NOTEVQA: BENCHMARKING VLMS ON REAL-LIFE QUESTIONS FROM HUMAN COMMUNITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) increasingly power consumer-facing AI search, yet evaluating them on the diversity of everyday visual questions remains challenging. Existing benchmarks often target predefined capabilities, such as multi-hop retrieval or long-form synthesis, whereas users ask photo-grounded questions spanning a long tail of everyday scenarios. Despite advances in VLMs, users continue to turn to other people for help with everyday visual questions on image-sharing platform. Motivated by this behaviour, we curate NoteVQA from these questions, yielding 252 items across 12 topical categories and 7 user intents. Each item includes a concise reference distilled from expert community responses and a human-audited interleaved reference answer that combines textual explanations with supporting visual evidence. We evaluate short-answer correctness and interleaved-answer quality. To support the latter, we introduce AgenticInterleave, a single-agent ReAct framework for retrieval-supported answer generation, together with IVR-12, a 12-dimensional rubric for assessing the content, presentation, and image quality of interleaved references and model outputs. Across 9 frontier VLMs, the highest short-answer accuracy is 52.8%, while adding agentic search to Qwen3.5-397B-A17B improves accuracy by only 2.0%. For interleaved answers, the same model running AgenticInterleave scores 3.52 under IVR-12, compared with 4.65 for the human-audited references, with the largest gap in content quality. These results highlight the challenges that everyday visual questions pose for current VLMs in both answer accuracy and the quality of visually grounded explanations.

1 INTRODUCTION

Vision-language models (VLMs) are rapidly becoming a new interface for information seeking, allowing users to ask questions directly about the visual world around them. Yet many existing benchmarks follow a capability-first paradigm: researchers define a target capability, such as compositional visual reasoning, knowledge-based question answering, or multimodal retrieval, and then construct questions to evaluate it [15, 13, 17, 4, 11]. While methodologically sound, this approach can produce queries that differ substantially from those users submit in practice. AI search systems and those embedded in user-generated content (UGC) platforms must address diverse everyday information needs. In image-based communities, users often pair a photo with a brief, naturally phrased question spanning the long tail of everyday scenarios. Strong performance on researcher-authored benchmarks may therefore not reliably translate into usefulness for real users.

We introduce NoteVQA, a benchmark built from visual questions posted by users on a user-generated image community. In the posts collected for NoteVQA, users share photos, ask questions, and seek answers from other community members. These questions provide a concrete setting for examining how well current VLMs address real-world visual information needs. Prior work identifies fine-grained entity recognition and knowledge-intensive visual questions as challenges for VLMs [**?**], motivating our focus on visual identification and local or niche knowledge. We also cover subjective judgment and authenticity checking to capture a broader range of everyday needs. Studies of social question answering highlight trust, personalised advice, and subjective recommendations as reasons for asking other people [18], underscoring the importance of answer usefulness alongside factual accuracy. Community question-answering research further associates topic-specific expertise with higher answer ratings in factual categories [1]. We therefore draw on community replies as a source of candidate answers and refine them through expert review, as illustrated in Figure 1(a).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Beyond the source of the questions, the answer format also matters. Benchmarks such as SimpleVQA [6], FVQA [25], MMSearch [14], and BrowseComp-VL (BC-VL) [9] use short reference answers, with median lengths of only a few words. Short-answer accuracy measures whether a model reaches the correct answer, but does not directly assess the quality of its explanation or supporting visual evidence. Community questions often call for more than identification: users also seek explanations, actionable guidance, and help distinguishing visually similar objects. To evaluate how well models meet these needs, NoteVQA pairs short references distilled from community responses with human-audited interleaved reference answers that combine textual explanations with supporting images. These two reference formats enable complementary evaluations of answer correctness and visually grounded explanation quality on the same questions.

Evaluating interleaved answers requires both a generation framework that can retrieve and incorporate visual evidence and criteria that assess how effectively the resulting text and images answer the question. We provide AgenticInterleave, a single-agent ReAct [26] framework that generates retrieval-supported interleaved answers with traceable image tags. We also introduce IVR-12, a 12-dimensional rubric covering content, presentation, and image quality. Applying the same criteria to human-audited references and model outputs provides a common scale for examining where generated answers succeed and fall short.

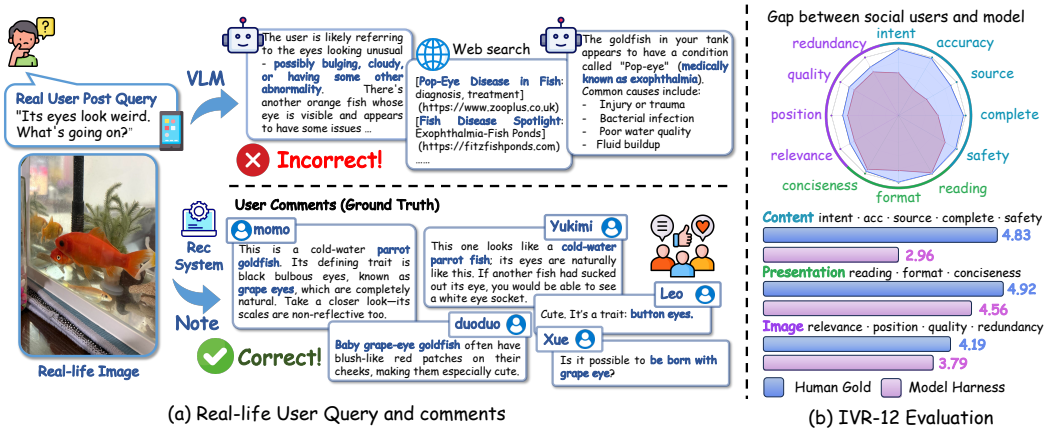


Figure 1: Overview of NoteVQA. (a) A real-user visual query challenging for a VLM with web retrieval but correctly answered by community experts. (b) IVR-12 comparison of the human-audited gold reference and AgenticInterleave output across content, presentation, and image quality.

