1.15	2.00	6.25	7.50	25.00	9.76	0.99	10.71
Diver-Embed-1.7B	35.59	2.17	0.00	0.00	26.51	25.00	2.86	11.84	2.00	23.75	1.15	4.00	8.75	6.25	38.75	4.88	0.00	11.38
Diver-Embed-4B	46.19	1.09	0.00	0.76	39.76	22.50	2.86	**14.47**	2.00	**55.00**	2.30	5.00	6.25	10.00	46.25	10.98	0.00	15.61
Qwen3-Embed-4B	42.37	2.17	**3.70**	0.76	27.71	21.25	1.90	10.53	3.00	36.25	1.15	4.00	5.00	3.75	33.75	10.98	2.97	12.43
Qwen3-Embed-8B	46.19	1.09	0.00	0.76	40.96	22.50	4.76	12.50	4.00	47.50	2.30	4.00	7.50	2.50	46.25	8.54	1.98	14.90
ReasonIR-8B	40.25	1.09	0.00	0.00	28.92	20.00	1.90	10.53	1.00	42.50	4.60	1.00	3.75	13.75	36.25	9.76	0.99	12.72
Multimodal Embedding Models																		
CLIP-0.4B	37.29	0.00	1.23	3.05	26.51	7.50	3.81	1.32	1.00	7.50	3.45	10.00	5.00	3.75	7.50	7.32	3.96	7.66
BGE-VL-0.4B	52.97	0.00	0.00	0.00	32.53	0.00	0.00	1.32	0.00	6.25	0.00	5.00	1.25	2.50	10.00	7.32	1.98	7.12
SigLIP2-0.9B	68.22	2.17	0.00	1.53	33.73	3.75	1.90	4.61	0.00	13.75	5.75	3.00	6.25	1.25	20.00	14.63	0.99	10.68
MetaCLIP2-2B	60.17	3.26	0.00	3.05	33.73	20.00	1.90	1.32	3.00	21.25	3.45	**20.00**	1.25	3.75	12.50	13.41	2.97	12.06
VLM2VecV2-2B	43.22	1.09	0.00	0.76	24.10	6.25	2.86	7.89	4.00	32.50	2.30	4.00	2.50	10.00	22.50	7.32	0.00	10.08
Ops-MM-Embed-v1-2B	63.14	1.09	1.23	2.29	38.55	10.00	3.81	6.58	3.00	21.25	8.05	7.00	3.75	13.75	23.75	9.76	2.97	12.94
Qwen3-VL-Embed-2B	65.25	1.09	0.00	4.58	27.71	21.25	0.95	7.24	4.00	38.75	4.60	3.00	2.50	6.25	26.25	8.54	3.96	13.29
LamRA-Ret-7B	60.17	**5.43**	0.00	0.00	38.55	1.25	2.86	9.87	6.00	35.00	3.45	8.00	2.50	5.00	5.00	7.32	0.99	11.26
RzenEmbed-7B	**69.92**	2.17	1.23	1.53	32.53	10.00	5.71	11.84	7.00	43.75	2.30	4.00	6.25	7.50	48.75	15.85	4.95	16.19
Ops-MM-Embed-v1-7B	62.29	1.09	1.23	**5.34**	42.17	15.00	3.81	9.21	6.00	41.25	6.90	13.00	5.00	**17.50**	42.50	12.20	5.94	17.08
Qwen3-VL-Embed-8B	69.07	0.00	0.00	3.82	39.76	**26.25**	3.81	9.87	14.00	51.25	11.49	4.00	3.75	7.50	46.25	13.41	**6.93**	18.30
EVA-CLIP-18B	63.98	2.17	0.00	2.29	**43.37**	13.75	**6.67**	0.66	3.00	18.75	0.00	18.00	3.75	3.75	16.25	**21.95**	3.96	13.08
Seed1.6-embedding	65.25	4.35	0.00	3.82	31.33	18.75	3.81	11.84	**18.00**	45.00	**12.64**	10.00	**13.75**	16.25	**57.50**	12.20	3.96	**19.32**

retrieval-style queries, keeping the underlying reasoning premises and constraint conditions required to identify the correct evidence. To discourage shortcut matching during evaluation, we construct targeted hard negatives whenever feasible. Specifically, for each rewritten query, we first mine visually or semantically similar candidates via web search, guided by keywords and topical cues suggested by the large model [44], and then manually select distractors that share superficial cues with the positive target while violating the underlying conditions. To further reduce false negatives [45, 46, 47], we embed each query with an up-to-date retriever [48], retrieve its top-50 nearest candidates, and manually remove ambiguous items, spurious negatives, and near-duplicates.

Human-in-the-loop curation. ARK construction involves two groups of human participants with distinct roles. First, trained contributors and domain-aware reviewers, including PhD students and undergraduate annotators from biology, mathematics, computer science, and related disciplines, participated in data collection, query refinement, hard-negative construction, and consistency checking. Second, reasoning-related labels, including cognitive demand and reasoning skills, were independently annotated by six PhD annotators under unified guidelines, with disagreements resolved through panel adjudication. This separation supports domain-aware data construction while promoting consistent multi-label annotation [49, 50, 51] of reasoning skills. Details of the annotation protocol and consistency analysis are provided in Appendix J.

4 Experiments

We first evaluate embedding-based retrievers on ARK, covering both text-only and multimodal models. We then examine two inference-time enhancement strategies, query rewriting and re-ranking, and further evaluate agentic retrievers in a multi-round retrieval setting in Appendix G. Details of evaluated models are provided in Appendix C. We use Recall@1 as the primary metric, as it directly measures whether the correct target is ranked first. To provide a more complete view of retrieval quality, we also report Recall@5, Recall@10, nDCG@5, nDCG@10, and nDCG@20 in Appendix E.

4.1 Evaluation across Knowledge Domains

We summarize the overall evaluation results of all investigated retrieval baselines on ARK in Table 2. All experiments are conducted within each individual subtype using its own dedicated retrieval corpus, ensuring that retrieval difficulty and domain coverage are controlled per subtype. For each subtype, we report Recall@1, together with the macro-average across all 17 subtypes. For text retrievers, we first convert visual inputs into textual descriptions using Qwen3-VL-8B-Instruct [52] (with the prompt shown in Table 26), and then perform retrieval in the text embedding space based on the generated captions. Based on these results, we draw several key observations:

Table 3: Recall@1 on ARK by cognitive demand and reasoning skill (reasoning axis). "Both" denotes instances that are both knowledge-intensive and reasoning-intensive.

Model	Cognitive Demand				Reasoning Skill						Average
	Perception-Only	Knowledge-Intensive	Reasoning-Intensive	Both	Knowledge Reasoning	Conceptual Abstraction	Fine-Grained Visual Reasoning	Logical Reasoning	Spatial Reasoning	Symbolic Reasoning	
Text Embedding Models											
BGE-M3-0.6B	37.29	12.09	7.89	10.12	11.70	18.34	3.08	4.29	2.31	3.90	7.27
Diver-Embed-1.7B	36.02	14.86	9.88	13.20	14.84	21.83	2.31	5.52	2.69	7.79	9.16
Diver-Embed-4B	46.19	17.73	11.94	15.95	18.12	27.07	3.85	6.13	1.54	9.31	11.00
Qwen3-Embed-4B	41.53	12.09	7.73	9.90	11.98	17.90	1.54	4.29	2.69	5.19	7.26
Qwen3-Embed-8B	45.76	15.67	9.57	13.09	15.98	23.58	0.77	4.91	1.54	6.71	8.92
ReasonIR-8B	40.25	15.31	11.18	14.41	16.41	28.38	3.85	3.07	4.23	7.58	10.59
Multimodal Embedding Models											
CLIP-0.4B	37.29	6.54	4.13	4.84	5.56	5.68	0.77	3.68	3.08	2.38	3.52
BGE-VL-0.4B	53.39	5.10	2.60	3.52	4.14	3.93	1.54	1.84	0.38	1.08	2.15
SigLIP2-0.9B	68.22	8.24	5.28	6.93	8.70	8.73	2.31	3.07	0.77	2.38	4.33
MetaCLIP2-2B	60.59	10.38	6.58	8.14	9.56	11.79	3.85	2.45	3.08	2.16	5.48
VLM2VecV2-2B	44.07	8.86	6.66	8.25	10.41	16.16	3.85	1.23	2.31	4.11	6.34
Ops-MM-Embed-v1-2B	63.98	10.92	7.73	9.24	11.13	13.97	4.62	3.68	4.23	4.55	7.03
Qwen3-VL-Embed-2B	65.68	12.00	8.58	11.11	12.98	23.58	3.08	4.29	3.08	5.19	8.70
LamRA-Ret-7B	59.32	9.67	6.81	8.69	10.27	14.41	4.62	2.45	1.15	5.41	6.39
RzenEmbed-7B	**70.34**	15.04	10.57	13.75	16.12	22.71	4.62	6.13	2.31	8.01	9.98
Ops-MM-Embed-v1-7B	63.56	16.11	12.02	14.63	17.12	23.58	**6.92**	6.13	5.77	6.93	11.08
Qwen3-VL-Embed-8B	68.64	17.73	11.94	15.95	17.55	30.13	0.77	6.75	3.85	8.87	11.32
EVA-CLIP-18B	63.98	11.46	6.43	8.14	9.42	9.61	2.31	3.68	2.31	3.46	5.13
Seed1.6-embedding	62.29	**22.33**	**16.85**	**21.70**	**22.08**	**33.33**	4.35	**10.71**	**6.19**	**15.16**	**15.30**

Observation 1: ARK is substantially more challenging and diagnostic than prior retrieval benchmarks. Performance on ARK remains low even for the strongest retrievers. The best embedding model, Seed1.6 [48], achieves only 19.32 R@1 in macro-average. This is in sharp contrast to its reported 75.59 R@1 on the embedding-focused MMEB-v2 leaderboard [53], where Qwen3-VL-Embed [32] also reaches 77.82 R@1. A similar gap appears for text retrieval: Qwen3-Embed-8B [54] obtains 54.1 nDCG@10 on the reasoning-oriented MRMR benchmark [9], but only 24.27 nDCG@10 on ARK (Table 14). These results suggest that ARK is not well captured by surface-level semantic matching and exposes failure modes that are less visible in existing benchmarks.

Observation 2: Performance varies sharply across knowledge domains, exposing clear bottlenecks. Across models, retrieval is relatively strong on daily-life natural images but drops markedly on domains requiring fine-grained perception or spatial grounding. The Fine-Grained (high-resolution image), Spatial-Cogmap, and Spatial-3D subsets are among the hardest, indicating model brittleness under detail-sensitive visual inspection and non-trivial spatial reasoning. Beyond visual cognition, performance is highly subtype-dependent: models perform reasonably well on some knowledge-heavy subtypes such as Biology, Economics, and Computer Science, but remain weak on Chemistry, Mathematics, and Code-Drawing. This suggests that ARK's bottlenecks are not determined by broad knowledge domains alone, but by the interaction between domain knowledge, visual format, and reasoning demands.

Observation 3: Model rankings are not stable across domains, reflecting differences in pretraining bias and representational strengths. While several multimodal embedding models rank highly on the overall average, their advantages are not consistent across subtypes. For instance, Qwen3-VL-Embed performs better than Seed1.6 on knowledge-heavy categories such as biology and physics, but is weaker on high-resolution subsets such as Metro and Fine-Grained. This variability indicates that "best overall" does not translate to "best everywhere", and ARK's domain-level breakdown provides more actionable diagnosis than single-number evaluations.

Observation 4: Transforming multimodal retrieval to text-space via captioning yields a strong baseline, but is limited by caption fidelity. Reducing multimodal retrieval to text by captioning images and retrieving in a text embedding space [55] can be competitive on some subtypes, as shown by the results in the top part of Table 2. However, this pipeline is computationally costly and frequently discards critical visual evidence (*e.g.*, fine-grained details, spatial relations, and diagram structure). This limitation helps explain why reasoning-oriented text retrievers (ReasonIR [35], Diver [34]) do not consistently dominate on ARK: when the caption omits key cues, downstream reasoning has insufficient evidence.

4.2 Evaluation along Reasoning Skill

To better understand where current retrievers succeed or fail in reasoning-intensive retrieval, we analyze ARK performance from a reasoning-centric perspective. Specifically, we classify instances along cognitive demand and reasoning skill. Cognitive demand categorizes instances as *perception-only*, *knowledge-intensive*, *reasoning-intensive*, or *both knowledge- and reasoning-intensive*, reflecting whether a query is mainly solvable by knowledge recall, multi-step inference, or requires both. Reasoning skill specifies the cognitive operation involved, covering *knowledge reasoning*, *spatial reasoning*, *logical reasoning*, *symbolic reasoning*, *fine-grained visual reasoning*, and *conceptual abstraction*. Based on the results in Table 3, we have the following observations:

Observation 5: Reasoning-intensive retrieval is substantially harder than knowledge-intensive retrieval. From the cognitive-demand breakdown in Table 3, *perception-only* instances are consistently the easiest: most models perform strongly when retrieval reduces to basic visual recognition and cross-modal alignment. However, performance degrades noticeably on *knowledge-intensive* instances and drops further on *reasoning-intensive* instances, suggesting that multi-step inference remains a critical bottleneck beyond domain-knowledge matching. Instances labeled as *both knowledge- and reasoning-intensive* also remain challenging overall, underscoring the compounded difficulty when retrieval requires knowledge and reasoning jointly.

Observation 6: Models exhibit a consistent performance pattern across reasoning skills. As shown in Table 3, most retrievers follow a similar trend across skill categories: they perform relatively better on knowledge reasoning and conceptual abstraction, but struggle markedly with fine-grained visual reasoning and spatial reasoning. Performance on logical and symbolic reasoning lies in between, suggesting substantial room for improvement beyond scaling alone. These results suggest that future work may benefit from integrating vision-centric reasoning mechanisms such as "thinking with images" [38, 56] approaches into embedding-based retrievers, so that fine-grained and spatial visual evidence can be more effectively preserved and exploited during retrieval.

Observation 7: Model scale correlates with stronger retrieval, but does not eliminate reasoning bottlenecks. Tables 2 and 3 show a clear trend that larger models tend to achieve higher Recall@1 across both the knowledge and reasoning axes, suggesting that increased capacity improves alignment. However, the benefits are uneven: gains are most visible on knowledge-intensive queries and conceptual abstraction, while fine-grained visual, spatial, logical, and symbolic reasoning remain difficult even for the largest models. This suggests that scaling alone is unlikely to close these gaps, pointing to the need for additional architectural or training improvements beyond model size.

4.3 Query Rewriting and Re-ranking

We study two inference-time interventions for improving retrieval: *query rewriting* and *re-ranking*. For query rewriting, we use GPT-5.2 [57] to transform the original query into a semantically equivalent but more explicit form that highlights reasoning cues (prompt provided in Table 27). For re-ranking, we first retrieve the top-50 candidates using an embedding retriever, and then apply the reranker for each query–candidate pair to produce a refined ranking. Based on results in Tables 9 and 10, we draw the following observations:

Observation 8: Query rewriting and re-ranking generally improve retrieval, and they are complementary. As shown in Table 9, query rewriting generally improves the Seed1.6 embedding baseline, increasing R@1 from 19.32 to 23.42 on the knowledge axis and from 15.30 to 17.49 on the reasoning axis. Re-ranking further improves retrieval when the reranker has sufficient capacity: Jina-Reranker and Qwen3-VL-Reranker outperform the embedding-only baseline, while the 600M BGE-Reranker underperforms it. Combining query rewriting with Qwen3-VL-Reranker achieves the best overall results on most metrics. These results suggest complementary effects: rewriting strengthens query-side reasoning, while capable rerankers enable stronger pairwise inference over the top retrieved candidates.

Observation 9: Query rewriting helps most when queries under-specify instructions. As shown in Table 10, rewriting yields the largest gains for logical and knowledge reasoning, where queries often contain implicit rules or abstract conditions that can be made more explicit through rewriting. In contrast, improvements are limited for fine-grained visual and spatial reasoning, where the main bottleneck lies in extracting subtle visual details or precise geometric relations from candidates rather

than clarifying the query. This suggests that beyond query-side rewriting, enriching gallery items with query-guided, reasoning-aware descriptions [55] before re-ranking could be a promising direction.

5 Conclusion

We introduce ARK, a knowledge- and reasoning-intensive benchmark for multimodal retrieval that enables systematic diagnosis along both the knowledge and reasoning axes. ARK spans 17 knowledge subtypes and six reasoning categories, and pairs most queries with targeted hard negatives to reduce shortcut matching, thereby revealing failure modes that are largely masked by conventional semantic-matching benchmarks. Evaluations on 25 representative retrievers show a pronounced gap between knowledge-intensive and reasoning-intensive retrieval, with fine-grained visual and spatial reasoning remaining persistent bottlenecks. Although interventions such as query rewriting, re-ranking, and agentic retrieval provide consistent improvements, substantial headroom remains for advancing reasoning-aware multimodal retrieval. We hope ARK will support more diagnostic evaluation and foster progress toward retrieval models that better integrate domain knowledge with visual reasoning.

Author Contributions

All authors contributed significantly to this work. Xi Peng conceived the study and supervised the project. Yijie Lin and Guofeng Ding led the design of the ARK dataset, while Yijie Lin, Guofeng Ding, Haochen Zhou, Haobin Li, and Mouxing Yang jointly organized and coordinated its construction with the annotators. Yijie Lin, Guofeng Ding, and Haochen Zhou developed the annotation protocol and analyzed the experimental findings. Guofeng Ding and Haochen Zhou implemented the evaluation pipeline. All authors reviewed and approved the final version.

Acknowledgments

This work was supported in part by Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China under Grant JYB2025XDXM610; in part by NSFC under Grants 624B2099 and U24B20174; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303, 20826041K4049; in part by Sichuan Science and Technology Planning Project under Grant 2024NSFTD0038; and in part by the Postdoctoral Fellowship Program and China Postdoctoral Science Foundation under Grant Number BX2026302.

References

[1] Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, et al. Towards ai search paradigm. *arXiv preprint arXiv:2506.17188*, 2025.

[2] Junjie Zhou, Ke Mei, Lei Li, Tianyi Wang, Fengyun Rao, and Jing Lyu. Wemm-embedding: Wechat multi-modal embedding technical report. *arXiv preprint arXiv:2608.24053*, 2026.

[3] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning*, pages 39755–39769. PMLR, 2023.

[4] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23369–23379, 2023.

[5] Zhe Xu, Zhengyu Zhang, Zhiyuan Cai, Jiahao Xu, Yijie Lin, Ziyi Liu, Junlin Hou, Hongyi Wang, Yuxiang Nie, Ling Liang, et al. A multi-modal agentic co-pilot for evidence grounded computational pathology. *arXiv preprint arXiv:2606.08093*, 2026.

[6] OpenAI. Deep research system card. `https://cdn.openai.com/deep-research-system-card.pdf`, 2025.

[7] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[8] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015.

[9] Siyue Zhang, Yuan Gao, Xiao Zhou, Yilun Zhao, Tingyu Song, Arman Cohan, Anh Tuan Luu, and Chen Zhao. Mrmr: A realistic and expert-level multidisciplinary benchmark for reasoning-intensive multimodal retrieval. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2026.

[10] Junjie Zhou, Ze Liu, Lei Xiong, Jin-Ge Yao, Yueze Wang, Shitao Xiao, Fenfen Lin, Miguel Hu Chen, Zhicheng Dou, Siqi Bao, et al. MR^2-bench: Going beyond matching to reasoning in multimodal retrieval. *arXiv preprint arXiv:2509.26378*, 2025.

[11] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 6761–6771, 2021.

[12] Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. Edis: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894, 2023.

[13] 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*, pages 11307–11317, 2021.

[14] 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*, pages 2125–2134, 2021.

[15] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

[16] David Chen and WilliamB. Dolan. Collecting highly parallel data for paraphrase evaluation. *Meeting of the Association for Computational Linguistics*, Jun 2011.

[17] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015.

[18] 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 *Proceedings of the 41st International Conference on Machine Learning*, pages 59403–59420, 2024.

