Vision-DeepResearch Benchmark: Rethinking Visual and Textual Search for Multimodal Large Language Models

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have advanced VQA and now support Vision-DeepResearch systems that use search engines for complex visual-textual fact-finding. However, evaluating these visual and textual search abilities is still difficult, and existing benchmarks have two major limitations. First, **existing benchmarks are not visual search-centric**: answers that should require visual search are often leaked through cross-textual cues in the text questions or can be inferred from the prior world knowledge in current MLLMs. Second, **overly idealized evaluation scenario**: On the image-search side, the required information can often be obtained via near-exact matching against the full image, while the text-search side is overly direct and insufficiently challenging. To address these issues, we construct the *Vision-DeepResearch benchmark* ***(VDR-Bench)*** comprising 2,000 VQA instances. All questions are created via a careful, multi-stage curation pipeline and rigorous expert review, designed to assess the behavior of Vision-DeepResearch systems under realistic real-world conditions. Moreover, to address the insufficient visual retrieval capabilities of current MLLMs, we propose a simple multi-round cropped-search workflow named MVF, which effectively improves model performance in realistic visual retrieval scenarios. Overall, our results provide practical guidance for future multimodal deep-research systems.

1 Introduction

The rapid progress of Multimodal Large Language Models (MLLMs) (Bai et al., 2025; Huang et al., 2025; Chen et al., 2024a,b; Wang et al., 2025; Yu et al., 2024; Zeng et al., 2025a,b) has enabled a new class of multimodal deep-research systems (Geng et al., 2025; Wu et al., 2025b; Narayan et al., 2025; Hong et al., 2025), which combine image understanding, web search, and multi-hop reasoning to solve complex visual–textual queries. These systems promise to automate deep visual fact-finding by grounding reasoning in real-world images and external knowledge sources. Despite this progress, evaluating the visual and textual search capabilities of multimodal deep-research systems remains challenging. Existing benchmarks for multimodal search and knowledge-seeking VQA (Geng et al., 2025; Jiang et al.; Fu et al., 2025; Wang et al., 2017; Cheng et al., 2025; Chen et al., 2023) suffer from two fundamental limitations, as illustrated in Fig. 1:

First, current benchmarks are not visual search–centric. Many benchmark questions can be answered without genuine visual verification, *i.e.*, some answers are inferred from language priors with cross-textual cues, and even MLLM's world knowledge. As a simple example (shown in Fig. 1), some VQA problems in these benchmarks can be resolved by exploiting cross-referencing among textual entities in the question: taking the intersection of multiple candidate sets that satisfy different textual constraints typically yields a small set, making it easy to pin down (or even guess) the answer. We view this as a shortcut phenomenon. Furthermore, as shown in Fig. 1, some VQA instance can simply solved by MLLM's prior world knowledge. As a result, benchmark performance often reflects memorization or text-only retrieval rather than the the necessary for gathering visual information.

Second, the evaluation scenario of existing benchmarks rely on overly idealized search settings. On the image-search side, required information is frequently obtained through near-exact whole-image matching, where querying the original image retrieves an identical or near-duplicate source along with its title or metadata. On the text-search side, questions are often formulated in a direct and shallow manner, underestimating the difficulty of multi-hop reasoning and evidence aggregation. These conditions fail to reflect the noisy, ambiguous, and iterative nature of real-world vi-

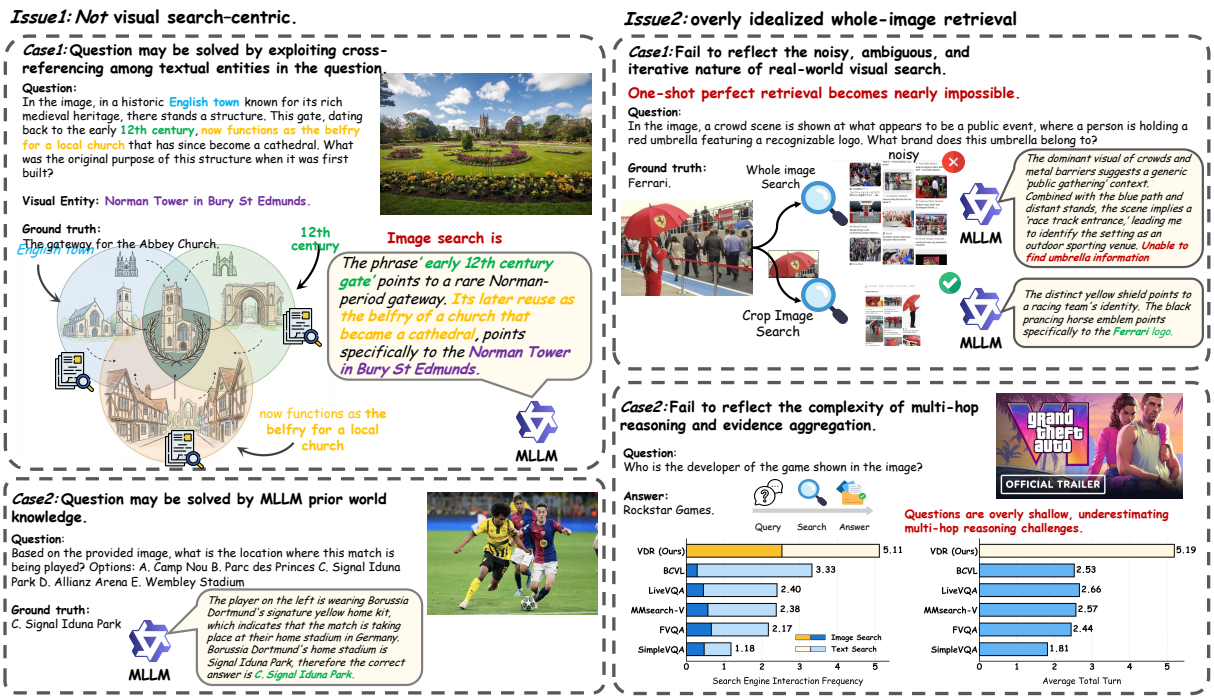

Figure 1: **Motivation**. Existing Vision-DeepResearch benchmarks often fail to measure realistic multimodal search. ***Left***: Many questions can be solved via text-only cues or model priors without genuine visual verification. ***Right***: Whole-image search frequently retrieves near-identical images with identifying metadata ("perfect retrieval"), and the search queries are overly direct and insufficiently challenging. VDR-Bench is designed to be visual-search–centric and to reflect real-world settings that require iterative, entity-level localization (e.g., multi-round cropping), cross-modal evidence collection, and multi-hop reasoning.

sual search. In realistic scenarios, visual search is inherently trial-based and iterative. Images often contain multiple entities, background distractors, and ambiguous visual cues, making whole-image retrieval unreliable. Effective multimodal deep-research systems should therefore localize candidate entities, issue multiple cropped image queries, refine hypotheses across rounds, and verify results through cross-modal reasoning, in order to adapt to real-world requirement. Multi-round, multi-scale cropping is not merely an implementation detail but a necessary strategy for robust visual retrieval.