Our main contributions are as follows.

- **A community-sourced visual QA benchmark with two reference formats.** We construct NoteVQA, comprising 252 real-world questions across 12 topical categories and 7 user intents. Each question is paired with an expert-distilled short reference and a human-audited interleaved reference, supporting evaluation of both answer correctness and visually grounded explanation quality.

- **A framework and rubric for interleaved answer evaluation.** We provide AgenticInterleave for generating retrieval-supported interleaved answers with traceable image tags, together with IVR-12 for assessing content, presentation, and image quality under the same criteria for references and model outputs.

- **An empirical analysis of accuracy, retrieval, and image use.** We evaluate 9 frontier VLMs, with the strongest achieving only 52.8% short-answer accuracy. Agentic retrieval improves Qwen3.5-397B-A17B by just 2.0 percentage points, while its interleaved answers show a substantial content-quality gap relative to the human-audited references. A controlled image-ablation study further quantifies the contribution of embedded visual evidence to answering image-critical probes.

2 Comparison with Existing Multimodal QA Benchmarks

We compare NoteVQA with 10 existing open multimodal QA benchmarks. The comparison considers three complementary perspectives: (1) query characteristics and domain coverage, (2) gains from agentic retrieval, and (3) reference-answer formats and evaluation scope. For query characteristics, we report mean query length and estimated reasoning depth. Query length provides a simple indicator of prompt complexity, while the mean number of reasoning hops, estimated by an LLM judge, approximates the depth of retrieval and reasoning required.

Table 1: Comparison of open multimodal VQA benchmarks with our proposed NoteVQA across key statistics.

Benchmark	Num.	Q. len.	Hops	Domain focus
SimpleVQA [6]	1013	10.2	1.52	Generic perception
FVQA [25]	1800	11.5	2.06	Factual single-entity QA
InfoSeek [5]	2000	8.2	1.95	KB entity attributes
MMSearch [14]	171	14.2	2.50	Web-search QA
LiveVQA [8]	300	52.1	2.52	News timeliness, MCQ
MMSearch+ [22]	222	12.0	2.12	arXiv multi-image QA
MMBC [16]	130	56.6	3.72	Web-browsing QA
BC-VL L1 [9]	199	20.0	3.27	Cross-modal multi-hop
BC-VL L2 [9]	200	74.0	3.88	Fuzzified long queries
VDR [28]	500	25.9	3.05	Fine-grained perception QA
NoteVQA	252	22.7	2.48	Consumer life

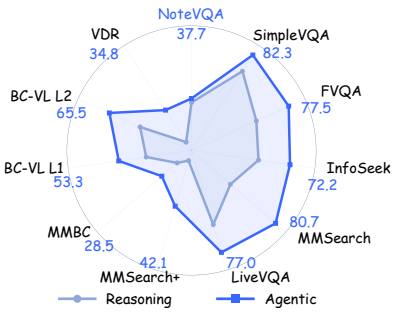


Figure 2: Accuracy comparison between reasoning and agentic settings.

(1) Query Characteristics and Domain Coverage. Table 1 compares the query length, estimated reasoning depth, and primary focus of the benchmark settings. BrowseComp-VL L2 and MM-BrowseComp have mean query lengths of 56 to 74 words and estimated reasoning depths of 3.7 to 3.9 hops, reflecting tasks with extended descriptions and explicit reasoning constraints. SimpleVQA, InfoSeek, and MMSearch have shorter queries, averaging 8 to 14 words and 1.5 to 2.5 estimated hops, and primarily assess generic visual perception, entity attributes, or web-search question answering. NoteVQA averages 22.7 words and 2.48 estimated hops, placing it between short factual questions and heavily specified multi-step tasks. Its questions are distinguished by their origin in an online user community and their focus on everyday visual information needs.

(2) Retrieval Gains across Benchmarks. Figure 2 compares Qwen3.5-397B-A17B [20] under two benchmark configurations. The single-pass setting generates answers directly from query and image without retrieval. The agentic setting runs the same model in a single-agent ReAct loop with web, page and image search. Agentic retrieval lifts accuracy 14–43 points across ten testbeds. On NoteVQA, accuracy rises modestly from 35.7% to 37.7% (Table 2), a gain of 2.0 points. This contrast demonstrates NoteVQA is a challenging setting where retrieval brings limited benefit. Section 4 performs instance-level analysis to identify successful, ineffective and error-inducing retrieval cases.