[19] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017.

[20] Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. In *Forty-second International Conference on Machine Learning*, 2025.

[21] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. Vista: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, 2024.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[23] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. Eva-clip-18b: Scaling clip to 18 billion parameters. *arXiv preprint arXiv:2402.04252*, 2024.

[24] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.

[25] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Multi-granularity correspondence learning from long-term noisy videos. In *The Twelfth International Conference on Learning Representations*, 2024.

[26] Jun Li, Jinpeng Wang, Chaolei Tan, Niu Lian, Long Chen, Yaowei Wang, Min Zhang, Shu-Tao Xia, and Bin Chen. Enhancing partially relevant video retrieval with hyperbolic learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23074–23084, 2025.

[27] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[29] Zhenyu Huang, Mouxing Yang, Xinyan Xiao, Peng Hu, and Xi Peng. Noise-robust vision-language pre-training with positive-negative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[30] Yikun Liu, Yajie Zhang, Jiayin Cai, Xiaolong Jiang, Yao Hu, Jiangchao Yao, Yanfeng Wang, and Weidi Xie. Lamra: Large multimodal model as your advanced retrieval assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4015–4025, 2025.

[31] 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 *European Conference on Computer Vision*, pages 387–404. Springer, 2024.

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

[33] Weiquan Huang, Aoqi Wu, Yifan Yang, Xufang Luo, Yuqing Yang, Liang Hu, Qi Dai, Chunyu Wang, Xiyang Dai, Dongdong Chen, et al. Llm2clip: Powerful language model unlocks richer visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[34] Meixiu Long, Duolin Sun, Dan Yang, Junjie Wang, Yecheng Luo, Yue Shen, Jian Wang, Hualei Zhou, Chunxiao Guo, Peng Wei, et al. Diver: A multi-stage approach for reasoning-intensive information retrieval. *arXiv preprint arXiv:2508.07995*, 2025.

[35] 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.

[36] Mingjun Xu, Jinhan Dong, Jue Hou, Zehui Wang, Sihang Li, Zhifeng Gao, Renxin Zhong, and Hengxing Cai. Mm-r5: Multimodal reasoning-enhanced reranker via reinforcement learning for document retrieval. *arXiv preprint arXiv:2506.12364*, 2025.

[37] Xuanming Cui, Jianpeng Cheng, Hong-you Chen, Satya Narayan Shukla, Abhijeet Awasthi, Xichen Pan, Chaitanya Ahuja, Shlok Kumar Mishra, Yonghuan Yang, Jun Xiao, et al. Think then embed: Generative context improves multimodal embedding. *arXiv preprint arXiv:2510.05014*, 2025.

[38] OpenAI. Thinking with images. `https://openai.com/index/thinking-with-images/`, 2025.

[39] Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

[40] Orion Weller, Benjamin Chang, Sean MacAvaney, Kyle Lo, Arman Cohan, Benjamin Van Durme, Dawn Lawrie, and Luca Soldaini. Followir: Evaluating and teaching information retrieval models to follow instructions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11926–11942, 2025.

[41] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S Siegel, Michael Tang, et al. Bright: A realistic and challenging benchmark for reasoning-intensive retrieval. *arXiv preprint arXiv:2407.12883*, 2024.

[42] Kuicai Dong, Yujing Chang, Derrick Goh Xin Deik, Dexun Li, Ruiming Tang, and Yong Liu. Mmdocir: Benchmarking multimodal retrieval for long documents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30959–30993, 2025.

[43] Yutong Yang, Lifu Huang, Yijie Lin, Xi Peng, and Mouxing Yang. Endowing vision-language models with System 2 thinking for fine-grained visual recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11802–11810, 2026.

[44] Google. Gemini. `https://gemini.google.com/`, 2026. Version: Gemini 3 Pro. Accessed: January 5, 2026.

[45] Yijie Lin, Mouxing Yang, Peng Hu, Jiancheng Lv, Hao Chen, and Xi Peng. Learning with partial and noisy correspondence in graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):8234–8249, 2026.

[46] Haobin Li, Yijie Lin, Peng Hu, Mouxing Yang, and Xi Peng. Learning with dual-level noisy correspondence for multi-modal entity alignment. In *International Conference on Learning Representations*, 2026.

[47] Yijie Lin, Mouxing Yang, Jun Yu, Peng Hu, Changqing Zhang, and Xi Peng. Graph matching with bi-level noisy correspondence. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23305–23314. IEEE, 2023.

[48] ByteDance. Seed-1.6-embedding (doubao-embedding-vision-251215). `https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-embedding-vision`, 2025. Accessed: January 10, 2026.

[49] Zhiqiang Kou, Jing Wang, Jiawei Tang, Yuheng Jia, Boyu Shi, and Xin Geng. Exploiting multi-label correlation in label distribution learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 4326–4334, 8 2024.

[50] Zhiqiang Kou, Si Qin, Hailin Wang, Jing Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations assisted by multi-label learning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, 8 2025.

[51] Zhiqiang Kou, Yucheng Xie, Hailin Wang, Junyang Chen, Jing Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, and Xin Geng. Rankmatch: A novel approach to semi-supervised label distribution learning leveraging rank correlation between labels. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*, San Diego, California, USA, December 2025. Neural Information Processing Systems Foundation.

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

[53] TIGER-Lab. Mmeb-leaderboard. `https://huggingface.co/spaces/TIGER-Lab/MMEB-Leaderboard`. Accessed: 2026-01-29.

[54] Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

[55] Guofeng Ding, Yiding Lu, Peng Hu, Mouxing Yang, Yijie Lin, and Xi Peng. Visual abstraction: A plug-and-play approach for text-visual retrieval. In *Forty-second International Conference on Machine Learning*, 2025.

[56] Haobin Li, Yutong Yang, Yijie Lin, Xiang Dai, Mouxing Yang, and Xi Peng. Reliable thinking with images. *arXiv preprint arXiv:2602.12916*, 2026.

[57] OpenAI. Introducing gpt-5.2. `https://openai.com/index/introducing-gpt-5-2/`, 2025. Accessed: 2026-01-29.

[58] 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*, pages 15338–15347, 2023.

[59] 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*, pages 16495–16504, 2022.

[60] 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 *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[61] Siwei Wu, Yizhi Li, Kang Zhu, Ge Zhang, Yiming Liang, Kaijing Ma, Chenghao Xiao, Haoran Zhang, Bohao Yang, Wenhu Chen, et al. Scimmir: Benchmarking scientific multi-modal information retrieval. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12560–12574, 2024.

[62] 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.

[63] 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.

[64] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE international conference on computer vision*, pages 1463–1471, 2017.

[65] Shuhang Gu, Andreas Lugmayr, Martin Danelljan, Manuel Fritsche, Julien Lamour, and Radu Timofte. Div8k: Diverse 8k resolution image dataset. In *ICCVW*, 2019.

[66] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

[67] Xinmiao Huang, Qisong He, Zhenglin Huang, Boxuan Wang, Zhuoyun Li, Guangliang Cheng, Yi Dong, and Xiaowei Huang. Spatial-dise: A unified benchmark for evaluating spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.13394*, 2025.

[68] 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. *Advances in Neural Information Processing Systems*, 37:126500–126514, 2024.

[69] Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark. *arXiv preprint arXiv:2501.05444*, 2025.

[70] Mehran Kazemi, Nishanth Dikkala, Ankit Anand, Petar Devic, Ishita Dasgupta, Fangyu Liu, Bahare Fatemi, Pranjal Awasthi, Sreenivas Gollapudi, Dee Guo, et al. Remi: A dataset for reasoning with multiple images. *Advances in Neural Information Processing Systems*, 37:60088–60109, 2024.

[71] Tim Sherratt. GLAM-Workbench/trove-journals, 2024.

[72] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *European Conference on Computer Vision*, pages 676–691. Springer, 2018.

[73] arXiv. arxiv.org e-print archive. `https://arxiv.org/`, 2026. Accessed: 2026-01-28.

[74] Farshid Ghezelbash, Amir Hossein Eskandari, and Amir J Bidhendi. A dataset for mechanical mechanisms. *arXiv preprint arXiv:2409.03763*, 2024.

[75] Pragati Shuddhodhan Meshram, Swetha Karthikeyan, Bhavya, and Suma Bhat. Electrovizqa: How well do multi-modal llms perform in electronics visual question answering?, 2024.

[76] Wikimedia commons. `https://commons.wikimedia.org/wiki/Main_Page`, 2026. Accessed: January 5, 2026.

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

[78] World bank open data. `https://data.worldbank.org/`, 2026. Accessed: January 5, 2026.

[79] 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, 2024.

[80] 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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19076–19095, 2025.

[81] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[82] Yung-Sung Chuang, Yang Li, Dong Wang, Ching-Feng Yeh, Kehan Lyu, Ramya Raghavendra, James Glass, Lifei Huang, Jason Weston, Luke Zettlemoyer, et al. Meta clip 2: A worldwide scaling recipe. *arXiv preprint arXiv:2507.22062*, 2025.

[83] 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.

[84] Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*, 2025.

[85] OpenSearch-AI. Ops-MM-Embedding-v1-7B. `https://huggingface.co/OpenSearch-AI/Ops-MM-embedding-v1-7B`, 2025. Hugging Face model.

[86] Jina AI. jina-reranker-m0. `https://huggingface.co/jinaai/jina-reranker-m0`, 2025. Hugging Face model.

[87] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Shaohan Wang, Pengyu Wang, Xiaorui Wang, and Zhendong Mao. A-rag: Scaling agentic retrieval-augmented generation via hierarchical retrieval interfaces. *arXiv preprint arXiv:2602.03442*, 2026.

[88] Wenhan Liu, Xinyu Ma, Yutao Zhu, Yuchen Li, Daiting Shi, Dawei Yin, and Zhicheng Dou. Agentic-r: Learning to retrieve for agentic search. *arXiv preprint arXiv:2601.11888*, 2026.

[89] Md Mahadi Hasan Nahid and Davood Rafiei. Prism: Agentic retrieval with llms for multi-hop question answering. *arXiv preprint arXiv:2510.14278*, 2025.

[90] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024.

APPENDIX

In this appendix, we present:

1. Comparison with Existing Benchmarks (Section A)
2. Dataset Details (Section B)
3. Details of Evaluated Models (Section C)
4. Evaluation with Query Rewriting and Re-ranking (Section D)
5. Results on ARK under Different Evaluation Metrics (Section E)
6. Bootstrap Uncertainty Analysis of Main Results (Section F)
7. Multi-round Retrieval Results for Agent (Section G)
8. Captioning Robustness and Prompt Details (Section H)
9. Impact Statement (Section I)
10. Annotation Protocol and Consistency Analysis for Reasoning-related Labels (Section J)
11. Case Study (Section K)

A Comparison with Existing Benchmarks

Table 4: Comparison of multimodal retrieval benchmarks. Benchmarks are grouped into three parts: (i) *Top*: traditional multimodal retrieval benchmarks dominated by everyday natural images and semantic matching; (ii) *Middle*: knowledge-oriented benchmarks that incorporate web-scale or expert content; (iii) *Bottom*: recent benchmarks that explicitly probe reasoning-intensive retrieval. For benchmarks containing mixed task formats, we report only the number of retrieval-related tasks.

Benchmarks	# Queries	# Tasks	Knowledge-Intensive	Reasoning-Intensive	Multi-Domain	Diverse Reasoning Type	Heterogeneous Visual Type	Fine-grained Visual Reasoning
COCO [7]	25K	1						
Flickr30K [8]	5, 000	1						
CIRR [14]	4,148	1						
FashionIQ [13]	60K	1						
CIRCO [58]	1,020	1						
WebQA [59]	7,540	1	✓					
EDIS [12]	3,241	1	✓					
InfoSeek [60]	1.35M	2	✓					
SciMMIR [61]	530K	11	✓		✓			
M-BEIR [31]	190K	10	✓		✓			
ViDoRe [62]	3,810	2	✓					
MMEB [63]	12K	12	✓		✓			
MRMR [9]	1,435	23	✓	✓	✓	–	–	
MR^2-Bench [10]	1,309	12	✓	✓	✓	–	–	
ARK (Ours)	1,750	17	✓	✓	✓	6	16	✓

As shown in Table 4, multimodal retrieval benchmarks have evolved from *surface-level semantic matching* to *knowledge-intensive retrieval*, and more recently to *reasoning-intensive retrieval*. Early multimodal retrieval benchmarks (top block) are largely built on natural images from daily-life scenarios, where relevance is often well captured by object-level semantics or short textual descriptions. Subsequent benchmarks (middle block) expand the scope to web-scale or expert content, placing stronger emphasis on knowledge-intensive queries and multi-domain coverage, but still provide limited diagnostics of the *reasoning process* required to resolve a query. The most recent benchmarks (bottom block) begin to explicitly incorporate reasoning requirements, yet they typically inherit query styles from knowledge-heavy sources and cover only a narrow slice of reasoning phenomena.

ARK differs from prior benchmarks along several axes reflected in Table 4. First, ARK is the first multimodal retrieval benchmark that explicitly separates knowledge and reasoning as two evaluation

Table 5: Data sources of ARK benchmark. ARK is structured by five knowledge domains and six reasoning skill categories, and covers 16 heterogeneous visual data types. For instances adapted from existing retrieval or QA datasets, we carefully rewrite queries to better elicit the intended reasoning skills, and we curate targeted hard negatives whenever feasible to reduce shortcut matching.

Meta-Task	Subtype	Visual Type	Reasoning Skill	Adapted From
Visual Cognition	Daily-Life	natural images	–	COCO [7] and Fashion200K [64]
	Fine-Grained	natural images	fine-grained visual reasoning	DIV8K [65]
	Spatial-Cogmap	cognitive maps	spatial reasoning	MindCube [66]
	Spatial-3D	3D images	spatial reasoning	Spatial-DISE-12K [67]
Natural Science	Biology	natural images	–	INQUIRE [68]
	Physics	physics phenomenon illustrations	knowledge reasoning	Manually collected
	Chemistry	chemical structure diagrams	knowledge reasoning, symbolic reasoning	EMMA [69]
Formal Science	Mathematics	symbolic mathematical plots	knowledge reasoning, symbolic reasoning	MR^2-Bench [10]
	Code-Drawing	code snippets, design drawings	symbolic reasoning	ReMI [70]
Humanity & Social Science	Economics	tables, line and bar charts	knowledge reasoning, conceptual abstraction	MR^2-Bench [10]
	Comic	comics	knowledge reasoning, conceptual abstraction	Bulletin Editorial Cartoons [71]
	Art	artworks	knowledge reasoning	SemArt [72]
	Calendar	calendars	knowledge reasoning, logical reasoning	Manually collected
	Metro	metro maps	spatial reasoning, fine-grained visual reasoning	Manually collected
Engineering & Technology	Computer-Science	tables, line and bar charts, design drawings	knowledge reasoning, conceptual abstraction	arXiv [73]
	Mechanical	design drawings	knowledge reasoning	Mechanical mechanism [74]
	Electronics	circuit schematics, tables	logical reasoning, symbolic reasoning	ElectroVizQA [75]

axes, enabling clearer diagnosis of model strengths and failures from both perspectives. In contrast, prior benchmarks lack an explicit taxonomy of reasoning skills, making it difficult to attribute errors to missing knowledge versus insufficient reasoning. Second, ARK goes beyond natural images by covering 16 heterogeneous visual formats spanning both natural and structured visuals (*e.g.*, line/bar charts, chemical structure diagrams, artworks, 3D renderings, code, and design drawings) as listed in Table 5. This diversity requires models to ground queries in visual evidence rather than relying on surface-level semantics. Third, ARK provides fine-grained reasoning annotations spanning a diverse set of reasoning types (*i.e.*, knowledge, spatial, logical, symbolic, fine-grained visual reasoning, and conceptual abstraction). In particular, ARK emphasizes richer vision-centric demands, including complex spatial inference and detail-sensitive visual inspection (e.g., identifying subtle objects in high-resolution images), which closely aligns with the recent interest in "thinking with images" [38, 39]. Together, these design choices enable more targeted analysis than prior benchmarks that primarily probe limited forms of visual relations.

B Dataset Details

The overall curation pipeline of ARK has been described in Section 3.3. In this section, we provide construction details for each knowledge subtype, including (i) data sources, (ii) the visual data types of queries and gallery items, (iii) the primary reasoning skills targeted, and (iv) hard-negative construction when applicable. Overall statistics are summarized in Table 5.

B.1 Visual Cognition

Daily-Life. This topic focuses on text-to-image retrieval over everyday visual content and primarily evaluates basic perceptual recognition and visual-text alignment. It serves as a baseline for multimodal retrieval without requiring domain knowledge or multi-step reasoning. Data are constructed from images in COCO [7] and Fashion200K [64], covering daily scenes, object interactions, and apparel combinations. As shown in Fig. 15, queries describe observable attributes such as object categories, colors, spatial arrangements, and clothing details (*e.g.*, style, color, and texture), and retrieval can be resolved through accurate attribute matching. Hard negatives are selected using embedding-based similarity retrieval followed by manual verification, resulting in visually similar but non-matching candidates that require precise perception rather than higher-level reasoning to disambiguate.

Fine-Grained. This topic focuses on retrieval over ultra-high-resolution images and evaluates fine-grained visual reasoning over subtle and localized visual evidence, rather than coarse semantic recognition. We construct this subset based on the DIV8K dataset [65], which provides native ultra-high-resolution images with rich fine-grained details. As shown in Fig. 18, the selected images

have an average resolution of 5960×3900, in contrast to the daily-life subset, where images are typically around 365×461, substantially increasing perceptual complexity and the need for precise inspection of local regions. Queries are natural language descriptions of small or easily overlooked objects and fine-grained visual attributes, often coupled with spatial constraints, requiring models to localize relevant regions and verify detailed conditions rather than relying on global appearance or object-level semantics. This setting closely relates to recent interest in high-resolution visual reasoning and "thinking with images," where correct retrieval depends on deliberate inspection and reasoning over visual details. Hard negatives are manually curated by selecting images that partially satisfy the query (*e.g.*, similar scene layout or object categories) but violate critical fine-grained constraints such as object presence, size, or precise location, ensuring that correct retrieval requires reasoning over local evidence instead of matching salient global features.

Spatial-Cogmap. Inspired by MindCube [66], this subtype targets the spatial reasoning ability of retrieval models, testing whether retrievers can infer global layout, topological relationships, and relative object positions from limited views, rather than relying on object recognition alone. In this subtype, the retrieval task is defined as follows: given a small set of perspective-view images captured from limited viewpoints (e.g., four views sampled along a 360° camera rotation, as shown in Fig. 24), the model must retrieve the corresponding bird's-eye cognitive map from a large gallery. Each cognitive map is rendered on a standardized 10×10 coordinate grid (top-left $[0, 0]$, bottom-right $[9, 9]$), where object locations are marked using a unified set of icons. To prevent models from relying on shallow matching based on object co-occurrence, we construct hard negative samples by randomly swapping the coordinates of adjacent objects while keeping object categories and counts unchanged, thereby perturbing the spatial structure. This design forces models to perform genuine spatial transformation and geometric reasoning from perspective views to top-down maps, effectively evaluating their ability to build internal spatial representations.

Spatial-3D. Beyond spatial reasoning in natural images, we include a synthetic retrieval topic that targets abstract 3D geometric reasoning. We build this subset on Spatial-DISE-12K [67], a Blender-generated dataset with geometric tasks such as 3D rotation, folding, and composition. To adapt it to retrieval, we convert the original multiple-choice questions into retrieval queries. The correct option is treated as the positive target, while the original distractor options are directly reused as hard negatives, preserving their carefully designed geometric confusability. As shown in Fig. 23, in a 3D projection task where the query requires finding the correct planar projection from a specified viewpoint, the model must distinguish it from distractors generated by alternative viewing directions, occlusion patterns, or depth-order permutations. Because these negative samples may appear visually similar while differing in subtle geometric relationships (*e.g.*, relative depth, overlap, and foreshortening), this task compels models to move beyond surface-level appearance matching and perform precise 3D-to-2D spatial transformations and viewpoint-aware geometric reasoning.