To address these challenges, we introduce **VDR-Bench**, a large-scale benchmark comprising 2,000 carefully curated VQA instances designed to evaluate realistic multimodal search behavior. VDR-Bench is constructed through a rigorous multi-stage pipeline that emphasizes visual-first entity discovery, human verification, and multi-hop query expansion, ensuring that solving each instance genuinely requires both visual search and textual reasoning. Unlike prior benchmarks that convert text-based QA into VQA through single-step image retrieval (Geng et al., 2025), our benchmark avoids shortcut-prone instances and minimizes perfect-retrieval bias. Furthermore, to improve the insufficient visual retrieval capabilities of current MLLMs, we propose a simple yet effective multi-round cropped-search workflow named MVF. This strategy iteratively refines visual queries by cropping regions of interest, enabling more accurate entity localization and reducing retrieval noise. Experimental results demonstrate that this workflow substantially improves performance on realistic visual retrieval tasks, highlighting a practical direction for building more capable multimodal deep-research systems. Our main contributions are summarized as follows:

- We identify critical limitations in existing multimodal search benchmarks, including the lack of **visual-search–centric evaluation** and **overly idealized retrieval settings**.
- We introduce **VDR-Bench**, a new benchmark with 2,000 instances curated through a multi-stage, human-verified pipeline that enforces realistic visual and textual search requirements.
- We propose a **multi-round cropped-search workflow** that significantly improves model performance in realistic visual retrieval scenarios, offering practical guidance for future multimodal deep-research systems.

2 Related Work

2.1 Vision-DeepResearch Systems.

Vision-DeepResearch agents are designed to autonomously search, read, and synthesize knowledge from the open web through multi-step reasoning and cross-modal grounding. While recent progress in text-only deep research agents (Team et al., 2025; Wu et al., 2025a; Li et al., 2025; Tao et al., 2025) has been substantial, the increasing visual complexity of real-world web environments has driven a paradigm shift toward multimodal Vision-DeepResearch systems. Compared with text-only counterparts, these systems face significantly greater challenges, as they must integrate visual evidence, perform reliable visual search, and conduct cross-modal verification under noisy and ambiguous conditions.

To address these challenges, several Vision-DeepResearch frameworks have been proposed. WebWatcher (Geng et al., 2025) reformulates text-based question answering into visual question answering (VQA) via reverse image search. MMSearch-R1 (Wu et al., 2025b) enhances multimodal search strategies through Group Relative Policy Optimization (GRPO) (Shao et al., 2024). Building upon these works, DeepMMSearch-R1 (Narayan et al., 2025) improves visual grounding by employing entity-level image cropping, reducing background noise and enabling more targeted visual retrieval. Despite these advances, existing systems still struggle with the complexity and noise of realistic visual search environments.

2.2 Evaluations of Vision-DeepResearch Systems.

The rapid development of Vision-DeepResearch systems requires evaluation benchmarks that accurately measure visual and textual search capabilities under realistic conditions. SimpleVQA (Cheng et al., 2025) assesses multimodal factuality and hallucination robustness, while FVQA (Wang et al., 2017) and InfoSeek (Chen et al., 2023) require models to ground visual evidence in external knowledge sources. LiveVQA (Fu et al., 2025) evaluates reasoning over dynamic, real-time visual information, whereas MMSearch (Jiang et al.) and BrowseComp-VL (Geng et al., 2025) examine agentic competence in multi-step multimodal search and deep research workflows.

Although these benchmarks provide valuable progress, they exhibit two critical limitations aligned with our analysis. First, most existing benchmarks are not visual search–centric: many instances can be solved through language priors or text-only retrieval without requiring genuine visual verification, resulting in shortcut solutions that fail to evaluate true visual retrieval and grounding abilities. Second, current benchmarks rely on overly idealized search settings: on the image side, near-exact whole-image matching often enables perfect retrieval, while on the text side, queries are frequently shallow and insufficiently challenging. These design choices underestimate the noise, ambiguity, and iterative nature of real-world visual–textual search.

3 A Quantitative Analysis of Existing Vision-DeepResearch Benchmarks

As discussed in the Fig. 1, existing Vision-DeepResearch benchmarks suffer from two core limitations: (1) many instances are not truly **visual-search–centric**, and (2) image retrieval is often conducted under **overly idealized evaluation scenario**. To quantify the severity of these issues, we conduct a controlled empirical study across representative multimodal search and knowledge-seeking VQA benchmarks, including SimpleVQA (Cheng et al., 2025), LiveVQA (Fu et al., 2025), FVQA (Wang et al., 2017), BrowseComp-VL (BCVL) (Geng et al., 2025), and MMSearch (Jiang et al.).

We evaluate representative open and closed-source multimodal large language models, including Gemini-2.5 Pro (Comanici et al., 2025) and Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025), as well as the text-based deep research model Tongyi-DeepResearch-30B-A3B (Team et al., 2025). These models are evaluated under controlled retrieval configurations to disentangle the independent contributions of whole-image search, text search, and model prior knowledge. We compare two experimental settings: (1) Given the original image and question: Direct Answer without external search, Whole-Image Search (WIS), Text Search (TS), and WIS + TS. (2) Replacing the image with caption (produced by Qwen3-VL-235B-A3B-Instruct) alongside the question: Direct Answer and Text Search only. This design allows us to evaluate whether benchmark instances genuinely require visual search, or can instead be solved via text-only retrieval or model priors. The results are reported in Tab. 1, where arrows (↑ / ↓) indicate

Table 1: Performance comparison across visual search benchmarks. WIS (Whole Image Search), TS (Text Search). Arrows (↑/↓) denote performance changes relative to the corresponding **No Search** baseline.

Type	Setting	SimpleVQA	LiveVQA	FVQA	BCVL	MMSearch	VDR-Bench
Gemini 2.5 Pro							
Image	Direct Answer	69.7	60.3	60.7	43.1	39.8	8.2
	WIS	75.3 ↑5.6	64.3 ↑4.0	66.7 ↑6.0	42.6 ↓0.5	50.3 ↑10.5	9.8
	TS	67.3 ↓2.4	72.7 ↑12.4	69.3 ↑8.6	49.9 ↑6.8	57.3 ↑17.5	5.4
	WIS+TS	74.7 ↑5.0	76.3 ↑16.0	75.7 ↑15.0	52.9 ↑9.8	69.6 ↑29.8	10.6
Capt.	Direct Answer	54.0	46.3	37.3	38.1	28.1	4.6
	TS	54.7 ↑0.7	56.0 ↑9.7	43.0 ↑5.7	43.9 ↑5.8	33.3 ↑5.2	5.2
Qwen3-VL-30B-A3B-Instruct							
Image	Direct Answer	56.3	42.7	34.7	29.6	18.7	3.8
	WIS	67.7 ↑11.4	48.3 ↑5.6	53.7 ↑19.0	35.1 ↑5.5	36.3 ↑17.6	5.2
	TS	48.3 ↓8.0	53.3 ↑10.6	45.0 ↑10.3	37.5 ↑7.9	40.9 ↑22.2	5.8
	WIS+TS	70.0 ↑13.7	62.7 ↑20.0	73.7 ↑39.0	46.6 ↑17.0	66.7 ↑48.0	6.0
Capt.	Direct Answer	51.7	42.0	31.0	31.3	15.8	4.6
	TS	49.0 ↓2.7	46.7 ↑4.7	43.0 ↑12.0	37.9 ↑6.6	38.0 ↑22.2	7.4
Tongyi-DeepResearch-30B-A3B							
Capt.	Direct Answer	49.7	44.7	31.7	29.4	14.6	4.0
	TS	46.3 ↓3.4	60.7 ↑16.0	45.7 ↑14.0	44.4 ↑15.0	45.0 ↑30.4	7.5

changes relative to the No Search baseline.