(3) Reference-Answer Formats and Evaluation Scope. Existing benchmarks judge final answer correctness via brief references or multiple-choice labels. Such formats verify target answers yet cannot assess full explanations grounded in visual evidence. Community queries demand richer outputs combining identification, practical advice, visual comparisons to tell similar objects apart. NoteVQA supplies short references and human-reviewed interleaved answers, enabling complementary measurement of answer accuracy and visually grounded explanation quality on identical questions. Three traits differentiate NoteVQA: real community-sourced everyday visual questions, modest gains with tested agentic retrieval, dual evaluation for short and interleaved responses. The next section details question curation, refinement, construction of two reference formats.

3 NoteVQA

Three principles govern NoteVQA. First, questions derive from user-generated image community posts. Second, a vision-necessity gate removes text-only samples. Third, references merge community knowledge: expert comments form references, model interleaved drafts receive annotator checks.

3.1 Data Curation Pipeline

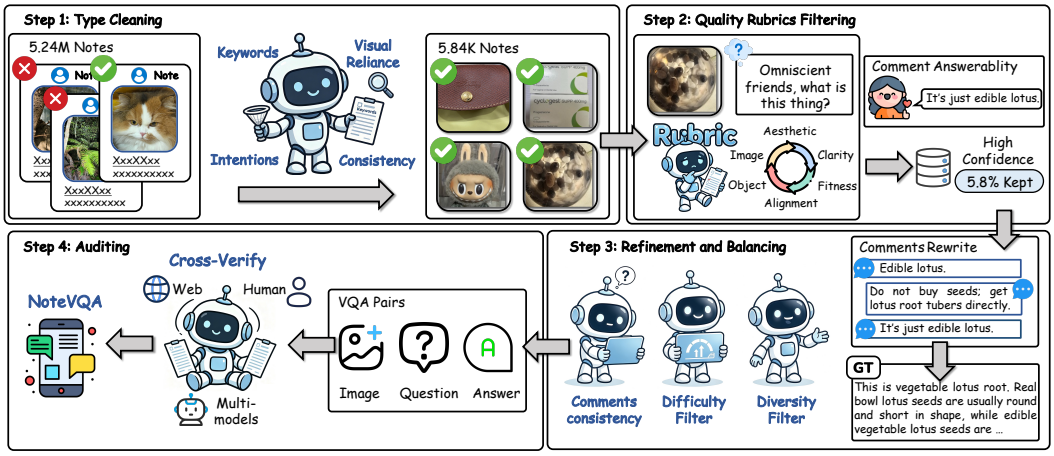


Figure 3: Data Curation Pipeline of NoteVQA. Raw notes flow through Step 1: Type cleaning to yield vision necessities, Step 2: Quality filtering to obtain high-confidence samples, Step 3: Refinement and balancing to produce the refined item set, and Step 4: Auditing to produce the final benchmark.

Step 1: Type Cleaning. We first apply keyword and length filters to retain image-bearing UGC posts with sufficient textual content, then use a VLM [20] to identify potential VQA candidates. We further remove items answerable from text alone and verify image–text consistency with a VLM, reducing 5.24M raw notes to 5.84K candidates.

Step 2: Quality Rubric Filtering. Each candidate is screened by a VLM aesthetic filter and an interleaved-suitability filter requiring the answer to cover at least three substantive aspects. We then apply a five-axis rubric covering question clarity, image quality, image–text alignment, interleaved fit, and answer objectivity, followed by per-tag top-K selection. A final VLM check verifies whether the comments provide a plausible answer, leaving 338 high-confidence samples.

Step 3: Refinement and Balancing. We standardise query and answer styles and rewrite queries to remove answer-revealing domain terms while preserving the original intent. We then run Qwen3.5-397B-A17B ten times per item and use response consistency as a difficulty probe. Samples are further pruned and balanced across category and intent groups, yielding a final set of 252 items.

Step 4: Auditing. We translate the refined items into English, cross-check the translations with multiple models, and have human reviewers proofread them for correctness and fluency. Each item is then assigned to its corresponding category and intent cell as a final consistency check.

3.2 Human-Verified Interleaved Reference Construction

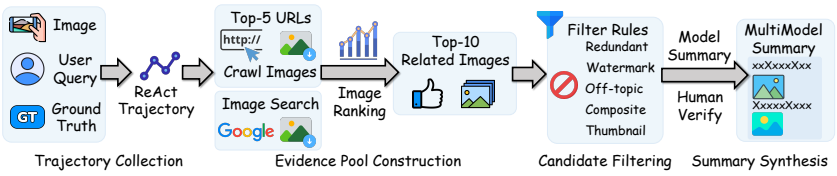


Figure 4: Four-stage pipeline for constructing interleaved multimodal summaries.

We implement these principles through the four-step curation pipeline shown in Figure 3.

As shown in Figure 4, we construct a human-verified interleaved reference for each NoteVQA item through a four-stage pipeline with human audit that combines agentic retrieval, multimodal ranking, and human verification to select informative images while retaining their source links.

Stage 1: Trajectory Collection. Multimodal Qwen3.5-397B-A17B agent runs a ReAct loop (web search, reverse image search, page visits), guided by the reference toward the intended answer; the trajectory and retrieved sources form the evidence record.

Stage 2: Evidence-Pool Construction. The candidate pool merges two sources: LLM-scored URLs (relevance, informativeness, directness, uniqueness, authoritativeness; top-5 per item; images crawled under minimum-size constraints, less equal than 20 per URL) and reverse-image-search images. Qwen3.5-397B-A17B then scores each image on visual relevance, informativeness, clarity, answer support, and necessity, keeping the top-10; named entities missing from the pool trigger forward image search for supplementary evidence.