B.2 Natural Science

Biology. Fine-grained biological identification is a long-standing challenge in multimodal retrieval, as many species exhibit highly similar visual appearances and differ only in subtle morphological traits. This subtype targets knowledge-intensive retrieval: queries are natural-language descriptions or common names referring to specific species, and candidates are natural images of organisms, where correct retrieval primarily relies on recognizing subtle morphological cues and recalling taxonomic knowledge rather than multi-step reasoning. We sample instances from the INQUIRE dataset [68], which contains images of over 10,000 species annotated by citizen scientists and experts, providing rich visual diversity and expert-level taxonomic labels. Hard negatives are constructed by first using LLMs to propose visually similar species as candidate distractors, and then manually selecting images that closely resemble the positive target in overall appearance or ecological traits but belong to different taxa. As shown in Fig. 9, for the given query "A resting Common Raven on the ground.", common raven and corvus macrorhynchos exhibit nearly identical color distributions and body shapes, yet differ in subtle morphological details. Such distractors ensure that successful retrieval depends on precise biological knowledge.

Physics. This topic targets both knowledge-intensive and reasoning-intensive retrieval in physics, evaluating whether models can retrieve the correct visual evidence by combining domain knowledge with diagram-based understanding. We curate undergraduate-level and advanced physics concepts and phenomenon illustrations spanning classical mechanics, electromagnetism, thermodynamics,

and modern physics. Queries are natural-language descriptions that often reference mechanisms, conditions, or observable outcomes, requiring models to interpret schematic diagrams or phenomenon figures and reason about the underlying physical process rather than relying on surface-level semantics. Hard negatives are constructed whenever feasible by first using an LLM to propose closely related physical concepts that are easily confusable with the target, and then collecting corresponding visual illustrations from Wikimedia Commons [76]. We restrict our collection to Commons assets with explicit reuse permissions and preserve attribution. For the example in Fig. 22, when the ground-truth corresponds to the effective potential contours and Lagrange points, hard negatives include topologically similar geometries. These candidates share similar nuclear contexts and visually similar diagrammatic patterns, yet differ in the underlying physical mechanism, ensuring that correct retrieval depends on physics-grounded reasoning and precise concept discrimination rather than superficial similarity.

Chemistry. This topic targets knowledge reasoning and symbolic reasoning in chemistry, where correct retrieval requires understanding structured molecular representations and chemistry-specific transformation principles. We source reaction simulation problems from EMMA [69], which are curated by domain experts who annotate reaction products and design questions with professional rigor, covering diverse reaction families and mechanistic patterns. Queries describe a reaction setup using chemical structure diagrams and symbolic descriptions, and candidates contain corresponding molecular structures or reaction outcomes. Hard negatives are constructed whenever feasible by manually selecting alternative reaction outcomes that are chemically plausible and structurally similar to the ground-truth products, but do not strictly satisfy the specific requirements implied by the query, as shown in Fig. 11. These distractors are designed to force models to rely on chemistry-specific reasoning over molecular structures instead of shortcut matching.

B.3 Formal Science

Mathematics. Mathematics is a core topic in formal science. In ARK, this subtype is designed to evaluate knowledge reasoning and symbolic reasoning. Following MR^2-Bench [10, 77], theorems and relations are expressed via geometric constructions and symbolic diagrams. As shown in Fig. 19, answering these queries requires models to combine domain knowledge (*e.g.*, mathematical definitions, theorems, and constraints) with symbolic manipulation, and to ground the reasoning on visual evidence such as geometric figures. To increase retrieval difficulty, we construct hard negatives by manually selecting diagrams or symbolic expressions that are highly similar to the query in appearance or surface form, but conflict with key mathematical constraints.

Code-Drawing. This subtype targets symbolic reasoning over code-conditioned visual generation, requiring models to ground structured program symbols into the corresponding visual structures. We use the TikZ code–image pairs from ReMI [70], where queries are raw TikZ snippets and gallery items are the rendered vector graphics produced by compiling the code as shown in Fig. 12. Correct retrieval therefore depends on understanding how syntactic primitives and parameters (*e.g.*, coordinates, commands, and constraints) deterministically specify geometric layouts. Hard negatives are constructed via a single-line perturbation strategy. Specifically, we minimally edit a critical line in the original TikZ code and compile the modified program to obtain a distractor image that remains visually similar but changes key geometric relations (*e.g.*, relative positions, connectivity, or angles). Such negatives enforce precise code-to-image grounding and discourage shortcut matching based on global appearance.

B.4 Humanities and Social Science

Economics. This subtype targets knowledge reasoning and conceptual abstraction over economic data visualizations, evaluating whether models can retrieve the correct chart by understanding both economic concepts and the abstract information conveyed by graphical trends. This subset follows the data sourcing of MR^2-Bench [10]: we use the curated economics reports and charts derived from World Bank Open Data [78], covering economic phenomena across countries and time periods. Queries are natural-language descriptions that reference economic concepts, comparative relations, and temporal or predictive conditions, requiring models to interpret chart semantics rather than match surface text or visual patterns. Hard negatives are constructed by selecting charts that share similar economic keywords or themes, but convey different economic mechanisms or implications than those required by the query. For example, as shown in Fig. 16, for the query "Assuming a host has a duty

to admit people escaping danger, but discretion over opportunity-driven entrants, which diagram best supports concluding within a generic two-by-two that crosses push vs. pull motives with the strength of the skills-demand alignment that the near-term net payoff is governed chiefly by the match, with an explicit side-by-side contrast of net gains versus net losses along the match gradient?", hard negatives may include diagrams that emphasize push–pull forces without an explicit skill–demand match dimension or charts that depict aggregate trends rather than a match-based contrast of gains versus losses.

Comic. This subtype targets knowledge reasoning and conceptual abstraction on historical cartoons, where meaning is conveyed implicitly through symbolism and metaphor rather than explicit visual descriptions. Accurate retrieval requires recognizing historically grounded cues and mapping concrete visual elements (*e.g.*, characters, icons, and allegories) to higher-level concepts and viewpoints, as illustrated in Fig. 13. We construct this topic based on the Bulletin Editorial Cartoons dataset [71], which covers cartoons published on the front pages of The Bulletin[3] between 1886 and 1952. We select 3,468 cartoons to form the gallery and manually annotate 87 instances to create the query set. Given the difficulty of constructing hard negative samples for this topic, we instead avoid shortcut learning by constraining the query format. Each query consists of only two components: a statement of the historical event and a revelation of the underlying viewpoint. This design deliberately avoids direct visual descriptions of the images, requiring models to genuinely understand the metaphors embedded in the cartoons rather than relying on superficial visual feature matching.

Art. This subtype targets knowledge reasoning in art, where queries provide indirect art-historical cues rather than explicit entity names, and correct retrieval requires inferring the most likely artist, style, or subject matter from visual and textual evidence. We use SemArt [72], which pairs paintings with expert-written descriptions and structured metadata (*e.g.*, artist and genre), enabling queries grounded in art history. As shown in Fig. 8, queries describe stylistic signatures, creative techniques, iconography, or thematic conventions. Successful retrieval therefore requires reasoning with art knowledge to connect these cues to the appropriate artwork, rather than relying on direct name matching or generic aesthetic similarity. Hard negatives are constructed by first retrieving style-neighbor candidates in the embedding space and then manually selecting artworks that are visually close in overall style but violate key knowledge constraints implied by the query (*e.g.*, period, technique, or iconographic elements). These distractors force models to perform knowledge-grounded inference over artworks rather than exploiting superficial stylistic resemblance.

Calendar. This subtype targets knowledge reasoning and logical reasoning over artifacts from the traditional Chinese calendar system. It evaluates cross-modal understanding and reasoning over traditional Chinese calendar artifacts, including the lunar calendar system and the 24 solar terms. We curate this subset by manually selecting representative calendar images from publicly available online resources with reuse permission, focusing on solar-term annotations, lunar–Gregorian correspondences, and culturally grounded visual elements. Queries require the model to analyze the text and visual information in calendar images (*e.g.*, solar term names, lunar/Gregorian date positions, zodiac animal information, etc.) and incorporate reasoning based on knowledge, such as lunar calendar rules and the sequence of solar terms. As shown in Fig. 10, a query like "A calendar month page from which one can determine that people born in the lunar second month can celebrate their lunar birthday twice." requires the model to understand that having two birthdays is possible only when there is a leap month, specifically a leap second lunar month, and thus infer that the calendar image in question corresponds to a year with a leap second lunar month.

Metro. This subtype targets spatial reasoning and fine-grained visual reasoning in metro diagrams. Unlike knowledge-intensive retrieval, the goal is to test whether models can find the correct map or station context by reasoning about connectivity and topology, rather than relying on domain knowledge or surface-level keyword matching. We manually curate high-resolution subway route maps from publicly available online resources with reuse permission. The maps have an average resolution of 9858×9057, substantially increasing the demand for precise visual inspection and spatial reasoning. As shown in Fig. 21, queries are natural-language descriptions that refer to route connectivity, transfer relations, and relative station positions, and candidates are the corresponding metro maps.

[3]The Bulletin (Australian periodical)

B.5 Engineering and Technology

Computer-Science. This subtype targets knowledge reasoning and conceptual abstraction over figures from computer science literature, testing whether retrievers can interpret technical visuals and recover the intended high-level takeaway instead of relying on superficial keyword overlap. We curate this subset from publicly available arXiv papers and complementary online resources with reuse permission, covering common CS figure types such as algorithm schematics, data-structure diagrams, and experimental plots (*e.g.*, scaling trends and complexity-related curves). As shown in Fig. 14, queries are natural-language statements describing methodological claims or figure-implied conclusions, and candidates are the corresponding paper figures. Correct retrieval requires grounding the query in structured visual evidence (*e.g.*, axes, legends, symbols, and trend patterns) and using CS knowledge to infer the conclusion supported by the figure.

Mechanical. This subtype is primarily knowledge-intensive, with a small portion requiring knowledge reasoning to disambiguate visually similar mechanisms. It evaluates cross-modal grounding between mechanical engineering terminology and technical imagery as shown in Fig. 20, where models must retrieve the correct part or mechanism given its name or functional description. We construct this subset from a mechanical mechanism dataset [74] containing $\sim$8K images with structured descriptions, covering common components and mechanisms (*e.g.*, gears, bearings, linkages, and fasteners). Hard negatives are obtained by retrieving the most similar candidates in the embedding space and then manually filtering to retain distractors that are category-consistent yet differ in key geometric or functional attributes.

Electronics. This subtype targets logical reasoning and symbolic reasoning in digital circuit understanding, evaluating retrieval over structured artifacts such as circuit diagrams, truth tables, and Karnaugh maps. We build this subset on ElectroVizQA [75], which covers both combinational and sequential logic with paired visuals and Boolean-form descriptions. As shown in Fig. 17, queries require models to interpret gate symbols and circuit structure, or to infer Boolean relations from truth tables, and then retrieve the matching visual evidence. Hard negatives are constructed by preserving highly similar visual layouts while perturbing the underlying logic (*e.g.*, swapping gate types, modifying truth-table entries, or changing Boolean operators), yielding distractors that look plausible but violate the intended logical function.

C Details of Evaluated Models

We summarize the evaluated embedding-based retrievers, re-ranking models, and agentic retrievers in Tables 6, 7, and 8, respectively, covering 25 representative retriever configurations in total. These configurations span text-only and multimodal retrieval paradigms, a wide range of parameter scales, and different training strategies, including models with and without explicit reasoning-oriented objectives. This breadth enables a comprehensive assessment of how modality, architecture, model capacity, and retrieval paradigm affect knowledge-intensive and reasoning-intensive multimodal retrieval.

All locally executed inference experiments were conducted on NVIDIA RTX 4090 GPUs. For each evaluated model, we use the officially released checkpoint and default inference configuration. For API-based models, computation is performed through the corresponding hosted service provider.

Table 6: Overview of Evaluated Embedding Models.

MODALITY	MODEL NAME	PARAM.	REASONING	Model Link
TEXT	BGE-M3 [79]	600M		HF: BAAI/bge-m3
	Qwen3-Embedding [54]	4B		HF: Qwen/Qwen3-Embedding-4B
	Qwen3-Embedding [54]	8B		HF: Qwen/Qwen3-Embedding-8B
	Diver-Embed [34]	1.7B	✓	HF: AQ-MedAI/Diver-Retriever-1.7B
	Diver-Embed [34]	4B	✓	HF: AQ-MedAI/Diver-Retriever-4B
	ReasonIR [35]	8B	✓	HF: reasonir/ReasonIR-8B
MULTI MODAL	CLIP [22]	428M		HF: openai/clip-vit-large-patch14-336
	BGE-VL [80]	400M		HF: BAAI/BGE-VL-large
	SigLIP2 [81]	0.9B		HF: google/siglip2-so400m-patch14-384
	MetaCLIP2 [82]	2B		HF: facebook/metaclip-2-worldwide-huge-quickgelu
	EVA-CLIP [23]	18B		HF: BAAI/EVA-CLIP-18B
	VLM2VecV2 [83]	2B		HF: VLM2Vec/VLM2Vec-V2.0
	LamRA-Ret [30]	7B		HF: code-kunkun/LamRA-Ret
	RzenEmbed [84]	7B		HF: qihoo360/RzenEmbed
	Ops-MM-Embedding [85]	2B		HF: OpenSearch-AI/Ops-MM-Embedding-v1-2B
	Ops-MM-Embedding [85]	7B		HF: OpenSearch-AI/Ops-MM-Embedding-v1-7B
	Qwen3-VL-embedding [32]	2B		HF: Qwen/Qwen3-VL-Embedding-2B
	Qwen3-VL-embedding [32]	8B		HF: Qwen/Qwen3-VL-Embedding-8B
	Seed1.6-embedding	-		API: Doubao-Embedding-Vision

Table 7: Overview of Evaluated Reranker Models.

MODALITY	MODEL NAME	PARAM.	Model Link
TEXT	BGE-Reranker [79]	600M	HF: BAAI/bge-reranker-v2-m3
MULTI-MODAL	Qwen3-VL-Reranker [32]	8B	HF: Qwen/Qwen3-VL-Reranker-8B
	Jina-Reranker [86]	2B	HF: jinaai/jina-reranker-m0

Table 8: Overview of Evaluated Agentic Models. A-RAG combines GPT with Qwen3 embedding models. Since GPT is shared across settings, we distinguish the two A-RAG variants by the parameter size of the underlying Qwen3 retriever.

MODALITY	MODEL NAME	PARAM.	Model Link
TEXT	A-RAG [87]	0.6B	GPT-5-mini + Qwen3-Embedding-0.6B
	A-RAG [87]	8B	GPT-5-mini + Qwen3-Embedding-8B
	Agentic-R [88]	7B	HF: liuwenhan/search-agent

D Evaluation with Query Rewriting and Re-ranking

Due to space constraints in the main paper, we report additional results for *query rewriting* and *re-ranking* in Tables 9 and 10.

Table 9: Performance of retrieval models with query rewriting and re-ranking on ARK. R@k and N@k denote Recall@k and nDCG@k, respectively. Scores are macro-averaged across subtypes under the knowledge-axis grouping and across reasoning skills under the reasoning-axis grouping.

Model	**Knowledge perspective**						**Reasoning perspective**					
	R@1	R@5	R@10	N@5	N@10	N@20	R@1	R@5	R@10	N@5	N@10	N@20
Query Rewrite												
Seed1.6-embedding	19.32	37.60	48.43	28.79	32.28	34.77	15.30	31.69	40.70	23.80	26.71	29.18
+ *Rewrite*	23.42	43.85	54.49	34.04	37.52	39.87	17.49	35.09	44.12	26.59	29.53	31.96
Reranker Models Based on Seed1.6-embedding												
BGE-Reranker	12.23	23.95	32.47	18.29	21.05	24.11	8.19	18.35	26.01	13.40	15.88	18.51
+ *Rewrite*	15.25	27.29	35.97	21.50	24.29	27.49	11.64	21.39	28.98	16.72	19.15	21.70
Jina-Reranker	23.90	43.06	52.88	33.95	37.16	39.46	17.14	33.48	42.91	25.74	28.81	30.91
+ *Rewrite*	27.77	47.47	57.69	38.17	41.50	43.32	21.21	38.27	47.32	30.20	33.14	34.76
Qwen3-VL-Reranker	29.33	49.20	58.25	39.92	42.86	44.46	23.89	41.76	49.56	33.39	35.90	37.28
+ *Rewrite*	**34.51**	**51.45**	**59.38**	**43.66**	**46.23**	**47.88**	**28.31**	**42.76**	**49.58**	**36.12**	**38.29**	**39.86**

Table 10: Recall@1 on ARK by reasoning skill with query rewriting and re-ranking.

Model	**Knowledge Reasoning**	**Conceptual Abstraction**	**Fine-Grained Visual Reasoning**	**Logical Reasoning**	**Spatial Reasoning**	**Symbolic Reasoning**	**Average**
Query Rewrite							
Seed1.6-embedding	22.08	33.33	4.35	10.71	6.19	15.16	15.30
+ *Rewrite*	24.96	31.00	10.00	18.40	5.00	15.58	17.49
Reranker Models Based on Seed1.6-embedding							
BGE-Reranker	14.55	21.40	3.85	3.07	1.54	4.76	8.19
+ *Rewrite*	16.98	21.40	5.38	17.18	3.08	5.84	11.64
Jina-Reranker	27.82	**41.48**	10.00	4.91	6.92	11.69	17.14
+ *Rewrite*	29.24	38.86	14.62	21.47	7.69	15.37	21.21
Qwen3-VL-Reranker	33.95	37.12	**30.00**	5.52	**10.77**	25.97	23.89
+ *Rewrite*	**38.09**	39.74	26.15	**23.31**	9.23	**33.33**	**28.31**

E Results on ARK under different evaluation metrics.

In addition to the main Recall@1 results reported in Tables 2 and 3, we provide a broader view of retrieval quality under multiple metrics. Tables 11–20 report Recall@5/10 and nDCG@5/10/20. Recall@k measures whether the ground-truth item appears within the top-k retrieved candidates, while nDCG@k further accounts for its rank position, assigning higher credit when the correct item is retrieved closer to the top.

Table 11: Domain-wise Recall@5 performance on ARK (knowledge axis). Columns correspond to ARK's five knowledge domains and their subtypes. "Avg." denotes the macro-average across all subtypes.