3.1 Finding 1: Existing Benchmarks Do Not Enforce Visual-Search–Centric Reasoning

Tab. 1 reveals a clear and concerning issue: a substantial portion of instances in existing Vision-DeepResearch benchmarks can be solved without meaningful visual search. Across multiple datasets, enabling TS alone yields significant performance gains, and in some cases achieves results comparable to or even exceeding those obtained with Image Search (e.g., for Gemini 2.5 Pro and Qwen3-VL-30B-A3B-Instruct, TS substantially outperforms WIS on MMSearch and LiveVQA). This indicates that many benchmark questions suffer from severe textual cue leakage, where answers can be inferred through cross-validation of textual evidence without requiring image search for visual entities.

Additionally, in several benchmarks, the Caption + Direct Answer setting attains competitive performance (e.g., Tongyi-DeepResearch-30B-A3B achieves a 30.4% gain via TS on MMSearch), suggesting that models can often bypass visual evidence entirely by relying on language priors and parametric world knowledge. Taken together, these findings demonstrate that many existing benchmarks fail to enforce a closed visual-evidence loop and systematically overestimate models' true visual search and verification capabilities. Current benchmark performance often reflects proficiency in text retrieval and prior memorization, rather than genuine visual-search–centric reasoning.

3.2 Finding 2: Overly Idealized Retrieval Evaluation Setting

Our analysis further uncovers a second limitation in current benchmarks: search is often conducted under unrealistically idealized conditions.

As shown in Tab. 1, enabling single-shot WIS alone frequently yields substantial performance gains relative to the Direct Answer baseline (e.g., Gemini 2.5 Pro and Qwen3-VL-30B-A3B-Instruct achieve gains of 10.5% and 17.6% on MMSearch, respectively). This behavior reveals a strong perfect-match bias, where querying the full image retrieves a near-duplicate copy accompanied by identifying titles or metadata, effectively reducing the task to a one-shot lookup. Under such conditions, benchmark difficulty is dominated by whether the search engine returns an exact or near-exact match, rather than evaluating an agent's ability to localize entities, refine visual queries, or verify evidence across modalities. Notably, this setting departs sharply from real-world visual search, where relevant evidence rarely appears as exact duplicates and is often distributed across diverse viewpoints, crops, resolutions, and contextual scenes. Our results indicate that many existing benchmarks contain a substantial proportion of images and metadata that can be located nearly exactly on the web, thereby overlooking the iterative localization, visual search, and entity-level cross-modal verification required in real-world scenarios.

Furthermore, as shown in the bottom-right of Fig. 1, the multi-hop component of most existing

benchmarks remains insufficiently challenging, *i.e.*, their deep-research setups are often overly simplistic, failing to effectively stress-test the increasingly strong deep-research capabilities of MLLMs.

3.3 Why VDR-Bench?

Our findings reveal a fundamental misalignment between existing benchmark protocols and the capabilities required for realistic multimodal deep-research systems. Many current benchmarks allow models to succeed through text-only shortcuts, language priors, or one-shot near-duplicate retrieval, rather than necessitating iterative visual search, fine-grained localization, and cross-modal evidence aggregation.

These empirical gaps directly motivate the design of VDR-Bench. By emphasizing visual-search–centric reasoning with the tailored designed challenging search multi-hop setting, mitigating perfect-retrieval bias, and evaluating agents under realistic multimodal search conditions, VDR-Bench offers a more faithful and rigorous testbed for assessing and advancing multimodal deep-research systems.

4 VDR-Bench

In Sec. 4.1, we present a detailed description of the data curation process for VDR-Bench, including a multi-stage pipeline and rigorous human experts review. In Sec. 4.2, we introduce two evaluation metrics specifically designed for this benchmark to assess the accuracy of deep-research and entity-level retrieval performance.

4.1 Data Curation Process

Our benchmark data is constructed through a strict vision-centric search pipeline and rigorous human experts review, beginning from raw images and culminating in multi-hop VQA instances. Throughout the process, we explicitly annotate retrieved entities, thereby providing each sample with grounded visual and textual evidence.

Step 0: Multi-Domain Image Pre-Filtering. We first collect images from multiple datasets spanning diverse domains and remove samples with insufficient resolution. We then employ Qwen3-VL-235B-A22B-Instruct as an image filtering model to further select the images that contain multiple entities and reflect realistic, visually rich scenes.

Step 1: Manual Cropping and Visual Search. Given an original image, annotators are instructed to extract local regions in the form of bounding boxes (*e.g.*, objects, logos, or individuals). Each cropped region is then used as a query image and submitted to an image search engine to produce visual search results for candidate entities.

Step 2: Visual Entity Extraction and Verification. We prompt Qwen3-VL-235B-A22B-Instruct with each cropped entity image and its corresponding search results to extract candidate entity names. The extracted entities are then filtered through two criteria: (1) we query multiple mainstream MLLMs with only the entity image to verify that the entity cannot be identified from the model's internal world knowledge alone; and (2) human experts further validate the correctness of each entity name.

Step 3: Seed VQA Generation from Visual Entities. For each verified visual entity, we prompt Gemini-2.5-Pro with the original image, the cropped entity image, and the entity name to generate simple seed VQA (*e.g.*, identifying the landmark, logo, or individual depicted in the image). Human experts then review these VQA to ensure unambiguous entity reference and answer uniqueness.

Step 4: Knowledge-Graph–Based Complexity Expansion. To address the overly direct and simplistic nature of VQA in existing benchmarks, we expand each seed VQA into multi-hop complex questions by linking visual entities to an external knowledge graph and performing random walks over related nodes.

We employ two complementary complexity expansion strategies: (1) **Answer Expansion**: starting from the known visual entity, we traverse the knowledge graph to identify a related external entity and extend the question along this new reasoning hop; (2) **Entity Expansion**: we replace explicit entity mentions in the question with indirect descriptive clues, requiring the solver to first resolve the reference before proceeding with reasoning.

A concrete example is provided in the Appendix D to further illustrate this process.

Step 5: Solvability and Quality Verification. All synthesized questions undergo a two-stage verification process. First, we prompt multiple MLLMs with the visual search evidence (cropped images and entity names) and knowledge-graph-expanded entities with their associated web context to verify that each answer is solvable. Second, human annotators further review each question to ensure that solving it genuinely requires visual entity recognition and that the question is valid and unambiguous.