Stage 3: Candidate Filtering. Rule-based and multimodal filtering removes redundant, composite, watermarked, or video-thumbnail images and off-topic content, then deduplicates, reducing noisy or misleading evidence while preserving complementary images.

Stage 4: Summary Synthesis. Qwen3.5-397B-A17B combines query image, short reference, trajectory evidence, filtered images into an interleaved reference, each embedded image indexed to its source. After multimodal consistency checks, PhD annotators label items *accept*, *minor fix*, or *reject*, correcting content and placement when necessary; only references passing audit enter benchmark. Benchmark contains 252 items across 12 categories and 7 intents. Queries average 22.7 words; short references and interleaved summaries average 107.5 and 219.4 words, with 378 embedded images total (1.5 per item). Figure 5 shows category distribution and representative queries.

3.3 Generating and Evaluating Interleaved Multimodal Answers

We provide two complementary components for the interleaved-answer track: **AgenticInterleave** generates answers by retrieving and integrating textual and visual evidence, while **IVR-12** (*Interleaved Visual-answer Rubric*) evaluates human-audited references and model outputs on a common scale.

AgenticInterleave. Given a query–image pair, AgenticInterleave follows a single-agent ReAct [26] loop with four retrieval tools: *text search*, *page visit*, *reverse image search*, and *forward image search*. At each turn, the agent uses the input and accumulated tool observations to decide whether to retrieve further evidence or produce the final interleaved answer. The loop ends when an answer is generated or repeated behaviour is detected. In our experiments, the agent uses Qwen3.5-397B-A17B as its backbone and receives only the query and image; the short reference is used separately for reference construction and content evaluation.

The system prompt governs image retrieval and placement via SUBJECT/BACKGROUND entity protocol. SUBJECT entities denote core objects/concepts of the query, BACKGROUND entities supply contextual supporting information. To visualize a SUBJECT entity, the agent fetches its image via forward image search, picking candidates from retrieved results with no dedicated image-selection model. Image tags are appended after entity descriptions. Each tag stores the search query and candidate index, validated against tool outputs by a post-hoc checker to confirm retrieval provenance.

IVR-12: Interleaved Answer Evaluation. We assess answer correctness and interleaved-answer quality separately. Short-answer accuracy is the share of responses marked correct by a SimpleQA-style judge. IVR-12 complements this metric by evaluating interleaved answers' content, presentation, and images, applying identical criteria to human-audited references and AgenticInterleave outputs.

IVR-12 has 12 dimensions scored 1–5. The five *content* dimensions evaluate intent understanding, objective accuracy, source fidelity, completeness, neutrality, and safety. The three *presentation* dimensions evaluate reading experience, formatting, highlighting, and conciseness. The four *image* dimensions evaluate relevance, necessity/placement, visual quality, and non-redundancy. Three binary gates check timeliness, safety redlines, and refusal correctness. Let $\bar{C}$, $\bar{P}$, $\bar{I}$ denote category-wise mean scores. The overall score is

$$S_{\text{IVR-12}} = \begin{cases} 0.50\,\bar{C} + 0.20\,\bar{P} + 0.30\,\bar{I}, & \text{if all three gates pass,} \\ 0, & \text{otherwise.} \end{cases}$$