Model	Visual Cognition				Natural Science			Formal Science		Humanities & Social Science					Engineering & Technology			Avg.
	Daily.	Fine.	SCog.	S3D.	Bio.	Phy.	Che.	Mat.	Cod.	Eco.	Comic.	Art	Cal.	Met.	Compu.	Mecha.	Ele.	
Text Embedding Models																		
BGE-M3-0.6B	57.63	6.52	0.00	1.53	63.86	30.00	8.57	10.53	4.00	56.25	3.45	9.00	7.50	31.25	31.25	10.98	0.99	19.61
Diver-Embed-1.7B	48.73	7.61	0.00	0.76	57.83	46.25	9.52	26.97	7.00	50.00	2.30	7.00	10.00	30.00	52.50	14.63	1.98	21.95
Diver-Embed-4B	61.02	6.52	0.00	0.76	68.67	40.00	15.24	31.58	15.00	75.00	5.75	11.00	7.50	38.75	60.00	15.85	4.95	26.92
Qwen3-Embed-4B	57.20	7.61	4.94	2.29	71.08	41.25	14.29	28.95	13.00	63.75	3.45	9.00	7.50	23.75	51.25	20.73	6.93	25.12
Qwen3-Embed-8B	61.86	10.87	3.70	2.29	73.49	50.00	17.14	30.92	14.00	**77.50**	6.90	7.00	16.25	27.50	58.75	21.95	9.90	28.82
ReasonIR-8B	59.75	8.70	0.00	1.53	62.65	38.75	15.24	21.71	6.00	73.75	10.34	5.00	7.50	41.25	58.75	17.07	3.96	25.41
Multimodal Embedding Models																		
CLIP-0.4B	61.44	8.70	3.70	10.69	61.45	20.00	11.43	5.92	2.00	27.50	9.20	29.00	10.00	23.75	16.25	17.07	11.88	19.41
BGE-VL-0.4B	72.46	9.78	0.00	0.00	67.47	8.75	0.00	7.24	8.00	10.00	1.15	13.00	3.75	13.75	16.25	12.20	9.90	14.92
SigLIP2-0.9B	**88.98**	6.52	2.47	4.58	66.27	15.00	15.24	11.84	1.00	31.25	9.20	15.00	11.25	5.00	35.00	26.83	5.94	20.67
MetaCLIP2-2B	78.39	**13.04**	4.94	6.87	78.31	38.75	8.57	9.21	8.00	42.50	4.60	**41.00**	7.50	12.50	28.75	25.61	10.89	24.67
VLM2VecV2-2B	65.25	8.70	13.58	6.11	63.86	27.50	23.81	28.95	10.00	58.75	3.45	10.00	7.50	30.00	37.50	20.73	5.94	24.80
Ops-MM-Embed-v1-2B	83.05	8.70	6.17	7.63	74.70	32.50	**40.95**	25.66	12.00	48.75	13.79	18.00	6.25	31.25	50.00	18.29	8.91	28.62
Qwen3-VL-Embed-2B	85.17	10.87	14.81	12.21	73.49	36.25	26.67	32.89	9.00	70.00	13.79	9.00	3.75	45.00	47.50	20.73	**16.83**	31.06
LamRA-Ret-7B	78.39	8.70	1.23	3.82	78.31	13.75	5.71	25.00	24.00	62.50	9.20	17.00	3.75	30.00	6.25	14.63	3.96	22.72
RzenEmbed-7B	84.75	10.87	20.99	4.58	71.08	45.00	35.24	33.55	30.00	67.50	14.94	19.00	12.50	43.75	**72.50**	23.17	**16.83**	35.66
Ops-MM-Embed-v1-7B	83.05	**13.04**	**25.93**	**13.74**	75.90	43.75	**40.95**	35.53	20.00	68.75	17.24	23.00	10.00	43.75	60.00	21.95	**16.83**	36.08
Qwen3-VL-Embed-8B	86.86	9.78	6.17	**13.74**	78.31	**51.25**	39.05	30.92	**45.00**	**77.50**	19.54	12.00	11.25	43.75	56.25	26.83	**16.83**	36.77
EVA-CLIP-18B	83.05	10.87	0.00	12.21	**86.75**	35.00	10.48	1.32	9.00	52.50	6.90	38.00	6.25	15.00	23.75	**29.27**	5.94	25.08
Seed1.6-embedding	84.75	**13.04**	7.41	8.40	66.27	45.00	**40.95**	**38.16**	36.00	75.00	**21.84**	26.00	**27.50**	**47.50**	71.25	23.17	6.93	**37.60**

Table 12: Domain-wise Recall@10 performance on ARK (knowledge axis). Columns correspond to ARK's five knowledge domains and their subtypes.

Model	Visual Cognition				Natural Science			Formal Science		Humanities & Social Science					Engineering & Technology			Avg.
	Daily.	Fine.	SCog.	S3D.	Bio.	Phy.	Che.	Mat.	Cod.	Eco.	Comic.	Art	Cal.	Met.	Compu.	Mecha.	Ele.	
Text Embedding Models																		
BGE-M3-0.6B	63.98	11.96	1.23	1.53	68.67	37.50	17.14	15.13	11.00	67.50	6.90	13.00	7.50	45.00	43.75	14.63	4.95	25.37
Diver-Embed-1.7B	56.78	9.78	0.00	1.53	72.29	50.00	19.05	31.58	14.00	57.50	2.30	12.00	12.50	50.00	58.75	19.51	7.92	27.97
Diver-Embed-4B	68.64	13.04	0.00	0.76	78.31	48.75	23.81	38.82	23.00	82.50	6.90	17.00	10.00	53.75	67.50	19.51	6.93	32.90
Qwen3-Embed-4B	65.68	15.22	16.05	3.82	79.52	48.75	19.05	37.50	19.00	76.25	5.75	13.00	8.75	50.00	58.75	23.17	14.85	32.65
Qwen3-Embed-8B	70.34	15.22	3.70	6.11	83.13	53.75	27.62	41.45	21.00	**85.00**	10.34	14.00	21.25	50.00	65.00	25.61	13.86	35.73
ReasonIR-8B	69.49	17.39	3.70	3.05	73.49	45.00	21.90	32.24	9.00	82.50	13.79	7.00	10.00	61.25	66.25	23.17	6.93	32.13
Multimodal Embedding Models																		
CLIP-0.4B	69.49	18.48	7.41	12.98	85.54	30.00	14.29	10.53	6.00	42.50	12.64	40.00	13.75	35.00	21.25	20.73	14.85	26.79
BGE-VL-0.4B	81.78	14.13	0.00	0.00	85.54	12.50	0.00	13.82	10.00	22.50	1.15	18.00	6.25	33.75	18.75	15.85	10.89	20.29
SigLIP2-0.9B	**92.80**	7.61	7.41	5.34	75.90	23.75	20.95	17.11	1.00	37.50	10.34	17.00	23.75	16.25	42.50	31.71	7.92	25.81
MetaCLIP2-2B	86.02	18.48	7.41	10.69	**96.39**	51.25	11.43	13.82	13.00	50.00	5.75	**55.00**	8.75	31.25	32.50	35.37	12.87	31.76
VLM2VecV2-2B	74.15	15.22	35.80	9.92	74.70	33.75	36.19	42.76	19.00	65.00	5.75	14.00	10.00	48.75	41.25	23.17	10.89	32.96
Ops-MM-Embed-v1-2B	90.68	16.30	23.46	12.21	90.36	45.00	55.24	36.84	20.00	58.75	16.09	25.00	6.25	56.25	57.50	23.17	17.82	38.29
Qwen3-VL-Embed-2B	89.83	16.30	30.86	14.50	83.13	48.75	42.86	40.79	17.00	76.25	19.54	17.00	5.00	63.75	50.00	24.39	**24.75**	39.10
LamRA-Ret-7B	88.14	14.13	12.35	7.63	91.57	27.50	8.57	31.58	37.00	75.00	13.79	27.00	3.75	45.00	7.50	15.85	3.96	30.02
RzenEmbed-7B	**92.80**	15.22	44.44	6.11	87.95	52.50	**63.81**	42.76	42.00	82.50	17.24	23.00	18.75	56.25	76.25	29.27	19.80	45.33
Ops-MM-Embed-v1-7B	89.83	**22.83**	**56.79**	19.08	92.77	51.25	**63.81**	45.39	30.00	75.00	19.54	32.00	13.75	**67.50**	62.50	28.05	16.83	46.29
Qwen3-VL-Embed-8B	91.95	14.13	34.57	**19.85**	92.77	**60.00**	58.10	44.74	55.00	82.50	21.84	20.00	15.00	58.75	61.25	31.71	16.83	45.82
EVA-CLIP-18B	88.56	21.74	2.47	16.03	93.98	46.25	14.29	5.26	11.00	61.25	8.05	50.00	6.25	26.25	33.75	**37.80**	11.88	31.46
Seed1.6-embedding	91.10	18.48	30.86	12.98	83.13	56.25	60.95	**46.71**	**58.00**	81.25	**27.59**	30.00	**38.75**	61.25	**82.50**	31.71	11.88	**48.43**

Table 13: Domain-wise NDCG@5 performance on ARK (knowledge axis). Columns correspond to ARK's five knowledge domains and their subtypes.

Model	Visual Cognition				Natural Science			Formal Science		Humanities & Social Science					Engineering & Technology			Avg.
	Daily.	Fine.	SCog.	S3D.	Bio.	Phy.	Che.	Mat.	Cod.	Eco.	Comic.	Art	Cal.	Met.	Compu.	Mecha.	Ele.	
Text Embedding Models																		
BGE-M3-0.6B	48.34	4.27	0.00	0.76	48.88	23.44	5.13	7.10	2.89	47.21	2.22	5.62	6.73	19.37	28.37	10.23	0.99	15.39
Diver-Embed-1.7B	42.55	4.62	0.00	0.30	41.85	36.70	6.37	19.97	4.27	37.19	1.64	5.43	9.54	17.94	45.96	10.33	0.88	16.80
Diver-Embed-4B	54.27	3.70	0.00	0.76	55.07	32.82	8.98	23.41	8.35	64.75	4.09	8.25	7.04	24.23	53.74	13.81	2.48	21.51
Qwen3-Embed-4B	50.07	5.20	4.32	1.63	50.08	32.21	8.31	20.23	8.10	51.32	2.60	6.31	6.16	13.77	42.66	16.43	4.97	19.08
Qwen3-Embed-8B	54.28	6.13	1.68	1.73	56.86	37.89	10.65	22.22	8.94	62.72	4.69	5.56	11.69	14.92	53.09	15.52	6.12	22.04
ReasonIR-8B	50.58	4.83	0.00	0.76	45.79	29.48	8.17	16.84	3.56	60.16	7.10	2.88	5.34	27.97	49.51	13.59	2.36	19.35
Multimodal Embedding Models																		
CLIP-0.4B	49.96	3.97	2.24	6.59	45.22	13.93	7.88	3.83	1.50	17.63	6.56	20.22	7.52	12.90	12.14	12.47	8.13	13.69
BGE-VL-0.4B	62.79	4.51	0.00	0.00	50.66	4.39	0.00	4.16	3.87	8.04	0.57	9.59	2.66	7.30	13.45	10.24	6.09	11.08
SigLIP2-0.9B	**79.40**	4.41	0.96	2.88	50.33	9.05	8.36	7.93	0.43	22.66	7.46	8.71	8.83	3.29	28.01	21.09	3.43	15.72
MetaCLIP2-2B	69.59	7.75	2.33	5.36	56.65	29.14	4.95	4.98	5.34	31.81	4.17	**31.26**	4.62	7.97	21.12	19.92	6.54	18.44
VLM2VecV2-2B	54.85	4.87	5.91	3.51	43.95	16.67	13.69	18.53	6.97	46.34	2.79	7.19	5.02	19.77	30.94	14.27	3.12	17.55
Ops-MM-Embed-v1-2B	73.93	4.42	3.59	5.00	57.18	22.02	21.73	15.73	7.39	36.53	10.99	12.19	5.16	22.86	38.74	14.82	5.75	21.06
Qwen3-VL-Embed-2B	76.09	5.82	6.62	8.76	50.68	29.89	13.72	20.43	6.47	56.29	9.26	6.01	3.12	25.73	37.81	14.53	10.13	22.43
LamRA-Ret-7B	69.57	6.87	0.53	1.95	58.92	7.57	4.47	17.63	14.50	50.39	6.13	12.35	3.12	17.71	5.79	11.45	2.67	17.15
RzenEmbed-7B	77.79	6.56	10.29	3.20	52.16	28.09	20.79	22.94	18.30	56.80	9.16	12.03	9.17	25.77	60.63	19.45	11.57	26.16
Ops-MM-Embed-v1-7B	73.37	6.66	**14.83**	**9.53**	60.12	30.13	22.34	22.84	12.76	55.37	12.05	17.90	7.58	30.96	52.47	17.40	11.15	26.91
Qwen3-VL-Embed-8B	78.35	5.19	2.83	8.95	59.87	**40.18**	20.87	20.89	**29.38**	**65.62**	15.73	8.54	7.44	25.26	51.54	19.93	**12.11**	27.80
EVA-CLIP-18B	74.28	6.60	0.00	7.61	**67.24**	24.70	8.65	0.91	5.84	36.89	3.97	28.31	5.33	9.89	20.65	**25.87**	5.01	19.51
Seed1.6-embedding	75.56	**8.86**	3.31	5.86	49.10	33.69	**22.43**	**25.96**	27.12	61.55	**17.45**	17.97	**20.44**	**32.13**	**64.57**	17.91	5.51	**28.79**

Table 14: Domain-wise NDCG@10 performance on ARK (knowledge axis). Columns correspond to ARK's five knowledge domains and their subtypes.

Model	Visual Cognition				Natural Science			Formal Science		Humanities & Social Science					Engineering & Technology			Avg.
	Daily.	Fine.	SCog.	S3D.	Bio.	Phy.	Che.	Mat.	Cod.	Eco.	Comic.	Art	Cal.	Met.	Compu.	Mecha.	Ele.	
Text Embedding Models																		
BGE-M3-0.6B	50.39	6.01	0.39	0.76	50.49	25.81	7.93	8.63	5.14	50.90	3.36	6.82	6.73	23.71	32.54	11.38	2.31	17.25
Diver-Embed-1.7B	45.15	5.25	0.00	0.55	46.45	37.91	9.47	21.46	6.48	39.67	1.64	7.10	10.35	24.25	48.01	11.89	2.85	18.73
Diver-Embed-4B	56.74	5.78	0.00	0.76	58.24	35.64	11.72	25.83	10.93	**67.15**	4.50	10.15	7.90	29.10	56.26	14.96	3.13	23.46
Qwen3-Embed-4B	52.81	7.54	7.82	2.12	52.82	34.58	9.87	22.94	10.01	55.59	3.26	7.55	6.52	22.21	45.21	17.18	7.51	21.50
Qwen3-Embed-8B	57.04	7.57	1.68	2.93	59.96	39.15	13.93	25.67	11.21	65.25	5.89	7.89	13.33	21.97	55.11	16.69	7.34	24.27
ReasonIR-8B	53.78	7.56	1.12	1.20	49.29	31.62	10.33	20.26	4.57	62.91	8.21	3.54	6.12	34.33	51.99	15.43	3.34	21.51
Multimodal Embedding Models																		
CLIP-0.4B	52.55	6.98	3.43	7.32	53.06	17.03	8.73	5.26	2.82	22.40	7.70	23.57	8.76	16.46	13.70	13.59	9.12	16.03
BGE-VL-0.4B	65.77	5.99	0.00	0.00	56.80	5.59	0.00	6.28	4.46	12.26	0.57	11.27	3.47	13.64	14.27	11.35	6.37	12.83
SigLIP2-0.9B	**80.65**	4.74	2.46	3.15	53.44	11.80	10.25	9.65	0.43	24.66	7.79	9.36	12.87	6.94	30.47	22.63	4.09	17.38
MetaCLIP2-2B	72.03	9.51	3.06	6.59	62.59	33.16	5.81	6.44	6.86	34.24	4.58	**35.79**	4.99	13.76	22.33	23.12	7.15	20.71
VLM2VecV2-2B	57.69	7.00	13.28	4.67	47.37	18.77	17.73	22.92	9.97	48.35	3.56	8.54	5.80	25.80	32.14	15.06	4.80	20.20
Ops-MM-Embed-v1-2B	76.43	6.83	9.20	6.44	62.20	26.11	26.32	19.39	9.99	39.68	11.67	14.36	5.16	30.75	41.13	16.43	8.52	24.15
Qwen3-VL-Embed-2B	77.62	7.58	11.75	9.49	53.89	33.81	18.99	23.13	9.07	58.39	10.98	8.68	3.57	31.77	38.58	15.70	**12.72**	25.04
LamRA-Ret-7B	72.74	8.66	3.97	3.26	63.35	11.91	5.41	19.70	18.61	54.46	7.63	15.45	3.12	22.42	6.23	11.83	2.67	19.50
RzenEmbed-7B	80.43	7.98	17.86	3.71	57.88	30.53	**30.09**	26.02	22.16	61.79	9.88	13.36	11.19	29.86	61.75	21.37	12.57	29.32
Ops-MM-Embed-v1-7B	75.54	9.85	**24.66**	**11.24**	65.70	32.52	29.72	26.04	15.95	57.37	12.87	20.80	8.71	**38.87**	53.31	19.36	11.15	30.22
Qwen3-VL-Embed-8B	80.00	6.63	11.91	10.97	64.67	**43.16**	27.19	25.46	32.69	**67.15**	16.47	11.19	8.72	30.27	53.17	21.53	12.11	30.78
EVA-CLIP-18B	76.10	10.27	0.80	8.81	**69.65**	28.30	9.90	2.24	6.48	39.87	4.38	32.02	5.33	13.34	23.90	**28.65**	6.80	21.58
Seed1.6-embedding	77.69	**10.58**	10.97	7.42	54.66	37.25	28.97	**28.69**	**34.21**	63.52	**19.32**	19.18	**23.90**	36.41	**68.36**	20.60	7.06	**32.28**

Table 15: Domain-wise NDCG@20 performance on ARK (knowledge axis). Columns correspond to ARK's five knowledge domains and their subtypes.

Model	Visual Cognition				Natural Science			Formal Science		Humanities & Social Science					Engineering & Technology			Avg.
	Daily.	Fine.	SCog.	S3D.	Bio.	Phy.	Che.	Mat.	Cod.	Eco.	Comic.	Art	Cal.	Met.	Compu.	Mecha.	Ele.	
Text Embedding Models																		
BGE-M3-0.6B	52.68	7.61	1.92	1.16	52.64	27.62	10.59	11.11	6.14	54.08	4.25	7.76	6.73	29.55	33.71	13.54	2.54	19.04
Diver-Embed-1.7B	47.83	6.64	0.60	1.13	48.88	39.80	11.89	23.79	8.49	42.31	1.95	8.58	10.94	29.00	49.27	12.49	4.07	20.45
Diver-Embed-4B	58.72	7.38	0.32	0.95	60.07	37.29	12.92	28.04	12.67	**68.75**	4.50	11.65	10.77	33.72	56.93	15.92	4.15	24.99
Qwen3-Embed-4B	54.43	7.84	9.41	2.89	54.90	35.86	11.78	23.98	11.27	57.47	3.57	8.84	7.10	28.58	46.74	18.12	8.99	23.05
Qwen3-Embed-8B	59.31	9.23	3.20	3.11	62.11	40.62	15.88	28.20	12.71	65.56	6.53	9.90	14.57	26.91	56.39	17.02	9.61	25.93
ReasonIR-8B	55.72	9.36	2.99	1.58	51.44	32.86	11.47	21.90	6.32	63.60	9.38	5.53	6.79	37.60	52.95	16.62	5.01	23.01
Multimodal Embedding Models																		
CLIP-0.4B	54.77	7.83	5.29	8.47	55.46	19.73	10.64	8.28	4.29	25.37	9.39	26.09	10.92	23.84	14.65	16.09	9.39	18.26
BGE-VL-0.4B	67.17	7.60	0.00	0.18	59.10	6.56	0.00	7.94	5.67	13.79	1.15	14.10	4.40	22.21	14.27	13.17	7.45	14.40
SigLIP2-0.9B	**81.93**	5.59	5.64	3.56	54.08	13.66	12.46	11.30	0.43	26.55	8.65	11.10	14.14	8.44	32.05	24.13	4.33	18.71
MetaCLIP2-2B	73.14	**12.89**	4.81	7.35	63.50	34.44	6.97	9.31	8.20	36.78	5.20	**37.83**	4.99	21.45	24.59	24.99	7.62	22.59
VLM2VecV2-2B	59.54	7.81	16.73	6.43	49.23	20.65	19.88	25.05	13.53	50.56	4.48	10.31	6.74	29.88	33.67	16.00	5.07	22.09
Ops-MM-Embed-v1-2B	77.19	7.63	15.44	8.25	63.78	27.44	29.64	20.59	12.49	41.65	12.52	17.53	6.74	39.23	41.79	18.32	9.31	26.44
Qwen3-VL-Embed-2B	78.84	8.64	18.74	11.23	56.70	35.05	22.15	25.60	11.00	59.06	12.46	10.16	3.87	36.01	40.44	16.88	**13.51**	27.08
LamRA-Ret-7B	74.52	11.09	6.73	3.82	64.32	14.50	7.29	21.50	21.62	55.78	9.65	17.18	4.15	27.59	8.07	12.40	2.93	21.36
RzenEmbed-7B	81.39	10.72	23.55	5.24	59.99	32.43	**32.79**	28.50	26.79	63.07	11.32	15.17	12.74	35.03	62.69	23.86	12.84	31.65
Ops-MM-Embed-v1-7B	76.94	10.99	**28.14**	**12.41**	66.97	34.45	31.69	27.87	20.74	59.88	13.78	22.79	10.26	**43.11**	56.17	20.85	11.66	32.28
Qwen3-VL-Embed-8B	81.17	9.30	13.81	11.74	65.90	**43.75**	30.95	26.98	37.35	67.82	18.77	14.75	10.98	36.58	55.67	22.74	12.11	32.96
EVA-CLIP-18B	77.06	11.92	1.08	10.39	**70.88**	28.91	12.27	4.02	7.20	41.81	5.83	34.04	6.55	20.62	25.17	**30.43**	7.79	23.29
Seed1.6-embedding	78.55	12.84	16.74	7.82	57.77	38.28	31.43	**30.50**	**38.50**	65.07	**20.27**	22.70	**25.76**	40.95	**69.62**	24.30	10.01	**34.77**

Table 16: Recall@5 on ARK by cognitive demand and reasoning skill (reasoning axis). "Both" denotes instances that are both knowledge-intensive and reasoning-intensive.