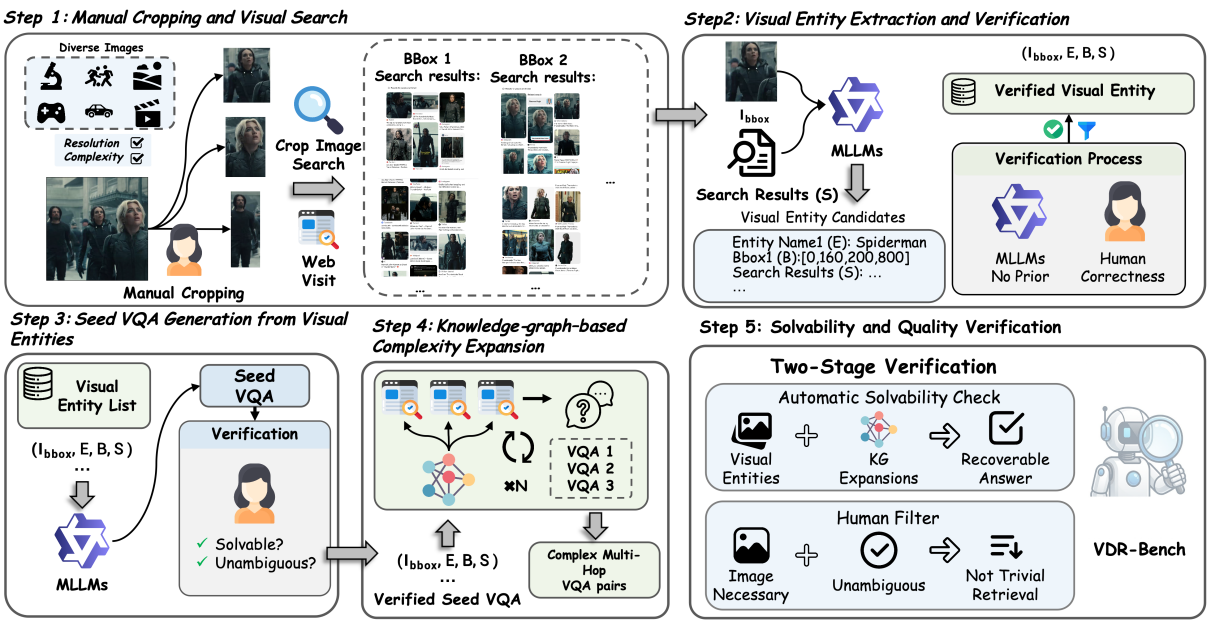

Figure 2: **VDR-Bench construction pipeline.** (1) Annotators extract local regions as bounding boxes and query an image search engine; (2) Candidate entity names are extracted from search results and verified via MLLMs and human experts; (3) Seed VQA pairs are generated from verified visual entities; (4) Questions are expanded into multi-hop form through Answer Expansion and Entity Expansion over a knowledge graph; (5) A two-stage check ensures each question is solvable, visually necessary, and unambiguous.

Further details of the above steps are provided in Appendix F.

4.2 Benchmark Composition and Metrics

Following the pipeline described above, we construct VDR-Bench, comprising 2,000 multi-hop VQA instances spanning 10 diverse visual domains. The benchmark covers a broad spectrum of visual complexity, entity density, and reasoning depth. Dataset composition and domain distribution are shown in Fig. 4.

We use two core metrics: answer accuracy and entity recall to comprehensively evaluate the correctness of the answer and the search engine hit rate at the entity level.

Answer Accuracy. We extract the agent's final answer from the trajectory last step and evaluate its correctness using Qwen3-VL-30B-A3B-Instruct as the judge model. The evaluation follows the judge prompt template from the Tongyi DeepResearch report (Team et al., 2025). See Appendix B.

Entity Recall. In addition to accuracy, we propose Entity Recall (ER) to evaluate whether a multimodal deep-research system successfully discovers the task-relevant entities. ER measures the alignment between the searched entities and the annotated entity set.

For each task i, let $\mathcal{E}_i = \{e_{i_1}, \ldots, e_{i_N}\}$ denote the annotated entity set, and let $\mathcal{S}_i = \bigcup_{t=1}^{T} s_{i_t}$ denote the set of entities extracted from the agent's search trajectory across T time steps. Specifically, ER is evaluated under an LLM-as-a-judge paradigm, which assesses the semantic equivalence between searched entities and annotated entities. Formally, the success of the i-th trajectory is defined as:

$$S_i = \mathbb{1}\left[\mathrm{LLM}_{\mathrm{eval}}\left(\mathcal{S}_i, \mathcal{E}_i\right) = 1\right], \qquad (1)$$

where the indicator function $\mathbb{1}[\cdot]$ yields 1 if the LLM judge confirms that the set of retrieved entities $\mathcal{S}_i$ sufficiently covers the annotated entity set $\mathcal{E}_i$ through semantic reasoning, and 0 otherwise. Unlike string matching, this approach recognizes semantic synonyms and respects the fact that an agent might follow several equally valid paths to reach the target entities.

5 Experiments

5.1 Experimental Setups

We evaluate a broad set of state-of-the-art MLLMs on VDR-Bench, including Gemini 2.5 Pro (Comanici et al., 2025), Gemini 3 Pro (Google DeepMind, 2026), GPT-5 (OpenAI., 2025), Claude-4-Sonnet (Anthropic, 2025a), Claude-4.5-Sonnet (Anthropic, 2025b), the Qwen3-VL series (Bai et al., 2025), the Qwen3.5 series (Team, 2026), and Kimi-K2.5 (Team et al., 2026). Full results across all 10 visual domains are reported

Table 2: **Answer accuracy (%) on VDR-Bench across 10 visual domains.** Models are evaluated under three settings: Direct Answer (no search), CIS+TS (cropped-image search + text search), and CIS+TS+MVF (with Multi-turn Visual Forcing). The best results are highlighted in **bold**, and the second-best results are underlined.

Model / Setting	People	Object	Arch.	Nature	Sci&Tech	Art&Music	Sports	Movie	Game	Other	Overall
Gemini 2.5 Pro											
Direct Answer	6.4	9.8	9.8	8.2	12.0	11.8	4.2	2.0	7.7	9.6	8.2
CIS+TS	14.9	15.7	27.5	12.2	24.0	17.6	12.5	10.2	1.9	25.0	16.2
CIS+TS+MVF	38.3	23.5	33.3	24.5	22.0	39.2	25.0	24.5	21.2	48.1	30.0
Gemini 3 Pro											
Direct Answer	14.6	2.0	14.0	17.8	15.1	9.6	4.3	8.8	7.5	15.0	10.7
CIS+TS	45.8	37.3	42.0	33.3	28.3	42.3	35.4	40.4	37.7	50.0	39.0
CIS+TS+MVF	41.7	51.0	50.0	35.4	37.7	48.1	37.5	45.6	39.6	**60.0**	44.4
GPT-5											
Direct Answer	4.4	9.8	11.7	12.3	10.0	7.8	8.4	8.2	3.8	13.5	9.5
CIS+TS	20.8	17.6	14.0	16.7	24.5	21.2	12.5	19.3	20.8	25.0	19.2
CIS+TS+MVF	23.4	25.5	23.5	20.4	18.0	27.5	22.9	30.6	30.8	42.3	26.6
Claude-4-Sonnet											
Direct Answer	2.1	3.9	7.8	6.2	10.0	7.8	2.2	0.0	3.8	11.5	5.6
CIS+TS	14.9	9.8	19.6	16.3	18.0	11.8	10.4	4.1	3.8	23.1	13.2
CIS+TS+MVF	12.5	17.6	24.0	35.4	15.1	26.9	16.7	12.3	23.1	24.4	20.6
Claude-4.5-Sonnet											
Direct Answer	14.6	2.0	14.0	17.8	15.1	9.6	4.3	8.8	7.5	15.0	10.7
CIS+TS	**60.4**	64.7	**74.0**	**77.1**	**71.7**	59.6	**68.8**	45.6	54.7	50.0	62.6
CIS+TS+MVF	58.3	**66.7**	62.0	75.0	69.8	**71.2**	**68.8**	**49.1**	**58.5**	57.5	**63.6**
Qwen3-VL-30B-A3B-Instruct											
Direct Answer	0.0	3.9	3.9	6.1	2.0	7.8	2.1	4.1	0.0	7.7	3.8
CIS+TS	17.0	19.6	17.6	16.3	20.0	5.9	14.6	10.2	5.8	44.2	17.2
CIS+TS+MVF	25.5	21.6	23.5	18.4	8.0	23.5	16.7	18.4	28.8	26.9	21.2
Qwen3-VL-235B-A22B-Instruct											
Direct Answer	6.2	3.9	10.0	22.9	7.5	13.5	6.2	3.5	7.5	7.5	8.8
CIS+TS	25.2	19.5	24.0	21.1	18.5	17.1	10.7	29.1	16.6	31.5	21.2
CIS+TS+MVF	25.0	23.5	30.0	31.2	30.2	28.8	20.8	22.8	30.2	32.5	27.4
Qwen3.5-35B-A3B											
Direct Answer	6.2	2.0	2.0	4.2	3.8	13.5	4.2	1.8	1.9	10.0	4.8
CIS+TS	19.4	16.6	23.1	25.0	21.1	19.9	28.5	21.8	11.1	27.5	21.4
CIS+TS+MVF	14.4	19.9	39.3	40.6	27.3	30.6	19.2	10.9	25.3	32.5	26.0
Qwen3.5-397B-A17B											
Direct Answer	10.4	3.9	12.0	10.4	9.4	15.4	8.3	5.3	0.0	7.5	8.2
CIS+TS	22.0	21.6	27.7	29.2	24.2	23.9	34.5	27.0	12.4	27.5	25.0
CIS+TS+MVF	32.8	26.9	28.0	33.1	35.2	34.4	24.4	34.1	18.1	35.0	30.2
Kimi-K2.5											
Direct Answer	16.7	2.0	4.0	18.2	16.4	9.6	12.8	17.5	5.7	27.5	9.8
CIS+TS	25.0	11.8	22.0	8.3	20.8	23.1	27.1	17.5	7.5	30.0	19.0
CIS+TS+MVF	20.8	15.7	30.0	33.3	30.2	28.8	27.1	28.1	11.3	37.5	26.0