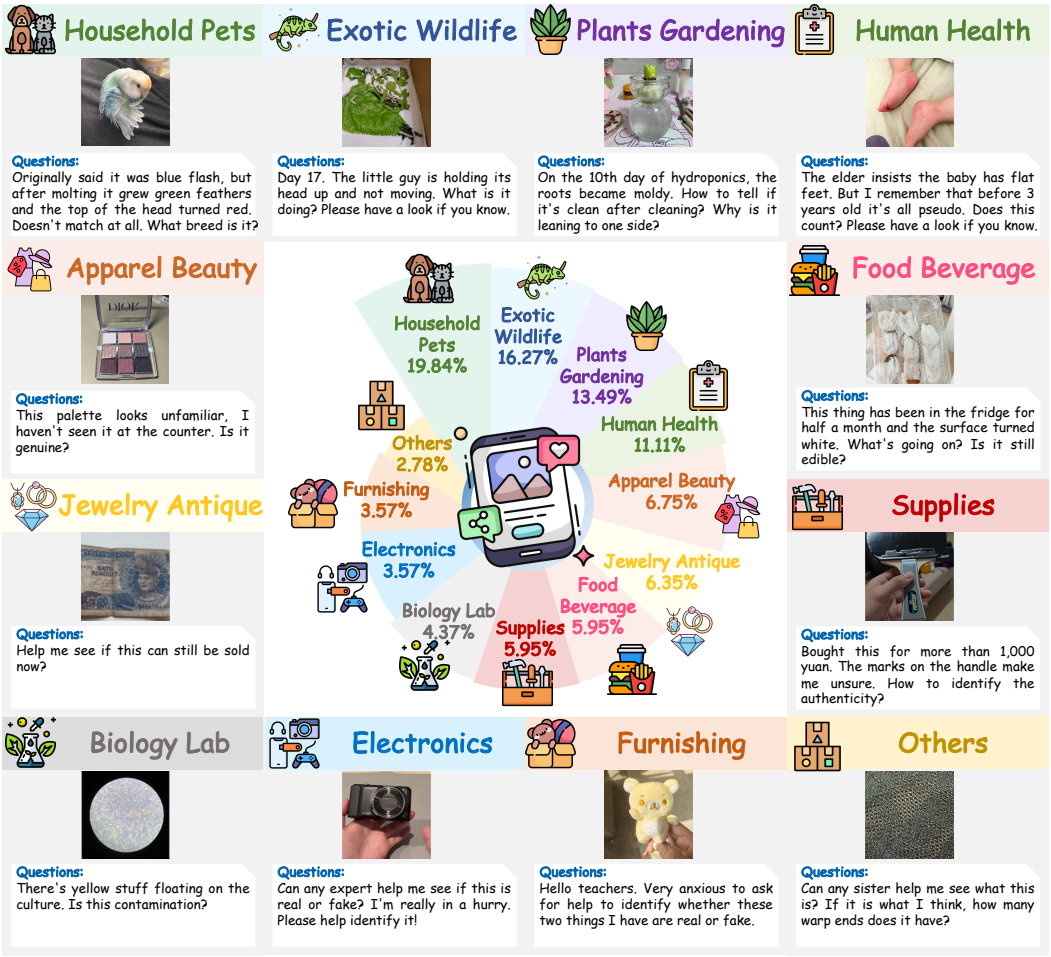


Figure 5: Dataset composition and representative examples of NoteVQA. The figure shows the distribution across 12 topical categories with representative real-user query instances for each category.

Thus, answers that pass all gates receive a weighted score, while failure on any gate yields zero.

We use Qwen3.5-397B-A17B for correctness judgment and IVR-12 evaluation. In IVR-12, answer text sets content and presentation scores, using human-vetted short references for content scoring. Image evaluation uses query and embedded images; gate checks use query image and answer text.

4 Experiments

We evaluate short-answer accuracy using nine models in thinking mode: Gemini-3.1-Pro [7], Doubao-Seed-2.0-Pro [21], GLM-5.3-Flash [10], Kimi-K2.6 [23], Kimi-K3 [24], Qwen3.8-Flash-Next [19], Claude Opus 4.7 [2], Claude Sonnet 4.6 [3] and Qwen3.5-397B-A17B [20]. For interleaved-answer quality, we use IVR-12 to compare outputs of Qwen3.5-397B-A17B running AgenticInterleave with human-audited references. Qwen3.5-397B-A17B is the default backbone for AgenticInterleave. It uses four retrieval tools described in Section 3.3, with Serper as the backend and a maximum of 10 ReAct rounds. The same model also serves as the IVR-12 judge.

4.1 Main Results

Table 2 reports short-answer accuracy overall and by category. Gemini-3.1-Pro achieves the highest overall accuracy at 52.8%. Adding agentic retrieval to Qwen3.5-397B-A17B raises its accuracy from 35.7% to 37.7%, a gain of 2.0%.

Table 2: NoteVQA short-answer accuracy (%) using thinking mode. ‡ denotes agentic retrieval.

Model	Overall	Pet	Exo	Pln	Hlt	Fas	Jwl	App	Fud	Bio	CE	Hom	Oth
Gemini-3.1-Pro	52.8	50.0	46.3	61.8	64.3	58.8	37.5	60.0	60.0	45.5	55.6	55.6	14.3
Doubao-seed-2.0-pro	49.6	48.0	53.7	47.1	64.3	41.2	18.8	73.3	53.3	45.5	55.6	55.6	14.3
GLM-5.3-Flash	48.0	46.0	48.8	52.9	53.6	23.5	43.8	53.3	53.3	45.5	55.6	44.4	57.1
Kimi-K2.6	47.6	42.0	43.9	50.0	57.1	52.9	56.2	60.0	46.7	36.4	44.4	44.4	28.6
Qwen3.8-Flash-Next	44.0	42.0	46.3	58.8	50.0	29.4	43.8	46.7	33.3	45.5	33.3	22.2	42.9
Claude-Opus-4.7	43.7	42.0	41.5	61.8	50.0	41.2	31.2	60.0	46.7	36.4	11.1	22.2	28.6
Kimi-K3	43.3	38.0	41.5	52.9	57.1	41.2	31.2	60.0	46.7	36.4	22.2	44.4	14.3
Claude-Sonnet-4.6	35.3	22.0	41.5	35.3	60.7	23.5	12.5	40.0	33.3	45.5	33.3	33.3	57.1
Qwen3.5-397B-A17B	35.7	28.0	31.7	38.2	42.9	47.1	25.0	53.3	40.0	27.3	33.3	33.3	42.9
Qwen3.5-397B-A17B‡	37.7	22.0	34.1	47.1	50.0	35.3	37.5	46.7	53.3	36.4	44.4	33.3	28.6

Categories: Pet, HomePet; Exo, Exotic and Wild; Pln, Plant and Garden; Hlt, Human Health; Fas, Fashion and Beauty; Jwl, Jewellery and Antique; App, Appliance, Kitchen, and Car; Fud, Food and Beverage; Bio, BioLab; CE, Consumer Electronics; Hom, Home; Oth, Other.