Model	Cognitive Demand				Reasoning Skill						Average
	Perception-Only	Knowledge-Intensive	Reasoning-Intensive	Both	Knowledge Reasoning	Conceptual Abstraction	Fine-Grained Visual Reasoning	Logical Reasoning	Spatial Reasoning	Symbolic Reasoning	
Text Embedding Models											
BGE-M3-0.6B	57.63	24.44	19.07	22.00	23.25	31.88	12.31	14.11	11.92	14.07	17.92
Diver-Embed-1.7B	49.15	28.74	21.82	26.84	28.67	37.55	13.85	15.34	8.46	21.21	20.85
Diver-Embed-4B	61.02	32.68	24.73	30.91	33.38	43.67	13.85	17.79	8.46	24.68	23.64
Qwen3-Embed-4B	57.63	25.96	18.38	22.66	26.68	35.81	7.69	7.36	8.46	16.23	17.04
Qwen3-Embed-8B	61.44	30.17	22.05	27.28	31.81	44.54	12.31	13.50	8.46	18.83	21.57
ReasonIR-8B	59.75	29.36	22.43	27.72	29.53	44.98	14.62	11.04	8.46	20.56	21.53
Multimodal Embedding Models											
CLIP-0.4B	61.44	17.46	13.02	13.75	15.69	17.47	11.54	10.43	11.54	7.58	12.37
BGE-VL-0.4B	72.46	12.53	8.04	9.24	9.27	8.73	12.31	7.36	1.92	6.28	7.65
SigLIP2-0.9B	**88.98**	18.98	13.09	16.39	20.40	19.65	6.92	7.98	4.62	9.31	11.48
MetaCLIP2-2B	78.39	23.46	16.16	18.59	21.68	24.02	13.85	9.20	8.46	8.87	14.35
VLM2VecV2-2B	64.83	23.99	19.60	22.22	27.10	31.44	13.85	7.98	13.08	18.40	18.64
Ops-MM-Embed-v1-2B	83.05	28.38	21.82	25.85	30.53	32.31	14.62	8.59	11.54	22.08	19.95
Qwen3-VL-Embed-2B	85.59	29.19	24.27	26.62	30.81	39.74	20.77	12.27	18.08	22.29	23.99
LamRA-Ret-7B	79.24	22.20	16.77	19.91	21.83	29.26	13.85	4.91	8.08	16.02	15.66
RzenEmbed-7B	84.75	36.08	29.56	34.76	38.52	45.41	**23.85**	15.95	14.23	29.65	27.94
Ops-MM-Embed-v1-7B	83.05	35.99	30.93	34.54	39.66	45.41	23.08	15.34	**22.31**	30.30	29.35
Qwen3-VL-Embed-8B	86.44	38.05	30.40	36.19	38.09	47.16	20.77	15.34	15.77	33.55	28.45
EVA-CLIP-18B	83.05	23.28	15.39	17.49	20.68	26.64	12.31	6.13	9.23	6.28	13.55
Seed1.6-embedding	84.32	**45.41**	**35.35**	**44.31**	**45.88**	**55.38**	14.13	**19.64**	14.16	**40.96**	**31.69**

Table 17: Recall@10 on ARK by cognitive demand and reasoning skill (reasoning axis). "Both" denotes instances that are both knowledge-intensive and reasoning-intensive.

Model	Cognitive Demand				Reasoning Skill						Average
	Perception-Only	Knowledge-Intensive	Reasoning-Intensive	Both	Knowledge Reasoning	Conceptual Abstraction	Fine-Grained Visual Reasoning	Logical Reasoning	Spatial Reasoning	Symbolic Reasoning	
Text Embedding Models											
BGE-M3-0.6B	63.98	29.81	23.74	27.17	27.67	37.12	17.69	17.18	14.62	20.56	22.47
Diver-Embed-1.7B	56.78	34.65	26.80	31.90	33.38	40.61	21.54	19.02	11.54	27.06	25.53
Diver-Embed-4B	69.92	39.03	30.70	37.29	39.94	49.34	23.08	23.31	11.54	30.52	29.62
Qwen3-Embed-4B	65.25	32.59	26.34	29.81	33.95	43.23	20.77	12.88	17.31	23.81	25.32
Qwen3-Embed-8B	70.34	37.42	29.56	35.20	40.37	50.22	23.85	18.40	13.08	27.49	28.90
ReasonIR-8B	69.49	35.90	29.48	34.32	36.38	48.91	27.69	17.18	13.46	28.35	28.66
Multimodal Embedding Models											
CLIP-0.4B	69.07	24.62	19.30	19.91	22.82	24.45	22.31	13.50	15.77	11.47	18.39
BGE-VL-0.4B	81.78	17.19	12.25	12.98	13.69	13.54	20.00	8.59	6.15	9.31	11.88
SigLIP2-0.9B	92.80	24.17	17.92	21.34	26.53	24.89	14.62	14.11	7.31	13.20	16.78
MetaCLIP2-2B	86.02	30.62	22.13	24.42	28.25	30.13	22.31	10.43	14.23	13.42	19.80
VLM2VecV2-2B	74.58	31.33	28.64	30.25	35.52	35.81	26.15	12.27	24.23	28.57	27.09
Ops-MM-Embed-v1-2B	90.68	37.51	31.39	34.32	39.23	38.43	29.23	14.11	22.69	33.12	29.47
Qwen3-VL-Embed-2B	90.68	37.15	32.77	34.76	38.94	44.54	27.69	18.40	28.85	33.12	31.92
LamRA-Ret-7B	88.56	28.65	23.43	26.07	27.96	35.81	21.54	4.91	16.15	21.65	21.34
RzenEmbed-7B	**93.22**	45.84	39.20	44.77	49.22	52.84	29.23	21.47	25.77	42.42	36.82
Ops-MM-Embed-v1-7B	90.25	44.05	40.74	42.35	48.36	49.78	**34.62**	15.95	**39.23**	39.83	37.96
Qwen3-VL-Embed-8B	91.53	46.37	39.74	44.22	47.08	51.53	25.38	17.18	31.92	43.94	36.17
EVA-CLIP-18B	88.56	29.27	20.83	22.55	26.11	32.75	23.08	9.82	13.85	10.17	19.30
Seed1.6-embedding	89.83	**56.30**	**45.51**	**55.03**	**55.83**	**61.29**	18.48	**30.36**	24.78	**53.46**	**40.70**

Table 18: NDCG@5 on ARK by cognitive demand and reasoning skill (reasoning axis). "Both" denotes instances that are both knowledge-intensive and reasoning-intensive.

Model	Cognitive Demand				Reasoning Skill						Average
	Perception-Only	Knowledge-Intensive	Reasoning-Intensive	Both	Knowledge Reasoning	Conceptual Abstraction	Fine-Grained Visual Reasoning	Logical Reasoning	Spatial Reasoning	Symbolic Reasoning	
Text Embedding Models											
BGE-M3-0.6B	48.34	18.58	13.65	16.35	17.73	25.26	7.34	9.35	7.26	9.27	12.70
Diver-Embed-1.7B	42.77	22.21	16.18	20.44	22.13	30.47	8.15	10.41	5.75	15.01	15.32
Diver-Embed-4B	54.21	25.68	18.65	23.84	26.17	36.12	8.97	12.15	5.07	17.20	17.61
Qwen3-Embed-4B	49.92	19.48	13.40	16.73	19.90	27.62	4.98	5.84	5.60	11.08	12.50
Qwen3-Embed-8B	53.86	23.16	15.99	20.44	24.20	34.66	6.41	9.15	5.09	12.98	15.42
ReasonIR-8B	50.58	22.52	16.97	21.28	23.25	37.02	9.20	7.22	6.37	14.19	16.21
Multimodal Embedding Models											
CLIP-0.4B	50.07	12.30	8.58	9.46	10.81	11.71	5.74	7.26	7.03	5.15	7.95
BGE-VL-0.4B	63.14	8.94	5.31	6.46	6.80	6.36	6.34	4.77	1.06	3.63	4.83
SigLIP2-0.9B	**79.42**	13.63	9.23	11.69	14.63	14.15	4.86	5.58	2.66	5.75	7.94
MetaCLIP2-2B	69.78	16.98	11.37	13.37	15.74	17.92	8.44	5.63	5.83	5.36	9.82
VLM2VecV2-2B	55.10	16.62	13.22	15.43	19.01	24.28	8.77	4.66	7.40	11.40	12.59
Ops-MM-Embed-v1-2B	74.19	19.98	14.99	17.86	21.28	24.10	9.10	6.06	8.03	13.16	13.62
Qwen3-VL-Embed-2B	76.35	20.90	16.63	19.17	22.34	32.50	11.99	8.12	10.48	13.81	16.54
LamRA-Ret-7B	69.65	16.08	11.89	14.43	16.26	22.18	9.22	3.73	4.62	10.75	11.13
RzenEmbed-7B	77.94	25.81	20.25	24.57	27.65	34.81	14.49	11.43	8.06	19.04	19.25
Ops-MM-Embed-v1-7B	73.97	26.32	21.65	24.76	28.69	35.02	**14.63**	10.71	**14.44**	18.68	20.36
Qwen3-VL-Embed-8B	78.00	28.24	21.49	26.51	28.37	39.53	10.93	11.30	9.77	21.38	20.21
EVA-CLIP-18B	74.28	17.75	11.21	13.13	15.49	18.82	7.66	4.96	6.01	4.80	9.62
Seed1.6-embedding	74.21	**34.29**	**26.52**	**33.54**	**34.60**	**45.18**	9.36	**14.99**	10.37	**28.33**	**23.80**

Table 19: NDCG@10 on ARK by cognitive demand and reasoning skill (reasoning axis). "Both" denotes instances that are both knowledge-intensive and reasoning-intensive.

Model	Cognitive Demand				Reasoning Skill						Average
	Perception-Only	Knowledge-Intensive	Reasoning-Intensive	Both	Knowledge Reasoning	Conceptual Abstraction	Fine-Grained Visual Reasoning	Logical Reasoning	Spatial Reasoning	Symbolic Reasoning	
Text Embedding Models											
BGE-M3-0.6B	50.39	20.33	15.18	18.04	19.19	26.99	9.04	10.37	8.17	11.41	14.20
Diver-Embed-1.7B	45.23	24.12	17.78	22.07	23.64	31.43	10.53	11.62	6.77	16.90	16.81
Diver-Embed-4B	57.07	27.73	20.60	25.90	28.29	37.94	12.02	13.92	6.09	19.12	19.56
Qwen3-Embed-4B	52.47	21.63	15.98	19.06	22.28	30.10	9.11	7.59	8.45	13.52	15.17
Qwen3-Embed-8B	56.74	25.52	18.41	23.01	27.00	36.55	10.05	10.70	6.55	15.77	17.77
ReasonIR-8B	53.78	24.66	19.26	23.44	25.49	38.36	13.34	9.20	8.02	16.72	18.52
Multimodal Embedding Models											
CLIP-0.4B	52.54	14.56	10.58	11.42	13.06	13.97	9.05	8.29	8.39	6.38	9.86
BGE-VL-0.4B	66.15	10.47	6.67	7.67	8.25	7.98	8.79	5.14	2.42	4.58	6.19
SigLIP2-0.9B	80.67	15.31	10.79	13.29	16.61	15.83	7.35	7.60	3.50	7.02	9.65
MetaCLIP2-2B	72.22	19.29	13.27	15.25	17.87	19.89	10.98	6.00	7.66	6.77	11.53
VLM2VecV2-2B	58.27	18.99	16.15	18.03	21.71	25.71	12.71	6.08	11.04	14.70	15.33
Ops-MM-Embed-v1-2B	76.64	22.91	18.06	20.58	24.08	26.00	13.71	7.77	11.65	16.73	16.66
Qwen3-VL-Embed-2B	77.98	23.52	19.41	21.86	25.04	34.09	14.17	10.16	13.96	17.39	19.13
LamRA-Ret-7B	72.71	18.12	14.01	16.38	18.20	24.28	11.72	3.73	7.18	12.52	12.94
RzenEmbed-7B	**80.71**	29.02	23.41	27.86	31.18	37.25	16.25	13.26	11.81	23.25	22.17
Ops-MM-Embed-v1-7B	76.29	28.94	24.83	27.30	31.53	36.45	**18.46**	10.88	**19.87**	21.75	23.16
Qwen3-VL-Embed-8B	79.65	30.98	24.55	29.15	31.32	40.94	12.48	11.93	15.01	24.79	22.75
EVA-CLIP-18B	76.10	19.68	12.94	14.74	17.22	20.82	11.20	6.07	7.44	6.03	11.46
Seed1.6-embedding	76.01	**37.85**	**29.84**	**37.04**	**37.87**	**47.03**	10.73	**18.30**	13.88	**32.44**	**26.71**

Table 20: NDCG@20 on ARK by cognitive demand and reasoning skill (reasoning axis). "Both" denotes instances that are both knowledge-intensive and reasoning-intensive.

Model	Cognitive Demand				Reasoning Skill						Average
	Perception-Only	Knowledge-Intensive	Reasoning-Intensive	Both	Knowledge Reasoning	Conceptual Abstraction	Fine-Grained Visual Reasoning	Logical Reasoning	Spatial Reasoning	Symbolic Reasoning	
Text Embedding Models											
BGE-M3-0.6B	52.68	22.31	17.11	19.98	21.26	28.72	11.54	12.22	9.52	13.67	16.16
Diver-Embed-1.7B	47.88	25.87	19.48	23.89	25.48	32.76	12.32	13.47	7.85	19.05	18.49
Diver-Embed-4B	58.53	29.15	22.03	27.36	29.81	38.41	14.16	14.87	7.09	20.95	20.88
Qwen3-Embed-4B	54.32	22.99	17.51	20.39	23.61	31.19	11.39	8.79	10.36	15.02	16.73
Qwen3-Embed-8B	59.11	27.11	20.04	24.56	28.46	37.21	12.48	12.57	8.02	17.93	19.44
ReasonIR-8B	55.73	26.16	20.62	24.84	26.98	39.46	15.22	10.28	9.03	18.08	19.84
Multimodal Embedding Models											
CLIP-0.4B	54.88	16.59	12.70	13.37	15.28	16.06	11.85	9.22	10.63	8.14	11.86
BGE-VL-0.4B	67.55	11.76	8.12	8.82	9.45	8.87	13.10	6.27	3.74	5.58	7.83
SigLIP2-0.9B	**81.85**	16.62	12.03	14.54	18.16	17.38	8.12	8.20	4.99	8.12	10.83
MetaCLIP2-2B	73.34	20.82	15.16	16.76	19.57	21.32	15.82	6.29	9.63	8.20	13.47
VLM2VecV2-2B	59.89	20.73	18.04	19.75	23.40	26.93	14.63	6.56	13.69	16.93	17.02
Ops-MM-Embed-v1-2B	77.42	24.73	20.48	22.43	25.97	27.48	16.58	8.56	16.12	18.73	18.91
Qwen3-VL-Embed-2B	79.08	25.40	21.66	23.81	27.21	35.97	16.01	10.65	17.59	19.60	21.17
LamRA-Ret-7B	74.20	19.68	15.85	17.96	19.82	25.97	16.26	4.06	8.61	14.31	14.84
RzenEmbed-7B	81.56	31.07	25.76	29.77	32.95	38.45	19.42	13.88	15.45	25.75	24.32
Ops-MM-Embed-v1-7B	77.47	30.97	26.85	29.20	33.20	38.01	**20.66**	11.65	**22.26**	24.03	24.97
Qwen3-VL-Embed-8B	80.94	33.00	26.78	31.21	33.34	42.49	16.91	12.73	16.76	27.10	24.89
EVA-CLIP-18B	77.06	21.16	14.72	16.22	18.87	22.39	14.33	6.83	9.68	7.53	13.27
Seed1.6-embedding	77.29	**40.25**	**32.17**	**39.33**	**39.64**	**48.52**	13.74	**21.63**	16.27	**35.28**	**29.18**

F Bootstrap Uncertainty Analysis of Main Results

To quantify uncertainty in the main retrieval results, we estimate 95% confidence intervals for Recall@1 using 1,000 query-level bootstrap resamples. Tables 21 and 22 report the point estimates and percentile-based bootstrap intervals for eight representative text and multimodal embedding models on the knowledge and reasoning axes, respectively. The intervals are reported as R@1 [95% CI], in percentage points.

Knowledge-axis reliability. The bootstrap results support the domain- and subtype-level findings underlying Observation 2. Across the evaluated models, performance remains substantially stronger on Daily-Life than on Fine-Grained, Spatial-Cogmap, and Spatial-3D, and is generally higher on Biology, Economics, and Computer Science than on Chemistry. Although confidence intervals are wider for some smaller subtypes, the large and consistently observed subtype-level gaps remain clear. Our conclusion therefore concerns recurring domain- and subtype-level variation across models.

Reasoning-axis reliability. The reasoning-axis intervals reinforce the main diagnostic conclusions. For every evaluated model, Perception-Only retrieval is markedly stronger than Reasoning-Intensive retrieval. Across reasoning skills, Knowledge Reasoning and Conceptual Abstraction consistently outperform Fine-Grained Visual Reasoning and Spatial Reasoning, whose estimates remain low even after accounting for bootstrap uncertainty. This recurring pattern supports the conclusion that vision-centric reasoning is a persistent bottleneck rather than an artifact of a single model. These marginal bootstrap intervals quantify sampling variability within each group; they should not be interpreted as formal paired significance tests between models.

Table 21: Recall@1 with 95% bootstrap confidence intervals on the knowledge axis. Each entry is reported as R@1 [lower, upper] in percentage points.