in Tab. 2. To evaluate the contribution of external search and visual grounding, we compare each model under three inference settings:

(1) Direct Answer. The model answers using only the input image and question, without any search. This setting measures how much of VDR-Bench can be solved by parametric knowledge.

(2) CIS+TS. The model is equipped with cropped-image search (CIS) and text search (TS), allowing it to crop local regions as visual queries, retrieve candidate entities, and gather textual evidence for multi-hop reasoning.

(3) CIS+TS+MVF. Building on CIS+TS, we further enable Multi-turn Visual Forcing (MVF), which decouples visual and textual search to prevent models from falling into shallow or skipped visual search behavior under standard CIS+TS. Before text search, the model is required to perform at least 3 rounds of visual search with no fewer than 12 bounding boxes at varying scales, ensuring sufficient coverage and recall of relevant visual entities. The discovered entities then serve as anchors for text search and final reasoning. Details of MVF are provided in the Appendix C.

All models are evaluated under identical tool interfaces and constraints. We report answer accuracy as the primary metric and Entity Recall (ER) as a complementary diagnostic, measuring whether the model successfully retrieves task-relevant annotated entities in its search trajectory.

5.2 Main Results

Tab. 2 reports the performance of all evaluated models across the 10 visual domains of VDR-Bench.

VDR-Bench requires active search. Under the Direct Answer setting, all models achieve consistently low accuracy, even the strongest model (Claude-4.5-Sonnet) reaches only 10.7%. This confirms that VDR-Bench cannot be solved through parametric knowledge alone; instead, models must

identify visual entities, retrieve external evidence, and perform multi-hop reasoning.

Multimodal search brings substantial gains. Enabling CIS+TS yields significant improvements across all models. For example, Gemini 2.5 Pro improves from 8.2% to 16.2%, GPT-5 from 9.5% to 19.2%, Qwen3-VL-30B-A3B-Instruct from 3.8% to 17.2%, and Kimi-K2.5 from 9.8% to 19.0%, demonstrating that CIS+TS provide essential evidence for solving VDR-Bench.

Search capability matters more than parametric knowledge. Notably, models differ substantially in how they leverage search results for downstream reasoning. Qwen3-VL-30B-A3B-Instruct lags far behind Gemini 2.5 Pro under Direct Answer, yet surpasses it under CIS+TS. This suggests that improving vision deep-research agents through stronger pretraining alone is insufficient; the key to solving such tasks lies in encouraging effective and extensive use of search tools.

Best performance and remaining challenges. Among all models, Claude-4.5-Sonnet achieves the highest overall accuracy of 62.6% under CIS+TS, showing that strong multimodal agents can make substantial progress when effectively combining localized visual search with text-based evidence aggregation. However, even this best result leaves considerable room for improvement. Performance varies noticeably across domains: categories with visually distinctive entities (*e.g.*, architecture, art & music) benefit most from search, while domains involving ambiguous or visually similar entities (*e.g.*, people, sports, games) remain challenging due to the difficulty of fine-grained localization and disambiguation.

5.3 Entity Recall Analysis

To better understand model behavior, we analyze the relationship between ER and answer accuracy. As shown in Fig. 3, a clear positive association emerges: models with higher ER consistently achieve higher accuracy, indicating that entity discovery is a central bottleneck on VDR-Bench.

ER as a diagnostic signal. The positive correlation between ER and accuracy reveals that many failures occur before final reasoning: models frequently fail to retrieve the key entities required for answering. This observation highlights that improving entity-level visual search may be as important as improving final-stage reasoning for future multimodal deep-research systems. We believe VDR-Bench can serve as a powerful diagnostic testbed for driving progress in this direction.

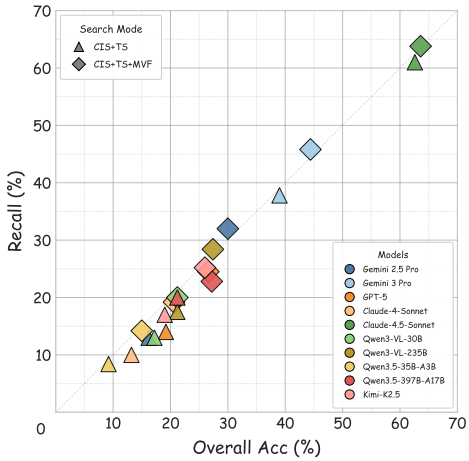

Figure 3: **Relationship between overall answer accuracy and entity-level recall on VDR-Bench.** Colors indicate model families; marker shapes distinguish CIS+TS from CIS+TS+MVF. Entity Recall strongly correlates with accuracy, and MVF consistently improves both metrics across models.

MVF improves both entity recall and accuracy. Compared to CIS+TS, enabling MVF shifts model performance toward both higher ER and higher accuracy. By enforcing multiple rounds of in-depth visual search, MVF prevents models from falling into shallow or skipped search behavior under standard CIS+TS. The improvement is especially pronounced for models with less effective baseline search: Gemini 2.5 Pro improves from 16.2% to 30.0% and Claude-4-Sonnet from 13.2% to 20.6%. These results confirm that explicitly encouraging multi-scale visual exploration helps models obtain more useful entity-level evidence before conducting textual reasoning.