Table 3 compares interleaved-answer quality. The references score 4.65 under IVR-12, compared with 3.52 for AgenticInterleave, yielding a 1.13-point gap. The references achieve higher scores across all 12 dimensions. Content exhibits the largest discrepancy, with references averaging 4.82 and AgenticInterleave 2.95. Accuracy carries the widest gap at 2.68 points, followed by source fidelity at 2.07, completeness at 1.90, and intent understanding at 1.76. Presentation and Image show smaller deviations from reference values, with mean gaps of 0.37 and 0.40 respectively. Formatting produces the largest gap within this pair at 0.86. Thus, the primary shortfall under IVR-12 lies in answer content.

Table 3: Mean IVR-12 scores for human-audited references and AgenticInterleave outputs.

System	$S_{\mathrm{IVR-12}}$	Content					Presentation			Image			
		Int	Acc	Src	Cmp	Sfy	Rd	Fmt	Cnc	Rel	Pos	Qty	NR
Human-audited gold	4.65	4.94	4.90	4.48	4.84	4.92	5.00	4.96	4.81	4.17	4.16	4.24	4.18
AgenticInterleave	3.52	3.18	2.22	2.41	2.94	4.00	4.94	4.10	4.63	3.79	3.79	3.81	3.76

Dimensions: Int, intent understanding; Acc, accuracy; Src, source fidelity; Cmp, completeness; Sfy, neutrality and safety; Rd, reading experience; Fmt, formatting; Cnc, conciseness; Rel, image relevance; Pos, necessity and position; Qty, image quality; NR, non-redundancy.

4.2 Analysis of Agentic Search Gains

To understand the limited aggregate gain from agentic retrieval, we group all 252 items by whether the reasoning-only Qwen3.5-397B-A17B baseline and its agentic counterpart answer correctly. Table 4 reports the four groups and their tool usage. Retrieval corrects 33 baseline errors but introduces 28 new errors, leaving a net gain of five correct answers. Both settings fail on 129 items (51.2%).

Table 4: Qwen3.5-397B-A17B answer outcomes with and without retrieval, and tool usage in agentic runs. $\surd$ and $\times$ denote correct and incorrect answers, respectively.

Bucket	Base	Agent	N	Share (%)	Rounds	Search	Visit	Img.
Both wrong	$\times$	$\times$	129	51.2	4.60	6.48	1.25	0.69
Both correct	$\surd$	$\surd$	62	24.6	4.31	5.81	1.34	0.55
Search helps	$\times$	$\surd$	33	13.1	4.73	6.91	1.24	0.70
Search hurts	$\surd$	$\times$	28	11.1	4.89	7.11	1.75	0.54
Total	–	–	252	100.0	4.58	6.44	1.33	0.64

Base: reasoning-only; Agent: agentic retrieval. Rounds, Search, Visit, and Img. report mean counts per agentic run; Img. denotes image search.

Correct and incorrect agentic runs use similar numbers of tool calls on average: 6.19 vs. 6.59 web searches, 1.31 vs. 1.34 page visits, and 0.60 vs. 0.66 image searches. The *Search helps* group averages 6.91 searches, compared with 7.11 for *Search hurts*. Useful-evidence hit rates are 0.12% across 1623 web-search calls and 5.69% across 334 page visits. To examine the failure patterns, a separate Qwen3.5-397B-A17B judge classifies the 157 incorrect agentic runs into tool-related (T1–T4) and non-tool (N1–N2) categories (Table 5).

Table 5: Judge-assigned failure categories for 157 incorrect agentic retrieval runs.

ID	Failure cause	BW	Loss	Total	Share (%)	Rounds	Search	Image Search
Tool-related failures								
T1	Reverse image search ineffective	8	2	10	6.4	5.30	6.70	1.00
T2	Insufficient text-search coverage	9	2	11	7.0	6.36	7.27	0.91
T3	Confirmation / anchoring loop	90	21	111	70.7	4.56	6.96	0.64
T4	Cross-lingual query dilution	3	1	4	2.5	4.50	7.00	0.75
Non-tool failures								
N1	Subjective reference or strict judging	13	2	15	9.6	4.13	4.47	0.53
N2	Other reasoning or execution errors	6	0	6	3.8	3.67	3.33	0.33
Total	—	129	28	157	100.0	—	—	—

BW: both wrong; Loss: Search hurts. Rounds, Search, and Image Search. are mean counts per run.

This classification attributes most incorrect runs to tool failures. Confirmation and anchoring loops T3 dominate failure cases. Agents in such traces lock into early wrong hypotheses and reinforce them via searches, ignoring contradictory evidence. The Search hurts group makes more web queries than the Both wrong group. This matches persistent retrieval along faulty directions, though query volume cannot prove failure origins. Two bottlenecks limit retrieval performance: extracting valid evidence from UGC and specialized communities, and updating flawed hypotheses mid-search. Combined results and failure analysis explain the marginal net gain of this agentic setup on NoteVQA.

4.3 Image Ablation of Interleaved References

We assess the contribution of embedded images using probes derived from 205 references that contain at least one image. A multimodal generator creates four probes per item: two *image-critical* probes, designed to require information from an image, and two *text-answerable* probes, designed to be answerable from the paragraph alone. A positive-control filter removes probes that the reader cannot answer with the full interleaved input, retaining 367 image-critical and 375 text-answerable probes.