Subtype	Text Embedding Models				Multimodal Embedding Models			
	Diver-Embed-4B	**Qwen3-Embed-4B**	**Qwen3-Embed-8B**	**ReasonIR-8B**	**RzenEmbed-7B**	**Ops-MM-Embed-v1-7B**	**Qwen3-VL-Embed-8B**	**Seed1.6-embedding**
Daily-Life	46.19 [39.83, 52.54]	42.37 [36.02, 48.73]	46.19 [39.83, 52.97]	40.25 [33.90, 46.61]	69.92 [64.41, 75.00]	62.29 [56.36, 68.22]	69.07 [63.56, 74.58]	65.25 [59.32, 71.61]
Fine-Grained	1.09 [0.00, 3.26]	2.17 [0.00, 5.43]	1.09 [0.00, 4.35]	1.09 [0.00, 3.26]	2.17 [0.00, 6.52]	1.09 [0.00, 3.26]	0.00 [0.00, 0.00]	4.35 [1.09, 8.70]
Spatial-Cogmap	0.00 [0.00, 0.00]	3.70 [0.00, 8.64]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	1.23 [0.00, 3.70]	1.23 [0.00, 3.70]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
Spatial-3D	0.76 [0.00, 2.29]	0.76 [0.00, 2.29]	0.76 [0.00, 3.05]	0.00 [0.00, 0.00]	1.53 [0.00, 3.82]	5.34 [1.53, 9.16]	3.82 [0.76, 7.63]	3.82 [0.76, 7.63]
Biology	39.76 [28.92, 50.60]	27.71 [18.07, 37.35]	40.96 [31.33, 51.81]	28.92 [18.07, 37.35]	32.53 [22.89, 42.17]	42.17 [31.33, 53.01]	39.76 [28.92, 50.60]	31.33 [21.69, 42.17]
Physics	22.50 [13.75, 31.25]	21.25 [12.50, 31.25]	22.50 [13.75, 32.50]	20.00 [12.47, 28.75]	10.00 [3.75, 17.50]	15.00 [7.50, 21.25]	26.25 [16.25, 36.25]	18.75 [11.25, 27.50]
Chemistry	2.86 [0.00, 5.71]	1.90 [0.00, 4.76]	4.76 [0.95, 9.52]	1.90 [0.00, 4.76]	5.71 [1.90, 10.48]	3.81 [0.95, 7.62]	3.81 [0.95, 7.62]	3.81 [0.95, 7.62]
Mathematics	14.47 [8.55, 20.39]	10.53 [5.92, 15.79]	12.50 [7.24, 18.42]	10.53 [5.92, 15.79]	11.84 [7.22, 17.76]	9.21 [4.61, 13.82]	9.87 [5.92, 15.13]	11.84 [6.58, 17.11]
Code-Drawing	2.00 [0.00, 5.00]	3.00 [0.00, 7.00]	4.00 [1.00, 8.00]	1.00 [0.00, 3.00]	7.00 [2.98, 12.00]	6.00 [2.00, 11.00]	14.00 [7.00, 21.00]	18.00 [11.00, 26.00]
Economics	55.00 [43.75, 65.00]	36.25 [26.25, 47.50]	47.50 [36.25, 60.00]	42.50 [32.50, 53.75]	43.75 [32.50, 55.00]	41.25 [30.00, 52.50]	51.25 [40.00, 62.50]	45.00 [33.75, 56.25]
Comic	2.30 [0.00, 5.75]	1.15 [0.00, 3.45]	2.30 [0.00, 5.75]	4.60 [1.15, 9.20]	2.30 [0.00, 5.75]	6.90 [2.30, 12.64]	11.49 [5.75, 18.39]	12.64 [5.75, 20.69]
Art	5.00 [1.00, 10.00]	4.00 [1.00, 8.00]	4.00 [1.00, 8.00]	1.00 [0.00, 3.00]	4.00 [1.00, 8.00]	13.00 [7.00, 20.00]	4.00 [1.00, 8.00]	10.00 [4.00, 17.00]
Calendar	6.25 [1.25, 12.50]	5.00 [1.25, 10.00]	7.50 [2.50, 13.75]	3.75 [0.00, 8.75]	6.25 [1.25, 12.50]	5.00 [1.25, 11.25]	3.75 [0.00, 8.75]	13.75 [7.50, 21.25]
Metro	10.00 [4.97, 17.50]	3.75 [0.00, 8.75]	2.50 [0.00, 6.25]	13.75 [6.25, 21.28]	7.50 [2.50, 13.75]	17.50 [9.97, 26.25]	7.50 [2.50, 13.75]	16.25 [8.75, 25.00]
Computer-Science	46.25 [36.25, 57.50]	33.75 [23.75, 43.75]	46.25 [36.25, 57.53]	36.25 [25.00, 46.25]	48.75 [37.50, 58.75]	42.50 [32.50, 53.75]	46.25 [35.00, 57.50]	57.50 [46.25, 67.50]
Mechanical	10.98 [4.88, 18.29]	10.98 [4.88, 18.29]	8.54 [3.66, 14.63]	9.76 [3.66, 15.85]	15.85 [8.54, 23.17]	12.20 [4.88, 19.51]	13.41 [6.10, 20.73]	12.20 [6.10, 19.51]
Electronics	0.00 [0.00, 0.00]	2.97 [0.00, 5.94]	1.98 [0.00, 4.95]	0.99 [0.00, 2.97]	4.95 [0.99, 9.90]	5.94 [1.98, 10.89]	6.93 [1.98, 11.88]	3.96 [0.99, 7.92]
Macro avg.	**15.61 [14.09, 17.10]**	**12.43 [10.96, 13.84]**	**14.90 [13.48, 16.45]**	**12.72 [11.13, 14.18]**	**16.19 [14.79, 17.55]**	**17.08 [15.50, 18.72]**	**18.30 [16.66, 19.94]**	**19.32 [17.52, 21.00]**

Table 22: Recall@1 with 95% bootstrap confidence intervals on the reasoning axis. Each entry is reported as R@1 [lower, upper] in percentage points.

Group	Text Embedding Models				Multimodal Embedding Models			
	Diver-Embed-4B	**Qwen3-Embed-4B**	**Qwen3-Embed-8B**	**ReasonIR-8B**	**RzenEmbed-7B**	**Ops-MM-Embed-v1-7B**	**Qwen3-VL-Embed-8B**	**Seed1.6-embedding**
Perception-Only	46.19 [40.08, 52.89]	41.53 [35.29, 47.93]	45.76 [39.20, 51.99]	40.25 [34.24, 47.01]	70.34 [64.38, 75.89]	63.56 [57.56, 69.49]	68.64 [62.60, 74.03]	62.29 [56.05, 68.02]
Knowledge-Intensive	17.73 [15.71, 19.98]	12.09 [10.27, 14.05]	15.67 [13.69, 17.91]	15.31 [13.36, 17.36]	15.04 [13.02, 17.33]	16.11 [14.07, 18.27]	17.73 [15.72, 20.02]	22.33 [20.05, 24.69]
Reasoning-Intensive	11.94 [10.38, 13.72]	7.73 [6.33, 9.21]	9.57 [8.00, 11.20]	11.18 [9.56, 12.85]	10.57 [8.91, 12.29]	12.02 [10.14, 13.80]	11.94 [10.32, 13.59]	16.85 [14.85, 18.90]
Both	15.95 [13.77, 18.35]	9.90 [8.11, 11.86]	13.09 [11.03, 15.37]	14.41 [12.27, 16.59]	13.75 [11.56, 16.06]	14.63 [12.27, 16.96]	15.95 [13.69, 18.29]	21.70 [19.08, 24.40]
Knowledge Reasoning	18.12 [15.54, 21.08]	11.98 [9.72, 14.48]	15.98 [13.49, 18.68]	16.41 [13.74, 19.03]	16.12 [13.56, 18.88]	17.12 [14.31, 20.03]	17.55 [14.91, 20.52]	22.08 [19.17, 25.29]
Conceptual Abstraction	27.07 [21.86, 32.87]	17.90 [13.15, 23.08]	23.58 [18.40, 29.32]	28.38 [22.77, 33.93]	22.71 [17.41, 28.76]	23.58 [17.70, 29.24]	30.13 [24.68, 36.41]	33.33 [27.62, 38.97]
Fine-Grained Visual Reasoning	3.85 [0.79, 7.14]	1.54 [0.00, 4.03]	0.77 [0.00, 2.52]	3.85 [0.81, 7.09]	4.62 [1.50, 8.57]	6.92 [2.88, 11.30]	0.77 [0.00, 2.56]	4.35 [0.70, 8.67]
Logical Reasoning	6.13 [2.65, 10.00]	4.29 [1.31, 7.74]	4.91 [1.95, 8.19]	3.07 [0.62, 5.85]	6.13 [2.86, 10.13]	6.13 [2.56, 10.00]	6.75 [3.31, 11.11]	10.71 [7.05, 14.84]
Spatial Reasoning	1.54 [0.37, 3.14]	2.69 [0.81, 4.89]	1.54 [0.37, 3.32]	4.23 [1.79, 6.85]	2.31 [0.75, 4.33]	5.77 [2.98, 8.66]	3.85 [1.76, 6.09]	6.19 [3.60, 9.24]
Symbolic Reasoning	9.31 [6.72, 12.12]	5.19 [3.23, 7.43]	6.71 [4.40, 9.21]	7.58 [5.26, 9.96]	8.01 [5.39, 10.55]	6.93 [4.64, 9.23]	8.87 [6.35, 11.51]	15.16 [12.19, 18.54]
Average	**11.00 [9.36, 12.67]**	**7.26 [5.89, 8.80]**	**8.92 [7.41, 10.54]**	**10.59 [8.89, 12.26]**	**9.98 [8.28, 11.80]**	**11.08 [9.23, 12.94]**	**11.32 [9.71, 13.01]**	**15.30 [13.38, 17.29]**

G Multi-round Retrieval Results for Agent

Recent agentic methods extend conventional one-shot retrieval by decomposing a user query into iterative search steps, inspecting intermediate results, and reformulating subsequent retrieval actions based on accumulated evidence [89, 87, 88]. This paradigm is particularly relevant to ARK, since many queries require multi-step reasoning and visual verification rather than simple semantic matching.

In this section, we evaluate two representative multi-round agent methods, A-RAG [87] and Agentic-R [88]. As shown in Table 23, these agentic retrievers do not outperform strong multimodal embedding models in per-round Recall@1, indicating that their single-step retrieval accuracy remains limited. However, their cumulative Hits@1 improves consistently as the number of retrieval rounds increases, showing that iterative search can recover additional correct targets over multiple attempts. These results suggest that current agentic retrievers are more promising as components in multi-stage retrieval pipelines than as direct replacements for strong multimodal retrievers.

Table 23: Multi-round retrieval performance. Text and multimodal embedding models are included as single-round baselines (reported with R@1 only). Agentic retrievers support multi-round retrieval, and we report both per-round R@1 and cumulative Hits@1 across rounds. Here, R@1 measures success in the current round only, while Hits@1 accumulates successful retrievals over all rounds. Agentic retrievers show consistent gains as the number of retrieval rounds increases.

Model	**Knowledge domain avg.**		**Reasoning skill avg.**	
	R@1	**Hits@1**	**R@1**	**Hits@1**
Text Embedding Models				
BGE-M3-0.6B	10.71	-	7.27	-
Diver-Embed-1.7B	11.38	-	9.16	-
Diver-Embed-4B	**15.61**	-	**11.00**	-
Qwen3-Embed-4B	12.43	-	7.26	-
Qwen3-Embed-8B	**14.90**	-	8.92	-
ReasonIR-8B	12.72	-	**10.59**	-
Multimodal Embedding Models				
CLIP-0.4B	7.66	-	3.52	-
BGE-VL-0.4B	7.12	-	2.15	-
SigLIP2-0.9B	**10.68**	-	**4.33**	-
LamRA-Ret-7B	11.26	-	6.39	-
RzenEmbed-7B	16.19	-	9.98	-
Ops-MM-Embed-v1-7B	17.08	-	11.08	-
Qwen3-VL-Embed-8B	**18.30**	-	**11.32**	-
Agentic Retrieval Models				
A-RAG-0.6B (round 1)	12.64	12.64	8.37	8.37
A-RAG-0.6B (round 3)	13.53	21.42	9.18	14.65
A-RAG-0.6B (round 5)	**14.39**	**27.66**	**9.74**	**20.48**
A-RAG-8B (round 1)	13.09	13.09	9.96	9.96
A-RAG-8B (round 3)	14.11	22.08	11.12	17.34
A-RAG-8B (round 5)	**14.87**	**28.42**	**11.90**	**22.83**
Agentic-R-7B (round 1)	12.81	12.81	8.73	8.73
Agentic-R-7B (round 3)	13.56	22.71	9.36	18.33
Agentic-R-7B (round 5)	**14.11**	**26.52**	**9.67**	**21.79**

H Captioning Robustness and Prompt Details

We first evaluate the robustness of text-only retrieval baselines to different image-captioning models. We then provide the prompts used for caption generation and query rewriting for inference-time enhancement in Tables 26 and 27, respectively.

H.1 Robustness to Captioning Models

To examine whether using Qwen-generated image captions systematically favors Qwen-family text embedding models, we reran all six text-only embedding models on the full benchmark using captions generated by two non-Qwen models: Kimi-VL-A3B-Instruct (16B parameters, 3B active) and GLM-4.6V-Flash (9B parameters).

Table 24 reports the macro-average Recall@1 on the knowledge axis. Model rankings remain highly stable across captioners: Diver-Embed-4B performs best under all three settings, and the Spearman rank correlations are $\rho = 1.00$ between Qwen and Kimi and $\rho = 0.94$ between Qwen and GLM. Moreover, Qwen3-Embed-4B performs better with GLM captions than with Qwen captions (13.30 vs. 12.43).

Table 24: Macro-average Recall@1 (%) on the knowledge axis using captions generated by different models. Differences are measured relative to Qwen captions.

Text embedding model	Qwen	Kimi	GLM	Kimi−Qwen	GLM−Qwen
BGE-M3-0.6B	10.71	8.97	10.08	−1.74	−0.63
Diver-Embed-1.7B	11.38	9.70	10.40	−1.68	−0.98
Diver-Embed-4B	**15.61**	**12.97**	**14.77**	−2.64	−0.84
Qwen3-Embed-4B	12.43	11.05	13.30	−1.38	+0.87
Qwen3-Embed-8B	14.90	11.95	13.67	−2.95	−1.23
ReasonIR-8B	12.72	11.09	12.37	−1.63	−0.35
Model average	12.96	10.96	12.43	−2.00	−0.53

To directly assess family-specific compatibility, we define the Qwen-caption gain for each embedding model as its performance with Qwen captions minus its average performance with the two non-Qwen captioners. On the knowledge axis, the mean gain is 1.17 percentage points for the two Qwen embedding models, compared with 1.31 points for the four non-Qwen models.

Table 25 reports the macro-average Recall@1 across the six reasoning skills. These fine-grained results are more sensitive to caption quality: the rank correlations decrease to approximately $\rho = 0.60$ for Qwen versus Kimi and $\rho = 0.49$ for Qwen versus GLM. Nevertheless, Diver-Embed-4B remains the strongest model with all three captioners. The Qwen-caption gain is 0.16 points for Qwen embedding models and 2.30 points for non-Qwen embedding models, indicating that the relative benefit of Qwen captions is not concentrated among Qwen-family retrievers.

Table 25: Macro-average Recall@1 (%) across the six reasoning skills using captions generated by different models.

Text embedding model	Qwen	Kimi	GLM
BGE-M3-0.6B	7.27	5.42	6.12
Diver-Embed-1.7B	9.16	5.29	5.64
Diver-Embed-4B	**11.00**	**9.29**	**9.54**
Qwen3-Embed-4B	7.26	7.18	8.35
Qwen3-Embed-8B	8.92	7.63	8.58
ReasonIR-8B	10.59	7.93	8.44
Model average	9.03	7.12	7.78

Overall, these results provide no evidence of a systematic Qwen-family compatibility advantage. The knowledge-axis conclusions and overall model rankings remain largely stable across captioners, whereas the fine-grained reasoning-skill results are more sensitive to caption quality. This contrast further highlights the value of separating knowledge and reasoning in ARK.

Table 26: The prompt for image caption generation.

```
prompt = f"""

Please generate a caption for this image, covering the scene context, the main objects, their relationships, actions, and attributes. The caption should be approximately 100 words and focus only on the visible content without adding speculation.

"""
```

Table 27: The prompt for query rewriting.

```
Input Format:
- Original Query: <query_text>
- Original Question Images: <query_image>

prompt = f"""

Task Description:
You are an expert in multimodal retrieval. Your goal is to rewrite the query to bridge the gap between a user's raw query and the complex visual target by explicitly externalizing implicit domain knowledge and reasoning. Perform the following execution flow to generate the refined query.

Execution Flow:
1. Dual-Axis Analysis: Identify the required knowledge domain and reasoning type.
2. Step-by-Step Reasoning:
   - Infer implicit background knowledge or domain concepts not explicitly stated in the text.
   - Translate the query's abstract reasoning logic into concrete, observable visual cues. Determine what specific objects, attributes, scene configurations, or structural patterns must be visibly present to support the query's intent.
   - Identify critical constraints that separate the ideal target from hard negatives.
3. Synthesize Target Image: Based on the reasoning above, describe the visual content and layout of the ideal target image.
4. Query Rewrite: Integrate the above information into a single, concise text search query. (~100 words)

Output Requirement:
Only output the final refined query.

"""
```

I Impact Statement

ARK is designed to support the development of more reliable multimodal retrieval. It is constructed exclusively from publicly available materials and is intended for non-commercial academic research. Following standard dataset curation practices [10, 90], annotators prioritize sources with explicit reuse permissions and refrain from using materials from websites that prohibit copying or redistribution. We also conduct best-effort screening to remove private or sensitive information and to filter out harmful content, aiming to facilitate socially beneficial research. Accordingly, ARK is primarily constructed from openly reusable resources, complemented by selected existing datasets released under compatible public licenses. ARK has been released under the CC BY-NC-SA 4.0 license to promote transparent and responsible downstream use. Overall, we do not anticipate significant ethical concerns.

J Annotation Protocol and Consistency Analysis for Reasoning-Related Labels

In this section, we provide additional details on both the data construction workflow and the instance-level annotation protocol. Figure 4 summarizes the overall curation pipeline, including data collection, reasoning-label annotation, and arbitration. Table 28 further specifies the data collection and curation rules used to construct ARK. The original data content and target pairs are constructed from reliable sources as described in Appendix B and the subsequent annotation protocol focuses on assigning and validating labels such as cognitive demand and reasoning skill. We use a structured human annotation pipeline. Annotators first receive unified training with shared annotation guidelines, then annotate instances independently, and finally resolve all disagreement cases through panel adjudication.

Instance-level annotation protocol. We provide the instance-level annotation guidelines and the corresponding agreement statistics for reasoning-related annotation in Tables 29–31. The annotation pipeline and user interface are shown in Figures 4–7.

Annotation process. The reasoning-related annotation was conducted by 6 PhD annotators with diverse academic backgrounds. Annotators assign reasoning-related labels based on the reasoning pattern required by each retrieval instance, including its query, target, and hard negatives. During annotation, annotators check whether the rewritten query is consistent with the target sample and whether the assigned reasoning-related labels reflect the intended reasoning requirement. Gemini 3 Pro is used as an auxiliary tool for difficult cases. In all cases, the final annotation decision is made by human annotators, not by the model.

Agreement and adjudication. The corresponding agreement statistics, including per-annotator accuracy and inter-annotator agreement such as Cohen's kappa, are reported in Table 32. Across annotators, Cohen's kappa values are consistently in the substantial-to-almost-perfect range (all > 0.61), indicating that the reasoning-related labels are generally reliable rather than heavily affected by annotation ambiguity. Remaining disagreement cases are resolved through panel adjudication.