6 Conclusion

We identify two critical deficiencies in existing Vision-DeepResearch benchmarks: reliance on text-only or prior-based shortcuts that bypass genuine visual verification, and overly idealized retrieval settings where whole-image search enables near-perfect matching. To address these issues, we propose VDR-Bench, a 2,000-instance benchmark curated through a vision-centric, multi-stage pipeline with rigorous expert review, and introduce Multi-turn Visual Forcing (MVF), a prompting workflow that encourages multi-scale visual exploration before textual reasoning. Experiments show that entity-level retrieval is a central bottleneck and that MVF yields substantial gains across models. Our results provide practical guidance for building more robust multimodal deep-research agents.

Limitations

While VDR-Bench provides a more realistic testbed for evaluating multimodal deep-research systems, several limitations remain. Our results depend on the specific search engines used, and different backends may affect model performance. Additionally, the current benchmark is constructed and evaluated exclusively in English, limiting its coverage of non-English visual search scenarios. Finally, MVF is a zero-shot prompting strategy bounded by the model's instruction-following and visual grounding capabilities; more advanced approaches such as reinforcement learning-based search policies remain promising future directions.

References

Anthropic. 2025a. Introducing claude 4. https://www.anthropic.com/news/claude-4.

Anthropic. 2025b. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, and 1 others. 2024b. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, and 1 others. 2025. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Mingyang Fu, Yuyang Peng, Benlin Liu, Yao Wan, and Dongping Chen. 2025. Livevqa: Live visual knowledge seeking. *arXiv preprint arXiv:2504.05288*.

Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Yida Zhao, Kuan Li, and 1 others. 2025. Webwatcher: Breaking new frontier of vision-language deep research agent. *arXiv preprint arXiv:2508.05748*.

Google DeepMind. 2026. Gemini 3 Pro model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf.

Jack Hong, Chenxiao Zhao, ChengLin Zhu, Weiheng Lu, Guohai Xu, and Xing Yu. 2025. Deepeyesv2: Toward agentic multimodal model. *arXiv preprint arXiv:2511.05271*.

Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Yao Hu, and Shaohui Lin. 2025. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Pengshuo Qiu, Pan Lu, Zehui Chen, Guanglu Song, Peng Gao, Yu Liu, and 1 others. Mmsearch: Unveiling the potential of large models as multi-modal search engines. In *The Thirteenth International Conference on Learning Representations*.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, and 1 others. 2025. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*.

Kartik Narayan, Yang Xu, Tian Cao, Kavya Nerella, Vishal M Patel, Navid Shiee, Peter Grasch, Chao Jia, Yinfei Yang, and Zhe Gan. 2025. Deepmmsearch-r1: Empowering multimodal llms in multimodal web search. *arXiv preprint arXiv:2510.12801*.

OpenAI. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, Yong Jiang, and 1 others. 2025. Webshaper: Agentically data synthesizing via information-seeking formalization. *arXiv preprint arXiv:2507.15061*.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

Qiuchen Wang, Ruixue Ding, Yu Zeng, Zehui Chen, Lin Chen, Shihang Wang, Pengjun Xie, Fei Huang, and Feng Zhao. 2025. Vrag-rl: Empower vision-perception-based rag for visually rich information understanding via iterative reasoning with reinforcement learning. *arXiv preprint arXiv:2505.22019*.

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, and 1 others. 2025a. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.

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

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

Yu Zeng, Wenxuan Huang, Shiting Huang, Xikun Bao, Yukun Qi, Yiming Zhao, Qiuchen Wang, Lin Chen, Zehui Chen, Huaian Chen, and 1 others. 2025a. Agentic jigsaw interaction learning for enhancing visual perception and reasoning in vision-language models. *arXiv preprint arXiv:2510.01304*.

Yu Zeng, Yukun Qi, Yiming Zhao, Xikun Bao, Lin Chen, Zehui Chen, Shiting Huang, Jie Zhao, and Feng Zhao. 2025b. Enhancing large vision-language models with ultra-detailed image caption generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 26703–26729.

Appendix

A Dataset Composition

VDR-Bench comprises 2,000 multi-hop VQA instances spanning 10 diverse visual domains: People, Object, Architecture, Nature, Sci & Tech, Art & Music, Sports, Movie, Game, and Others. Fig. 4 shows the full domain distribution along with representative question–answer examples.

B Eval Prompt Template

We evaluate answer accuracy using Qwen3-VL-30B-A3B-Instruct as the judge model. The structured evaluation prompt is shown in Fig. 5. The judge reads the full research report and determines whether it contains the correct answer, independent of format or position.

C Multi-turn Visual Forcing (MVF) Details

MVF is a two-phase structured prompting workflow. **In Phase 1 (Visual Exploration)**, the model is instructed to use `crop_and_search` to perform broad visual search over the input image at multiple scales, extracting local regions as bounding boxes and submitting them to an image search engine. The model iterates until sufficient visual entity information is gathered (model is required to perform at least 3 rounds of visual search with no fewer than 12 bounding boxes at varying scales). **In Phase 2 (Web Search & Synthesis)**, the model transitions to text-based evidence collection: it uses the discovered visual entities as anchors to issue text search queries and visit relevant webpages, then synthesizes all gathered evidence to produce a final answer.

The full system prompts for both phases are shown in Fig. 6 and Fig. 7.

D Knowledge-Graph Complexity Expansion Example

The following example illustrates how Answer Expansion and Entity Expansion progressively increase question difficulty from a simple seed question.

Seed Question: *"What brand is the football in the picture?"*

↓ **Answer Expansion** (link to the brand's founder)