The reader, Qwen3.5-397B-A17B-FP8, answers each probe with the same paragraph text under three conditions: image tags removed (y_{text}), tags replaced by `[image N]` placeholders (y_{ph}), or images displayed inline (y_{full}). The reader receives neither the original query image nor the probe reference answer. A text-only judge grades each response.

Table 6: Reader accuracy on retained probes under image ablation (%; 95% bootstrap CIs).

Probe type	Cond.	Hits/N	Acc %	95% CI
Image-crit.	y_{text}	76/367	20.7	[16.6, 25.1]
Image-crit.	y_{ph}	115/367	31.3	[26.7, 36.2]
Image-crit.	y_{full}	356/367	97.0	[95.4, 98.6]
Text-answ.	y_{text}	369/375	98.4	[97.1, 99.5]
Text-answ.	y_{ph}	367/375	97.9	[96.3, 99.2]
Text-answ.	y_{full}	368/375	98.1	[96.5, 99.5]

Table 7: Image-critical accuracy by intent (%). $\Delta_{\text{f-t}}$: full minus text (%).

Intent	N	y_{text}	y_{ph}	y_{full}	$\Delta_{\text{f-t}}$
Authenticity	66	13.6	19.7	98.5	+84.8
Diagnosis	65	30.8	36.9	100.0	+69.2
Identification	53	22.6	30.2	98.1	+75.5
Object cond.	35	8.6	28.6	88.6	+80.0
Physiology	57	28.1	40.4	94.7	+66.7
Procedural	49	12.2	26.5	98.0	+85.7
Spoilage	42	23.8	38.1	97.6	+73.8
Overall	367	20.7	31.3	97.0	+76.3

Table 6 shows that image-critical accuracy rises from 20.7% with text alone to 97.0% with the full interleaved input, a gain of 76.3%. Placeholder markers raise accuracy to 31.3%, a gain of 10.6% over text alone; displaying the images adds a further 65.7%. Text-answerable accuracy remains between 97.9% and 98.4%, varying by at most 0.5% across conditions.

The full-minus-text accuracy gain is positive across all seven intents, ranging from +66.7 to +85.7% (Table 7). In the 281 probes corrected by y_{full}, the text-only responses report missing information that the images subsequently provide. These results indicate that embedded images supply useful information beyond the paragraph text on the selected image-critical probes.

5 Related Work

Multimodal QA benchmarks. Existing multimodal QA benchmarks evaluate factual knowledge, visual understanding, and information seeking through answer-level correctness. SimpleVQA [6] and FVQA [25] focus on factual visual question answering, while LiveVQA [8] examines the acquisition and use of up-to-date visual knowledge. MM-BrowseComp [16], MMSearch [14], and MMSearch-Plus [22] extend evaluation to multimodal browsing and retrieval, with MMSearch-Plus emphasizing fine-grained visual cues and provenance verification. BC-VL [9] and VDR-Bench [28] further construct challenging visual search tasks through entity fuzzification and knowledge-graph-based query expansion, respectively, reducing opportunities for text-only shortcuts or simple image matching. These benchmarks advance the evaluation of visual reasoning and search, but their answer-level metrics do not directly assess the quality of a complete explanation supported by embedded images. NoteVQA complements this line of work with community-sourced visual questions and paired short and interleaved references, enabling both correctness and explanation-quality evaluation.

Multimodal deep research. Recent work extends multimodal evaluation from factual answers to long-form research reports. MMDR-Bench [12] comprises 140 expert-crafted tasks across 21 domains and evaluates multimodal understanding, citation grounding, and report synthesis using image–text task bundles. MiroEval [27] contains 100 tasks, including 30 multimodal tasks, constructed from real user query patterns and web trends; its evaluation covers synthesis quality, factuality, and the research process. Complementing these benchmarks, PTAH [29] coordinates planning, research, writing, and verification agents to generate interleaved reports, with additional evaluation of image content and multimodal presentation quality. These efforts primarily target extended research workflows and report generation. NoteVQA focuses on everyday visual questions posted in online communities and provides human-audited interleaved references alongside short answers, connecting factual QA evaluation with the quality of visually supported responses.

Rubric design and model-based evaluation. Multimodal report evaluation accounts for textual statements and their visual backing. MMDR-Bench [12] measures visual evidence fidelity and consistency between image-linked claims and visual content. PTAH [29] judges image content and layout via cross-modal alignment, information complementarity, and density-legibility tradeoff. Extending these works, IVR-12 assesses interleaved answers across 12 metrics under content, presentation and image quality. Its image metrics cover relevance, necessity and placement, visual quality and non-redundancy; content and presentation metrics check answer accuracy, utility and readability. We use this unified rubric for human-reviewed references and model outputs, enabling fair comparison of interleaved answer quality on real visual queries.

6 Conclusion

We introduce NoteVQA, a benchmark of 252 community-sourced visual questions that connects real user needs with evaluations of answer correctness and visually grounded explanation quality. Each question is paired with a short reference and a human-audited interleaved reference, complemented by AgenticInterleave for answer generation and IVR-12 for evaluation. Across 9 frontier VLMs, the highest short-answer accuracy is only 52.8%, highlighting the substantial challenge posed by everyday visual questions. Agentic retrieval improves Qwen3.5-397B-A17B by just 2.0 percentage points, while its interleaved answers score 3.52 under IVR-12 compared with 4.65 for the references, with the largest gap in content quality. In the image-ablation study, including embedded images raises reader accuracy on image-critical probes from 20.7% to 97.0%, while performance on text-answerable controls remains stable. This demonstrates that the images contribute information absent from the accompanying text. Together, these findings highlight the need for VLMs that can both answer real-world visual questions accurately and communicate the supporting evidence effectively.