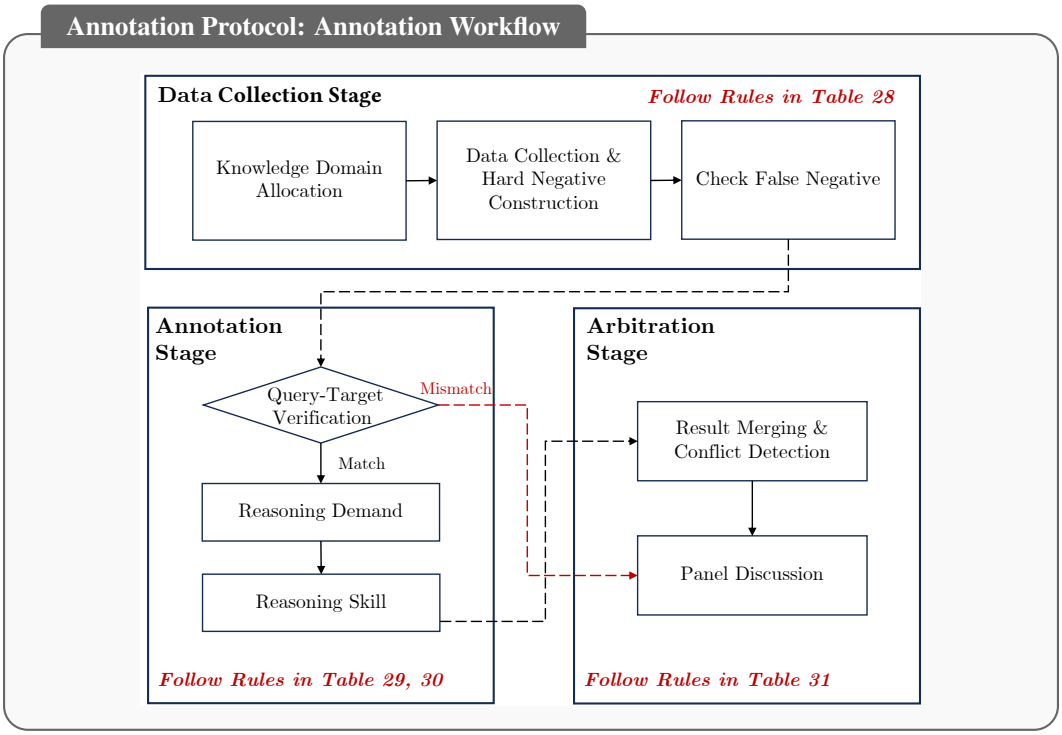


Figure 4: The three-stage construction workflow of the ARK benchmark. The pipeline begins with data collection and targeted hard negative construction. During the Annotation Stage, instances first undergo a strict Query-Target Verification step where mismatched or erroneous pairs are immediately flagged and routed to a Panel Discussion. Valid matches then proceed through systematic instance-level labeling (Reasoning Demand and Skill). Finally, the Arbitration Stage merges independent results to identify annotation conflicts. Any flagged instances undergo a Panel Discussion to resolve discrepancies and finalize the ground-truth labels. Each phase strictly adheres to predefined protocol rules (detailed in Tables 28–31) to ensure maximum dataset consistency and validity.

Table 28: The comprehensive data collection and curation protocol for the ARK benchmark.

Annotation Protocol: Data Collection

The principles for collecting and curating the multimodal query-target pairs for the ARK benchmark is given as:

1.1 Source of Data

- **Authoritativeness:** Annotators must collect instances from authoritative open-source platforms (e.g., Wikipedia, World Bank Open Data, arXiv) or expert-level academic datasets.
- **Diversity:** Annotators should prioritize finding diverse sources instead of collecting query-target pairs from a single domain to prevent structural bias.
- **Domains:** To comprehensively evaluate multimodal retrieval, annotators must ensure the collected data align with our 5 predefined meta-domains.

1.2 Ethical Considerations

- **Copyright:** Annotators should prioritize collecting materials with explicit open-source reuse permissions, such as the CC BY-NC-SA 4.0 license. Data from sources that explicitly prohibit copying, scraping, or redistribution **MUST** be strictly avoided.
- **Privacy and Safety:** Annotators must conduct a best-effort manual screening during the collection phase to identify and permanently filter out any images or texts containing personally identifiable information, sensitive content, or harmful materials to ensure the dataset remains safe for socially beneficial research.

1.3 Hard Negatives

- **High Surface-Level Similarity:** A valid hard negative must share significant semantic or visual overlap with the ground truth target.
- **Strict Logical Violation:** The negative **MUST** definitively violate at least one core constraint specified in the query. (e.g., a slightly swapped coordinate in a cognitive map)
- **No Ambiguity:** Annotators must verify that the negative is unambiguously incorrect based only on the provided textual and visual evidence. If a distractor could be arguably correct under a different interpretation, it must be discarded.

Table 29: Protocol for labeling specific reasoning demand in ARK dataset.

Annotation Protocol: Decision Rules for Reasoning Demand

1. Knowledge-Intensive

- **Keyword:** Expert-level, domain-specific concepts
- **Decision Rules:** Assign this label if distinguishing the target from hard negatives primarily depends on recognizing **expert-level, domain-specific concepts rather than general, everyday knowledge**. The matching process relies heavily on internal factual recall rather than multi-step logical deduction.
- **Examples:** Retrieving an image of an "Amur leopard". Distinguishing it from hard negatives (like jaguars, cheetahs, or other leopard subspecies) requires expert-level biological knowledge of specific coat patterns and taxonomy, rather than basic "animal" recognition.

2. Reasoning-Intensive

- **Keyword:** Explicitly stated characteristics; common sense; complex reasoning
- **Decision Rules:** Assign this label if the required domain knowledge is everyday common sense or if the target's characteristics are explicitly provided in the query. The core bottleneck is not recalling facts, but **applying multi-step inference, relational understanding, spatial alignment, or structural abstraction** to select the correct target over confusing distractors.
- **Examples:** A query such as: "Find the desk where the notebook is between the mug and the lamp, and the number of red pens is greater than the number of blue pens." Here, the model does not need specialized domain knowledge. Instead, it must integrate multiple relational and counting constraints to distinguish the target from highly similar distractors.

3. Both

- **Keyword:** Two-step cognitive process; expert-level knowledge; complex reasoning
- **Decision Rules:** Assign this label if the task strictly requires a **two-step cognitive process**: the model must first recall the correct domain rules or specialized concepts mentioned in the query, and then apply those rules through a deductive, spatial, or symbolic process to distinguish the ground truth from highly plausible hard negatives.
- **Examples:** Mathematical proofs, physics phenomena, or chemistry reactions where the query involves symbolic formulas or specialized terms.

4. None

- **Keyword:** Perception; common sense
- **Decision Rules:** Assign this label if the task requires neither specialized domain knowledge nor complex reasoning. The retrieval can be easily solved via **basic surface-level semantic matching**. It relies entirely on everyday common sense, general object recognition, and basic vocabulary, with no deceptive hard negatives that require multi-step verification.
- **Examples:** In a daily-life image retrieval scenario, the query is simply "a man riding a bicycle in a park." The model relies entirely on general common sense to align the basic text concepts with the visual elements. There are no challenging distractors that require domain expertise or structural deduction to filter out.

Table 30: Protocol for labeling specific reasoning skills in ARK dataset.

Annotation Protocol: Decision Rules for Reasoning Skills

1. Knowledge Reasoning

- **Decision Rules:** Assign this label if the retrieval requires inference grounded in background domain knowledge, **going explicitly beyond pure recognition or factual recall**. This skill frequently co-occurs with other reasoning types. The model must apply domain rules, facts, or historical contexts to deduce the correct target, rather than simply identifying a named entity.
- **Example:** In a chemistry retrieval scenario, the query asks for the immediate product of a specific chemical reaction. This inherently involves **both** Knowledge Reasoning (recalling and applying specific chemical transformation principles) and Symbolic Reasoning (deducing the resulting structural changes).

2. Spatial Reasoning

- **Decision Rules:** Assign this label if the task strictly requires understanding geometric transformations, topological relations, or precise relative positioning. The bottleneck **must involve mentally manipulating spatial structures, such as 3D folding/rotation**.
- **Examples:** In a 3D geometry scenario, the query provides a 3D block structure and a viewing direction arrow. The model must mentally perform a 3D-to-2D projection (Spatial Reasoning) to distinguish the correct 2D silhouette.

3. Fine-Grained Visual Reasoning

- **Decision Rules:** Assign this label if the target needs to be identified by subtle, localized, or detail-sensitive visual cues (e.g., tiny objects, specific textures, or small text) within high-resolution images. The model must perform a meticulous "**thinking with images**" inspection process to verify these localized conditions.
- **Examples:** In a high-resolution natural landscape scenario, the query asks for a specific "German nature park" featuring a small, localized informational sign. The model cannot rely on the overall forest appearance; it must zoom in to verify the tiny text on the signboard (Fine-Grained Visual Reasoning) to reject hard negatives showing similar ruins or parks.

4. Logical Reasoning

- **Decision Rules:** Assign this label if correct retrieval depends on **structured deduction over explicit conditions, relations, or Boolean logic** (e.g., conjunction, disjunction, negation, implication).
- **Examples:** In a digital circuit scenario, the query describes a logical expression `(A xor B) nand (C xor D)`. The model must trace the logic gates (Logical Reasoning) to select the correct circuit diagram over hard negatives that have identical visual layouts but swap an XOR gate for an OR gate.

5. Symbolic Reasoning

- **Decision Rules:** Assign this label if the task involves interpreting **structured symbolic representations** (e.g., mathematical formulas, chemical equations, or code snippets). This typically requires mapping structured symbolic representations to diagrams or graphical plots.
- **Examples**: In a code-driven graphics scenario, the query provides a raw TikZ code snippet specifying a geometric shape. The model must interpret the symbolic coordinate syntax and programmatic commands to map them to the correct rendered drawing.

Table 30: Protocol for labeling specific reasoning skills in ARK dataset. Continued.

Annotation Protocol: Decision Rules for Reasoning Skills Continued

6. Conceptual Abstraction

- **Decision Rules:** Assign this label if the task requires higher-level inductive generalization, interpreting metaphors, or summarizing abstract takeaways from visual trends. The model must **move from concrete visual elements** (like chart lines or cartoon characters) **to abstract concepts** (like economic shifts or historical viewpoints).
- **Examples:** In an economics scenario, the query asks for a chart demonstrating how "near-term net payoff is governed chiefly by the match between skills and demand." The model must abstract the high-level economic conclusion from numerical trends.

Table 31: Conflict resolution pipeline and guidelines for handling annotation ambiguities.

Annotation Protocol: Handling Ambiguities

To ensure the high validity and consistency of the ARK benchmark, we adopt the following conflict resolution pipeline for any disagreements:

- **Merging:** Once all annotators have completed their batches, any instance with conflicting labels is automatically flagged and exported to a centralized merging tool for arbitration by the core authors.
- **Arbitration:** For each disputed instance, the core authors review the conflicting labels alongside the raw data to reach a consensus based on the established decision rules.
- **Discarding:** If the core authors determine during the discussion that an instance is inherently ambiguous, the query and target are fundamentally mismatched, or the hard negative cannot be definitively eliminated without subjective assumptions—the instance is immediately discarded from the benchmark. **We prioritize annotation certainty over dataset size.**

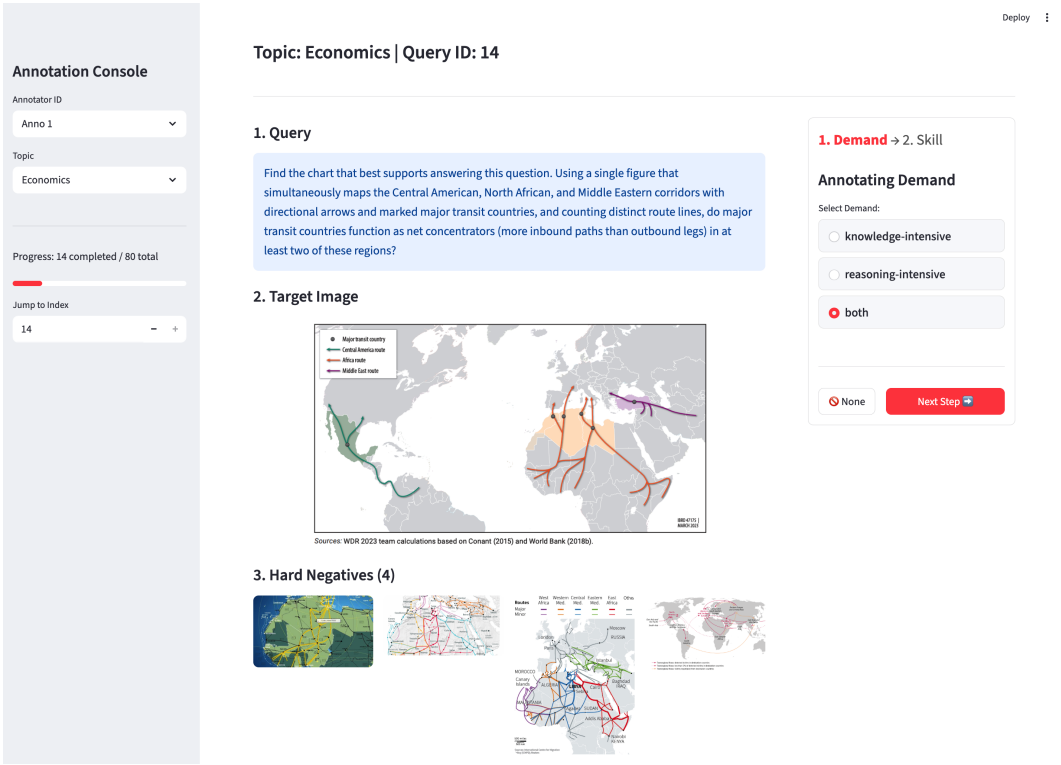


Figure 5: **Annotation Interface: Reasoning Demand**. The interface displays the text query, target image, and corresponding hard negatives. Annotators select the appropriate demand label on the right and click "Next Step" to proceed. Instances labeled as "Reasoning-Intensive" or "Both" proceed to the Reasoning Skill annotation stage, while instances labeled as "Knowledge-Intensive" or "None" immediately advance to the next sample without requiring reasoning-skill annotation. The left panel provides topic navigation and tracks overall annotation progress.

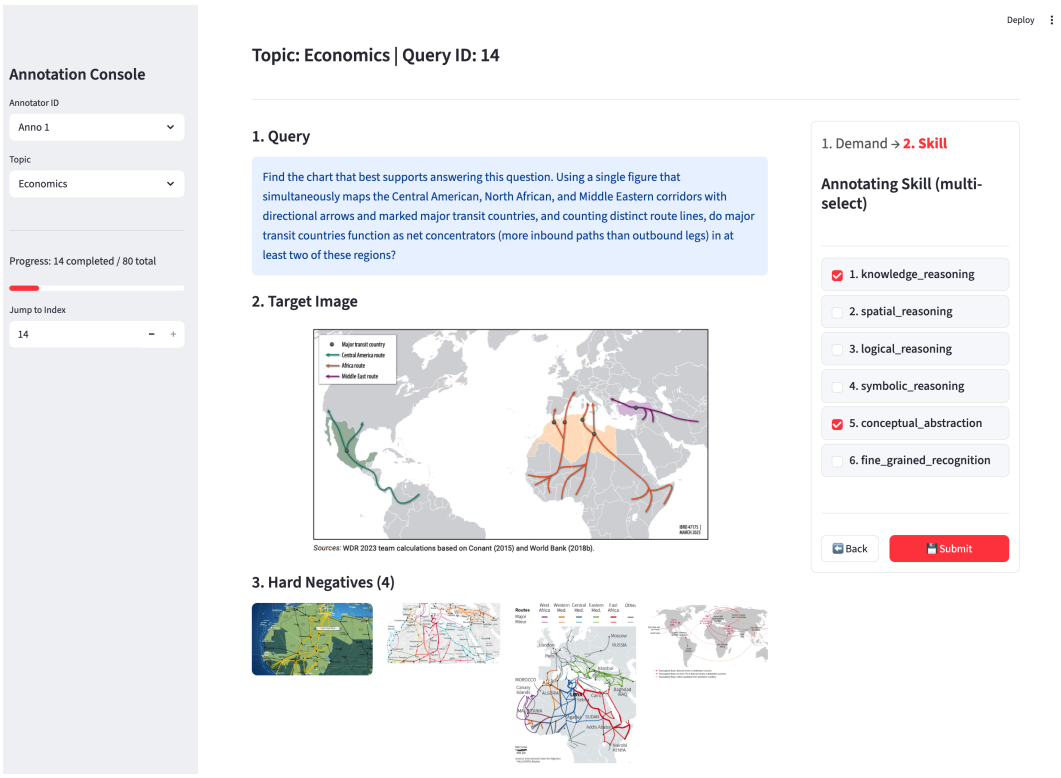


Figure 6: **Annotation Interface: Reasoning Skill.** The interface displays the text query, target image, and corresponding hard negatives. On the right panel, the annotators must select at least one applicable reasoning skill to activate the "Next Step" button and proceed to the Primary Evidence annotations.

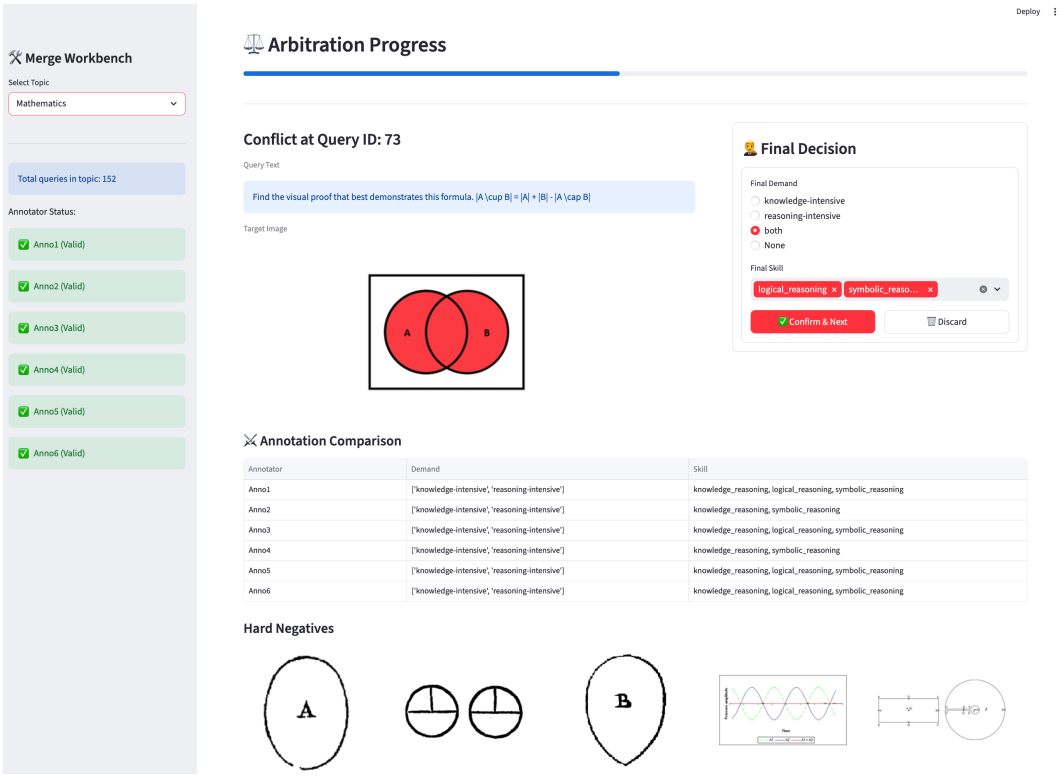


Figure 7: **Arbitration Interface**. It displays the text query, target image, hard negatives, and the conflicting label records from different annotators. Authors use this interface to finalize the ground-truth labels based on established guidelines. If an instance is inherently ambiguous or mismatched, the panel clicks "Discard" to eliminate the sample and ensure dataset validity.

Table 32: **Annotation consistency analysis**. The table reports the accuracy, Cohen's kappa coefficients, and qualitative agreement levels of six annotators across two distinct tasks: Reasoning Demand (4-class) and Reasoning Skill (6-class). Here, Cohen's kappa measures the level of agreement between the annotators' assigned labels and the final consensus labels. Qualitative agreement levels follow standard interpretation guidelines (Poor: < 0.00, Slight: 0.00-0.20, Fair: 0.21-0.40, Moderate: 0.41-0.60, Substantial: 0.61-0.80, Almost Perfect: 0.81-1.00). The leftmost column lists the annotators and their academic backgrounds: CS (Computer Science), BE (Biomedical Engineering), and ME (Mechanical Engineering). Confidence intervals for the kappa coefficients are provided in parentheses.