Hop 1: *"What is the name of the founder of the football brand in the picture?"*

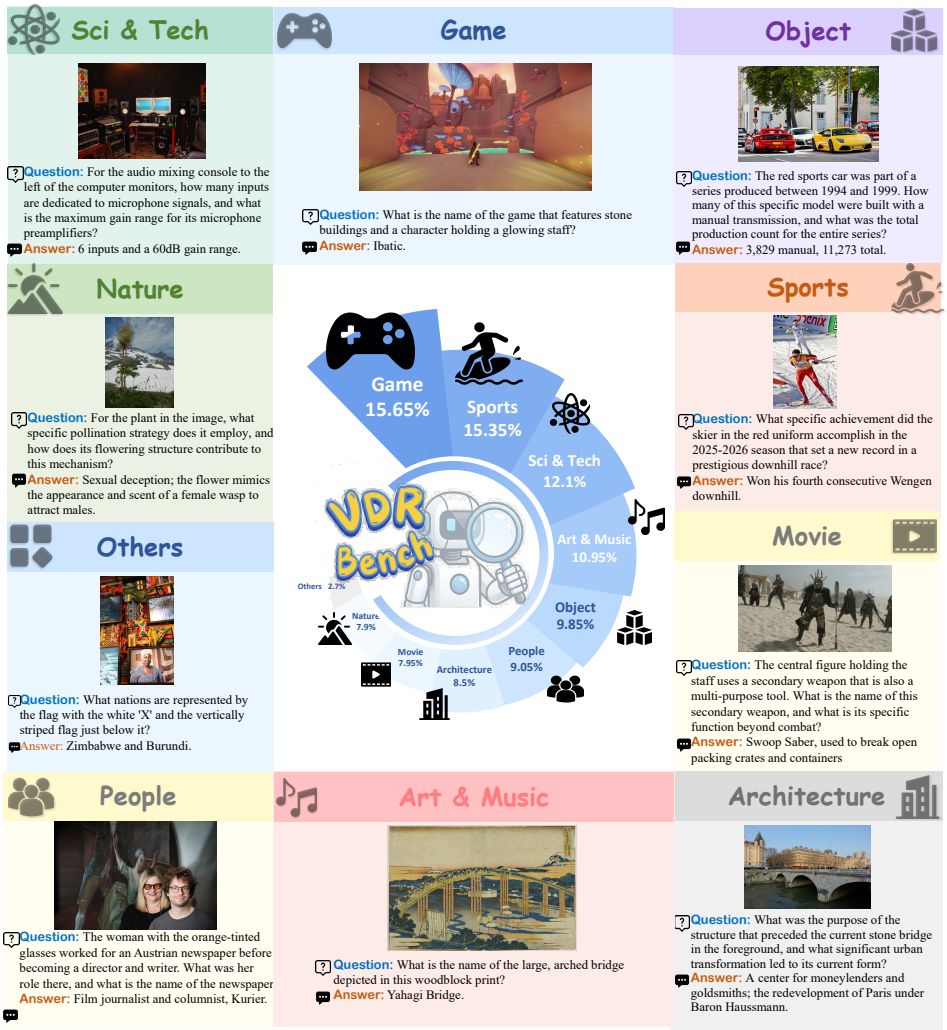

Figure 4: **Dataset composition and example instances from VDR-Bench.** The distribution across ten visual domains is shown, along with representative question–answer examples for each category.

↓ **Answer Expansion** (extend to a specific factual detail)

Hop 2: *"What did the founder of the football brand in the picture do in Wilmslow, Cheshire, England?"*

↓ **Entity Expansion** (replace "Wilmslow" with indirect clue)

Hop 3: *"There is a town in Cheshire, England, located 11 miles south of Manchester. What is the famous thing the founder of the football brand in the picture did in this town?"*

By combining both strategies, the resulting questions require multi-hop reasoning grounded in visual entity recognition, making them significantly more challenging and aligned with real-world visual deep-research scenarios.

E Annotation Details

E.1 Annotator Team.

Our team consists of 10 annotators, all with prior experience in multimodal data annotation. Before formal annotation, all annotators completed a structured training session covering annotation guidelines, entity taxonomy, and quality standards, and passed a qualification test to ensure consistent task understanding.

E.2 Workflow and Quality Control.

The pipeline operates under a dual-role workflow: each instance is first annotated by one annotator and then independently reviewed by a quality inspector. Instances flagged during inspection are returned for revision until they meet the required standards. The full annotation cycle spanned 14 days, during which the team processed approxi-

Prompt for **Deep Research** Answer Verification

[ROLE & TASK]
You are an impartial judge evaluating whether a deep research report contains the correct answer.
Task: Determine if the deep research report contains the correct answer anywhere in its content.

[EXECUTION RULES]

- **Thoroughness:** Read through the entire research report carefully.
- **Search Strategy:** Look for the correct answer anywhere in the report (it may be embedded in paragraphs, tables, or sections).
- **Consistency Check:** Check if the information in the report is consistent with the correct answer.
- **Format Independence:** The answer does NOT need to be in a specific format or labeled as "final answer".
- **Verdict Criteria:** Answer with "yes" if the report contains the correct answer, "no" if it doesn't or contradicts it.
- **Reasoning:** You must provide your reasoning to justify your final verdict.

[INPUT DATA]
Question: {question}
Correct Answer: {reference_answer}
Deep Research Report: {assistant_answer}

[OUTPUT FORMAT]
```
correct: [yes/no]
reasoning: [your explanation]
```

Figure 5: The structured evaluation prompt for determining if a deep-research report contains the correct answer.

mately 16,000 candidate instances. After multi-stage filtering and quality verification (Steps 1-5 in Sec. 4.1), 2,000 instances were retained in the final VDR-Bench, yielding an overall retention rate of approximately 12.5%.

F Data Curation Details

This section provides further implementation details and prompt templates for each step of the data curation pipeline described in Sec. 4.1.

F.1 Step 0: Multi-Domain Image Pre-Filtering.

We use Qwen3-VL-235B-A22B-Instruct to perform initial image filtering. The prompt instructs the model to assess whether the image contains at least 3–5 distinct, recognizable knowledge entities suitable as starting points for visual search. Images must also meet a minimum resolution of 512×512 pixels. Only images satisfying both criteria pass the pre-filtering stage.

F.2 Step 1: Manual Cropping and Visual Search.

Annotators follow a detailed guideline specifying target entity categories, including: *objects* (e.g., products, vehicles, instruments), *logos and brand marks*, *landmarks and architectural structures*, *individuals* (e.g., public figures, characters), *text and signs* (e.g., labels, posters), and *artworks* (e.g., paintings, sculptures, prints). For each image, annotators identify all visually salient entities, draw tight bounding boxes, and assign brief descriptive labels, focusing on entities whose identity would benefit from visual search verification.

F.3 Step 2: Visual Entity Extraction and Verification.

To verify entity correctness, annotators submit each cropped region to Google Reverse Image Search and cross-check the returned webpage content to confirm whether results consistently identify the same entity. This search-based verification ensures that entity names are grounded in retrievable web evidence rather than subjective judgment. Entities that cannot be reliably matched are discarded.

Prompt for Deep Research Assistant: **Phase 1 (Visual Search)**

[ROLE & CORE FUNCTION]
You are a deep research assistant. Your core function is to conduct thorough, multi-source investigations into any topic. You must handle both broad, open-domain inquiries and queries within specialized academic fields. For every request, synthesize information from credible, diverse sources to deliver a comprehensive, accurate, and objective response.

[ACTION DIRECTIVE]
First, you MUST use `crop_and_search` based on the given question to perform a broad search of the target image at **different scales** to obtain sufficient visual information.

[TOOLS]
You are provided with function signatures within `<tools></tools>` XML tags:

```
<tools>
{
  "type": "function",
  "function": {
    "name": "crop_and_search",
    "description": "Crop some important local regions from an image and search...",
    "parameters": {
      "type": "object",
      "properties": {
        "image_id": {"type": "string", "description": "..."},
        "bbox": {"type": "array", "description": "Bounding box coordinates [x1, y1, x2,
    y2]."},
        "goal": {"type": "string", "description": "The specific purpose of the visual
    search."}
      },
      "required": ["image_id", "bbox", "goal"]
    }
  }
}
</tools>
```

[EXECUTION RULES & OUTPUT FORMAT]

- Directly output `<tool_call>...</tool_call>`.
- Output only **one** function call or answer at a time. No further explanation is needed.
- For each function call, return a JSON object with function name and arguments:

```
<tool_call>
{
  "name": "crop_and_search",
  "arguments": {
    "image_id": "image_001.jpg",
    "bbox": [[120, 200, 480, 520]],
    "goal": "Identify the brand logo shown in the cropped region"
  }
}
</tool_call>
```

Strictly follow the instructions.

Figure 6: The structured system prompt for MVF (Phase 1: Visual Exploration).

Prompt for Deep Research Assistant: **Phase 2 (Web Search & Synthesis)**

[ROLE & STATUS]
You are a deep research assistant. Your core function is to conduct thorough, multi-source investigations into any topic... *(Role description identical to Phase 1)*
Status: You have now gathered sufficient visual information; next, use `visit` and `search` tools to get the answers.

[TOOLS]
You are provided with function signatures within `<tools></tools>` XML tags:

```
<tools>
{"type": "function", "function": {"name": "crop_and_search", ...}}
{
  "type": "function",
  "function": {
    "name": "search",
    "description": "Perform Google web searches then returns a string of the top search
     results.",
    "parameters": {
      "type": "object",
      "properties": {
        "query": {"type": "array", "items": {"type": "string"}, "description": "The list
    of search queries."}
      },
      "required": ["query"]
    }
  }
}
{
  "type": "function",
  "function": {
    "name": "visit",
    "description": "Visit webpage(s) and return the summary of the content.",
    "parameters": {
      "type": "object",
      "properties": {
        "url": {"type": "array", "description": "The URL(s) of the webpage(s) to visit
    ."},
        "goal": {"type": "string", "description": "The specific information goal for
    visiting webpage(s)."}
      },
      "required": ["url", "goal"]
    }
  }
}
</tools>
```

[EXECUTION RULES & OUTPUT FORMAT]

- Directly output `<tool_call>...</tool_call>` to invoke the tool.
- Output only **one** function call or answer at a time. No further explanation is needed.

Example of a correct call:

```
<tool_call>
{
  "name": "search",
  "arguments": {
    "query": ["Transformer model first introduced"]
  }
}
</tool_call>
```

Strictly follow the instructions.

Figure 7: The structured system prompt for the MVF (Phase 2: Web Search & Synthesis).

F.4 Step 3: Seed VQA Generation.

We prompt Gemini-2.5-Pro with the original image, the cropped entity image, and the verified entity name to generate simple, single-hop VQA pairs whose answers directly depend on recognizing the visual entity (e.g., *"What is the name of the landmark shown in the image?"*). The full prompt template is shown in Fig. 8.

F.5 Step 4: Knowledge-Graph–Based Complexity Expansion.

For each visual entity, we retrieve its associated webpage content and linked knowledge-graph entries, then prompt Qwen3-VL-235B-A22B-Instruc with the entity name, webpage context, and seed question to rewrite it into a multi-hop form via Answer Expansion or Entity Expansion. The model is instructed to (1) identify related factual details from the web context to extend the reasoning chain, and (2) replace explicit entity mentions with indirect descriptive clues to further increase difficulty. The full prompt template is shown in Fig. 9.

F.6 Step 5: Solvability and Quality Verification.

In the automatic stage, we provide multiple MLLMs with the full entity trajectory (cropped images, entity names, knowledge-graph expansions, and associated webpage content) and prompt them to answer the synthesized question. If the models arrive at the correct answer using only the provided evidence chain, the question is deemed solvable. The verification prompt is shown in Fig. 10. In the human stage, annotators receive the visual and textual entity list along with the step-by-step expansion question trace. They verify two criteria: (1) the question *genuinely requires* the visual entity (removing the image would make it unanswerable); and (2) the reasoning path from visual entity to final answer is *clear and unambiguous*. Questions failing either criterion are removed.

Prompt for **Seed VQA Generation** (Step 3)

[ROLE & TASK]
You are a visual–language reasoning assistant. You will be given: (1) selected entities from an image, and (2) external knowledge retrieved for each entity, which may include their real names. Your task is to generate **one complex, deep-reasoning question** that connects these entities.

[REQUIREMENTS]

- **No Name Leakage:** Do NOT use or reveal any real entity names. Use natural visual descriptions instead (e.g., "the person holding...", "the structure in the background").
- **Cross-Modal Dependency:** The question must require *both* visual information from the image *and* the provided external knowledge. It must be unanswerable using only text or only the image.
- **Depth:** The question must be specific, challenging, and require multi-step reasoning.
- **Uniqueness:** The answer should be unique or nearly unique given the combined image and knowledge context.

[FORBIDDEN QUESTION TYPES]

- Pure text-lookup: e.g., "In what year was the Eiffel Tower built?" (answerable without the image)
- Pure perception: e.g., "What color is the rabbit in the picture?" (answerable without external knowledge)

[INPUT]
Selected Entities & Knowledge: {context}

[OUTPUT FORMAT]

```
{
"question": "A deep reasoning question without revealing entity names",
"ground_truth": "The correct answer using both image and knowledge",
"reasoning": "How the question depends on both visual and textual cues"
}
```

Figure 8: The prompt template for generating seed VQA pairs (Step 3). The model receives visual entities and their retrieved knowledge, and generates a cross-modal reasoning question that avoids name leakage.

Prompt for **Knowledge-Graph Complexity Expansion** (Step 4)

[ROLE & TASK]
You are given: (1) an image, (2) a Target Entity (the focus), (3) the Original Question previously asked about this image, and (4) search results about the Target Entity. Your task is to create a new, complex QA pair that evolves from the Original Question by digging deeper into the Target Entity.

[QUESTION RULES]

- **Contextual Consistency:** The new question must follow the semantic logic of the Original Question but probe deeper into the Target Entity.
- **Visual Start:** Begin with visual referents (e.g., "This person", "The building in the background") to refer to the Target Entity. Do NOT use its real name.
- **Knowledge-Dependent:** The question must require the provided search results to answer; common sense alone must not suffice.
- **No Leakage:** Avoid names, dates, or affiliations in the question that could create a one-to-one mapping to the entity. Such details should appear only in the answer.
- **Complexity:** Focus on specialized, non-obvious facts (e.g., historical background, specific achievements, technical specs). At most two sub-questions are allowed.
- **Open-Ended:** No yes/no questions. The question must have a clear, specific answer rather than inviting opinions.

[ANSWER RULES]

- **Accuracy:** Base the answer strictly on the provided search context.
- **Specificity:** Use the entity's real name and concrete data (dates, names, figures) in the answer. Keep it concise.

[INPUT]
Image: {image} **Target Entity:** {entity}
Original Question: {original_question}
Search Results: {search_context}

[OUTPUT FORMAT]
```
{
"complex_question": "...",
"complex_answer": "..."
}
```

Figure 9: The prompt template for knowledge-graph–based complexity expansion (Step 4). The model rewrites a seed question into a deeper, multi-hop form grounded in retrieved web knowledge.

Prompt for **Solvability Verification** (Step 5)

[ROLE & TASK]
You are a verification assistant. You will be given a complex question about an image, along with the full evidence chain used to construct it. Your task is to determine whether the question is **answerable** using only the provided evidence.

[PROVIDED EVIDENCE]

- **Image:** The original image containing the target visual entities.
- **Cropped Regions:** Bounding-box crops of each annotated entity.
- **Entity Names:** The verified identity of each cropped entity.
- **Search Results:** Web content retrieved via cropped-image search for each entity.
- **KG Expansion:** Knowledge-graph–expanded entities and their associated webpage content.

[VERIFICATION RULES]

- **Evidence Sufficiency:** Determine whether the correct answer can be derived by combining the visual entities, search results, and KG-expanded knowledge. No external information beyond what is provided should be needed.
- **Reasoning Validity:** Verify that there exists a clear, step-by-step reasoning path from the visual entity to the final answer.
- **Answer Uniqueness:** Confirm that the answer is unique and unambiguous given the provided evidence.

[INPUT]
Image: {image} **Cropped Entities:** {crops}
Entity Names: {entity_names}
Search Results: {search_results}
KG Expansion: {kg_context}
Question: {question}
Expected Answer: {ground_truth}

[OUTPUT FORMAT]
```
solvable: [yes/no]
reasoning: [Step-by-step explanation of how the answer
           can/cannot be derived from the evidence]
```

Figure 10: The prompt template for automatic solvability verification (Step 5). The model receives the full evidence chain and determines whether the question is answerable without external information.