Domain	Reasoning Demand			Reasoning Skill		
	Accuracy (4-class)	Cohen's kappa	Agreement level	Accuracy (6-class)	Cohen's kappa	Agreement level
Annotator 1 (CS)	0.82	0.74 (0.70, 0.78)	Substantial	0.82	0.81 (0.78, 0.83)	Almost perfect
Annotator 2 (CS)	0.88	0.84 (0.81, 0.87)	Almost perfect	0.80	0.78 (0.75, 0.81)	Substantial
Annotator 3 (CS)	0.89	0.83 (0.81, 0.86)	Almost perfect	0.85	0.83 (0.82, 0.85)	Almost perfect
Annotator 4 (CS)	0.83	0.77 (0.76, 0.81)	Substantial	0.82	0.78 (0.74, 0.81)	Substantial
Annotator 5 (BE)	0.88	0.81 (0.78, 0.82)	Almost perfect	0.81	0.75 (0.73, 0.78)	Substantial
Annotator 6 (ME)	0.85	0.79 (0.76, 0.81)	Substantial	0.80	0.78 (0.75, 0.82)	Substantial

K Case Study

In this section, we present representative examples for each task in ARK. Figures 8–24 provide qualitative examples. These cases highlight the diversity of knowledge domains, visual data types, and reasoning skills covered by the benchmark.

Art

Query

A Flemish school interior piece by Jan Jozef Horemans II, representing a characteristic work by this Antwerp genre painter depicting a patrician setting.

Ground Truth



Hard Negative

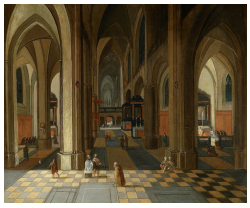





Figure 8: Visualized example of Art subtype.

Biology

Query

A resting Common Raven on the ground.

Ground Truth



Hard Negative

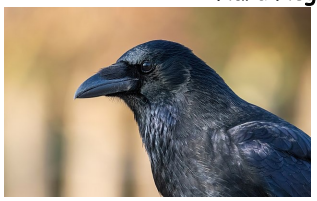








Figure 9: Visualized example of Biology subtype.

Calendar

Query

A calendar month page from which one can determine that people born in the lunar second month can celebrate their lunar birthday twice.

Ground Truth

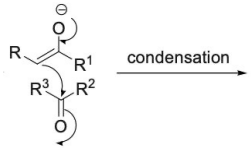


Figure 10: Visualized example of Calendar subtype.

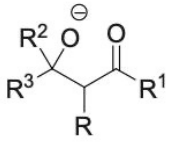
Chemistry

Query

An 'arrow-pushing' diagram is a common type of chemical image used to illustrate electron flow in mechanistic steps. The transition-state structure undergoes changes after the electron has been relocated or reacted. Which of the following options shows the structure after the change? Note: Disregard the arrows in the options. The selected option should represent the immediate product of the next step, not the final product of the reaction.\n<|image|>



Ground Truth



Hard Negative

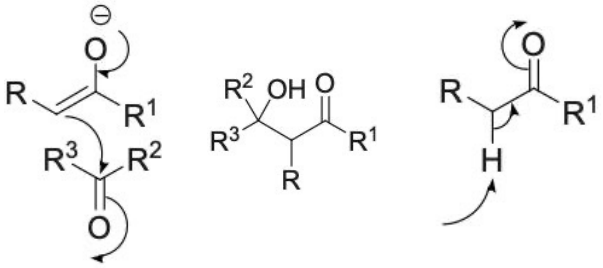


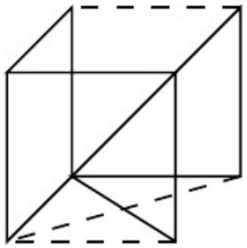
Figure 11: Visualized example of Chemistry subtype.

Code-Drawing

Query

\documentclass[tikz,border=2pt,png]{standalone} \usepackage{tkz-euclide} \begin{document} \begin{tikzpicture} \coordinate (O) at (0,0,0); \coordinate (A) at (1,0,0); \coordinate (B) at (1,1,0); \coordinate (C) at (0,1,0); \coordinate (D) at (0,0,1); \coordinate (E) at (1,0,1); \coordinate (F) at (1,1,1); \coordinate (G) at (0,1,1); \draw (O) — (A) — (B) — (F) — (E) — cycle; \draw (O) — (C) — (G) — (F) — cycle; \draw (O) — (D) — (G); \draw[dashed] (E) — (D); \draw[dashed] (B) — (C); \draw[dashed] (A) — (D); \end{tikzpicture} \end{document

Ground Truth



Hard Negative

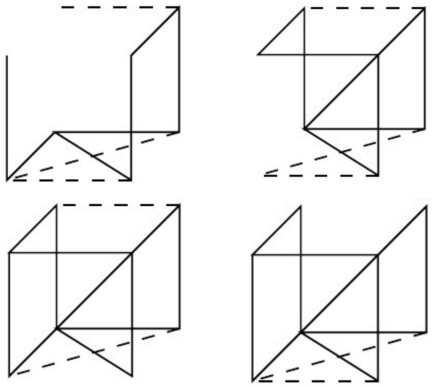


Figure 12: Visualized example of Code-Drawing subtype.

Comics

Query

Find a cartoon from the 1890s that illustrates the fragile and temporary truce between employers and labors immediately preceding a major industrial conflict.

Ground Truth

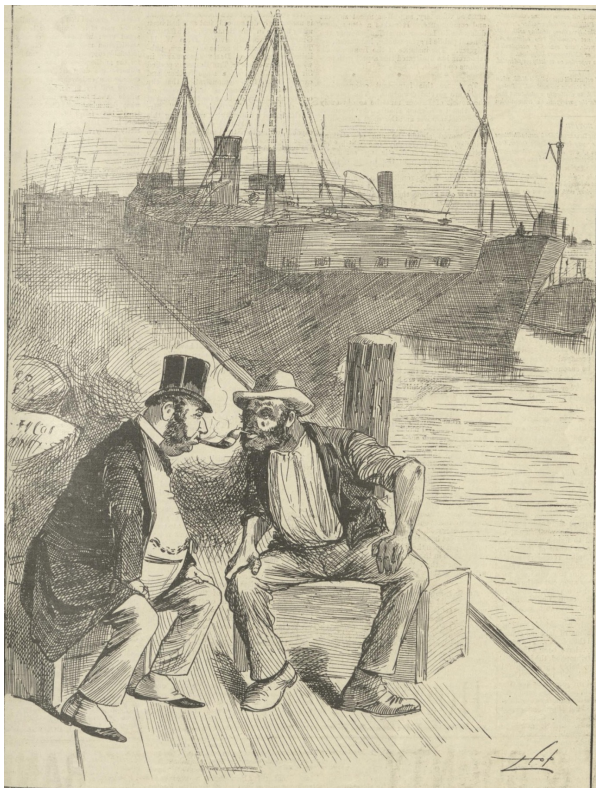


Figure 13: Visualized example of Comic subtype.

Computer-Science

Query

A conceptual diagram of Canonical Correlation Analysis (CCA) inference illustrating the cross-modal mapping process, where an input image and a candidate text are projected into a shared normalized vector space, establishing semantic correspondence by measuring the cosine similarity represented by the angle between their respective unit embeddings.

Ground Truth

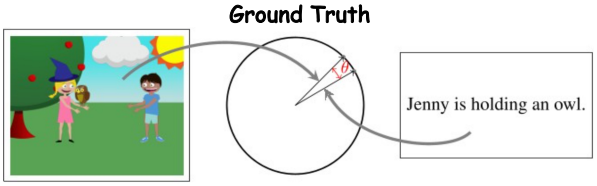


Figure 14: Visualized example of Computer-Science subtype.

Daily-Life

Query

A black Honda motorcycle is parked on the outdoor grass, with houses and surrounding green plants visible in the background.

Ground Truth



Hard Negative









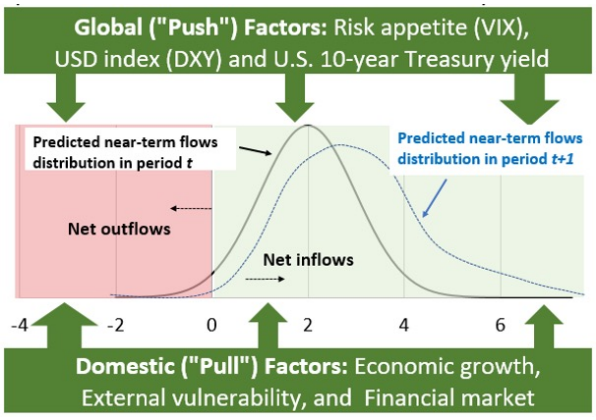
Figure 15: Visualized example of Daily-Life subtype.

Economics

Query

Find the chart that best supports answering this question. Assuming a host has a duty to admit people escaping danger but discretion over opportunity-driven entrants, which diagram best supports concluding-within a generic two-by-two that crosses push vs. pull motives with the strength of the skills-demand alignment-that the near-term net payoff is governed chiefly by the match (with an explicit side-by-side contrast of net gains versus net losses along the match gradient), while avoiding country-specific examples and omitting prescriptive policy to-do lists?

Ground Truth



Hard Negative

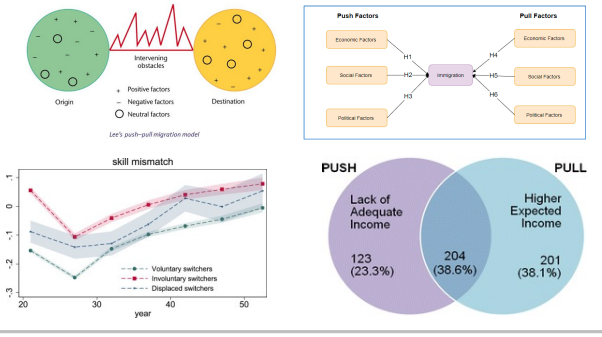


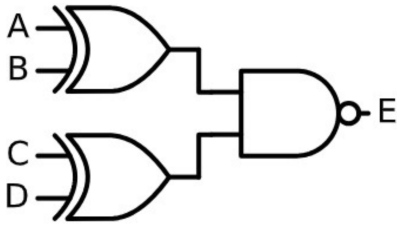
Figure 16: Visualized example of Economics subtype.

Electronics

Query

A circuit diagram illustrating the logical expression (A xor B) nand (C xor D).

Ground Truth



Hard Negative

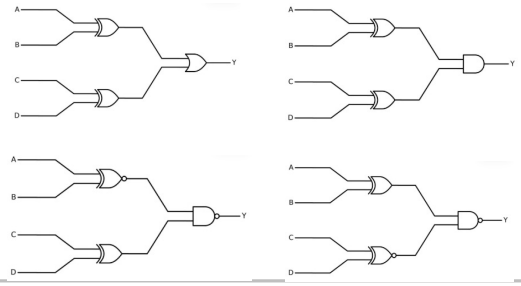


Figure 17: Visualized example of Electronics subtype.

Fine-Grained

Query

A German nature park.

Ground Truth



Hard Negative



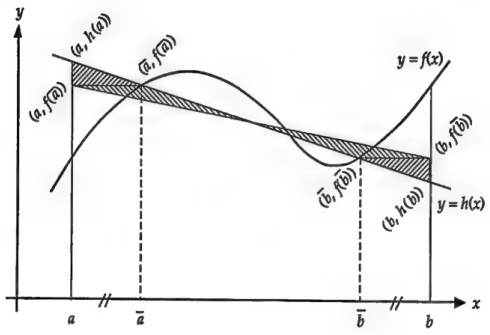
Figure 18: Visualized example of Fine-Grained subtype.

Mathematics

Query

Find the visual proof that best demonstrates this formula. \frac{1}{2}(b-a)(f(\bar{a})+f(\bar{b}))=\frac{1}{2}(b-a)(h(a)+h(b)).

Ground Truth



Hard Negative

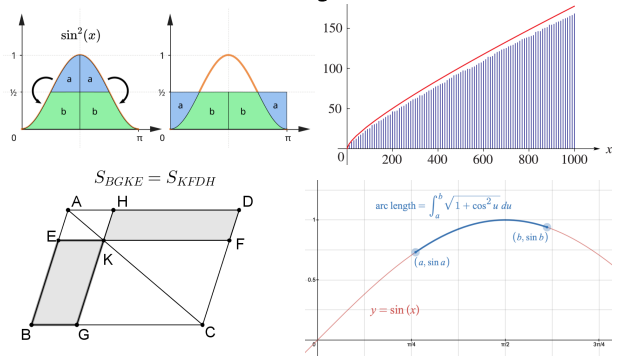


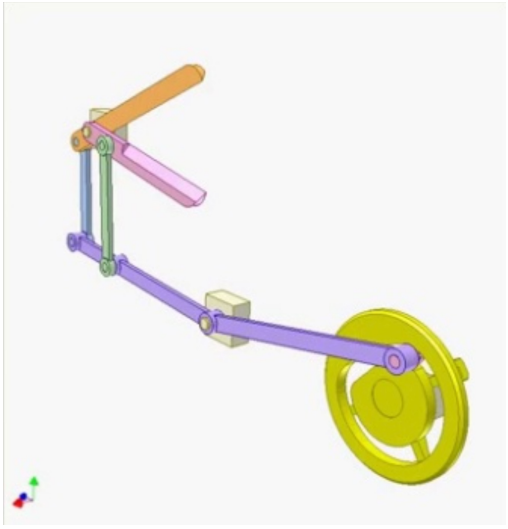
Figure 19: Visualized example of Mathematics subtype.

Mechanical

Query

Cam-driven scissors.

Ground Truth



Hard Negative

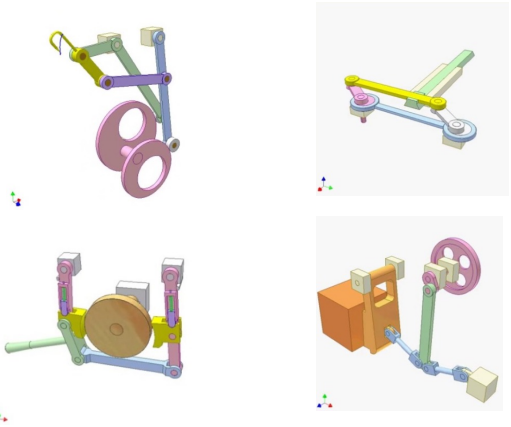


Figure 20: Visualized example of Mechanical subtype.

Metro

Query

A Metro map where only one line crosses the same river twice.

Ground Truth



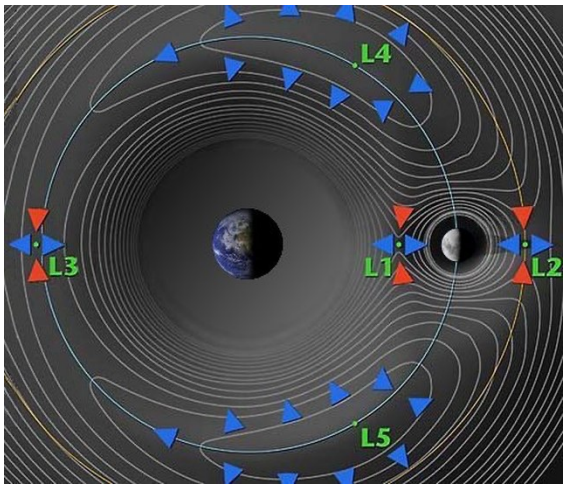
Figure 21: Visualized example of Metro subtype.

Physics

Query

Effective Potential Contours and Lagrange Points.

Ground Truth



Hard Negative

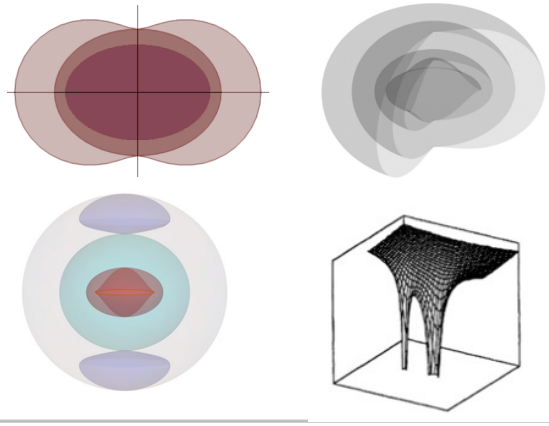
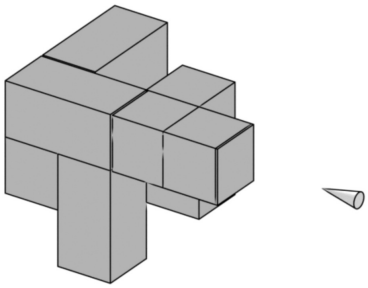


Figure 22: Visualized example of Physics subtype.

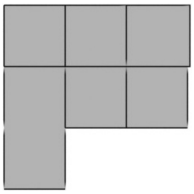
Spatial-3D

Query

Given the arrow indicating the viewing direction on the 3D figure above, find the 2D projection image that best represents the projection of the structure from that perspective.



Ground Truth



Hard Negative

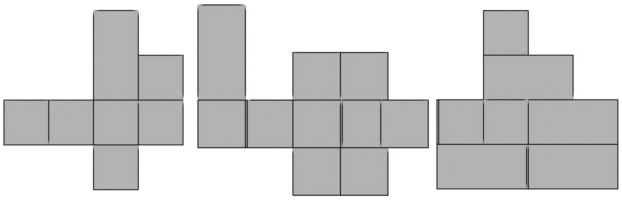


Figure 23: Visualized example of Spatial-3D subtype.

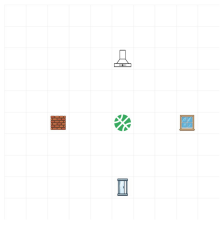
Spatial-Cogmap

Query

Retrieve the corresponding bird's-eye view cognitive map for the provided scene images. The map represents the spatial layout on a 10x10 coordinate grid (from [0,0] at the top-left to [9,9] at the bottom-right), where object positions accurately reflect their relative distances and spatial relationships as seen in the partial views.



Ground Truth



Hard Negative

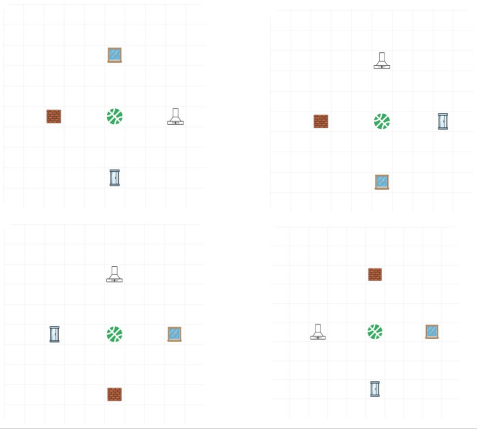


Figure 24: Visualized example of Spatial-Cogmap subtype.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses several limitations and remaining bottlenecks, such as low performance on reasoning-intensive cases, fine-grained visual and spatial reasoning failures, and the fact that rewriting and re-ranking leave substantial headroom.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not claim new theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper. The dataset and evaluation pipeline have been released.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset and evaluation pipeline have been released.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the retrieval corpus per subtype, the metrics, evaluated models and model links. Since the work is an evaluation benchmark rather than a trained model, optimizer and training hyper-parameters are not applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Because retrieval inference is deterministic, we quantify sampling uncertainty using 1,000 query-level bootstrap resamples within each subtype or reasoning group. Appendix F reports percentile-based 95% confidence intervals for the main Recall@1 results on both the knowledge and reasoning axes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix C states that locally executed inference experiments were conducted on NVIDIA RTX 4090 GPUs, that no training or fine-tuning was performed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available or permission-compatible materials, screens for sensitive and harmful content. We have reviewed the NeurIPS Code of Ethics and are not aware of deviations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the societal impacts of the work in Appendix I.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The paper describes safeguards including reuse-permission checks, exclusion of sources prohibiting redistribution, best-effort screening for sensitive information and harmful content.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Proper citations and links are provided throughout the document.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.

- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper documents the task formulation, taxonomy, data statistics, subtype-level construction details, annotation protocol, and agreement analysis. The Impact Statement also states that ARK has been released under CC BY-NC-SA 4.0.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing platforms. The data curation and annotation are conducted by a small group of trained annotators as part of the research process.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: The work does not involve crowdsourcing or research with human subjects that would require IRB-style approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are used as auxiliary tools during data curation and for image captioning, query rewriting, and agentic retrieval, as described in the paper. All instances and annotations are verified by human annotators through independent review and panel adjudication.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.