Ethics Statement

This work adheres to the ICLR Code of Ethics. All data in this paper were obtained in accordance with relevant usage guidelines and processed with anonymization and de-identification to prevent disclosure of personal privacy. We commit to maintaining transparency and academic integrity throughout the research process.

Reproducibility Statement

We fully report model configurations, hyperparameters and evaluation protocols. All public datasets used are named and accessible for replication.

AI Assistance Statement

LLMs were used for manuscript polishing, grammar revision, readability improvement, data cleaning and data labeling, as well as partial data analysis and experimental implementation. They were not involved in hypothesis formulation and overall experimental design. All scientific content was created and verified by authors, who take full responsibility for the paper.

References

[1] Lada A Adamic, Jun Zhang, Eytan Bakshy, and Mark S Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *Proceedings of the 17th international conference on World Wide Web*, pp. 665–674, 2008.

[2] Anthropic. Introducing claude opus 4.7, 4 2026.

[3] Anthropic. Introducing claude sonnet 4.6, 2 2026.

[4] 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 (CVPR)*, 2022.

[5] 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14948–14968, 2023.

[6] Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, et al. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

[7] Google DeepMind. Gemini 3.1 pro, 2 2026.

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

[9] Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang, Jialong Wu, Kuan Li, Yida Zhao, et al. Webwatcher: Breaking new frontiers of vision-language deep research agent. In *International Conference on Learning Representations (ICLR)*, 2026.

[10] GLM-5-Team, :, Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chenghua Huang, Chengxing Xie, Chenzheng Zhu, Congfeng Yin, Cunxiang Wang, Gengzheng Pan, Hao Zeng, Haoke Zhang, Haoran Wang, Huilong Chen, Jiajie Zhang, Jian Jiao, Jiaqi Guo, Jingsen Wang, Jingzhao Du, Jinzhu Wu, Kedong Wang, Lei Li, Lin Fan, Lucen Zhong, Mingdao Liu, Mingming Zhao, Pengfan Du, Qian Dong, Rui Lu, Shuang-Li, Shulin Cao, Song Liu, Ting Jiang, Xiaodong Chen, Xiaohan Zhang, Xuancheng Huang, Xuezhen Dong, Yabo Xu, Yao Wei, Yifan An, Yilin Niu, Yitong Zhu, Yuanhao Wen, Yukuo Cen, Yushi Bai, Zhongpei Qiao, Zihan Wang, Zikang Wang, Zilin Zhu, Ziqiang Liu, Zixuan Li,

Bojie Wang, Bosi Wen, Can Huang, Changpeng Cai, Chao Yu, Chen Li, Chengwei Hu, Chenhui Zhang, Dan Zhang, Daoyan Lin, Dayong Yang, Di Wang, Ding Ai, Erle Zhu, Fangzhou Yi, Feiyu Chen, Guohong Wen, Hailong Sun, Haisha Zhao, Haiyi Hu, Hanchen Zhang, Hanrui Liu, Hanyu Zhang, Hao Peng, Hao Tai, Haobo Zhang, He Liu, Hongwei Wang, Hongxi Yan, Hongyu Ge, Huan Liu, Huanpeng Chu, Jia'ni Zhao, Jiachen Wang, Jiajing Zhao, Jiamin Ren, Jiapeng Wang, Jiaxin Zhang, Jiayi Gui, Jiayue Zhao, Jijie Li, Jing An, Jing Li, Jingwei Yuan, Jinhua Du, Jinxin Liu, Junkai Zhi, Junwen Duan, Kaiyue Zhou, Kangjian Wei, Ke Wang, Keyun Luo, Laiqiang Zhang, Leigang Sha, Liang Xu, Lindong Wu, Lintao Ding, Lu Chen, Minghao Li, Nianyi Lin, Pan Ta, Qiang Zou, Rongjun Song, Ruiqi Yang, Shangqing Tu, Shangtong Yang, Shaoxiang Wu, Shengyan Zhang, Shijie Li, Shuang Li, Shuyi Fan, Wei Qin, Wei Tian, Weining Zhang, Wenbo Yu, Wenjie Liang, Xiang Kuang, Xiangmeng Cheng, Xiangyang Li, Xiaoquan Yan, Xiaowei Hu, Xiaoying Ling, Xing Fan, Xingye Xia, Xinyuan Zhang, Xinze Zhang, Xirui Pan, Xu Zou, Xunkai Zhang, Yadi Liu, Yandong Wu, Yanfu Li, Yidong Wang, Yifan Zhu, Yijun Tan, Yilin Zhou, Yiming Pan, Ying Zhang, Yinpei Su, Yipeng Geng, Yong Yan, Yonglin Tan, Yuean Bi, Yuhan Shen, Yuhao Yang, Yujiang Li, Yunan Liu, Yunqing Wang, Yuntao Li, Yurong Wu, Yutao Zhang, Yuxi Duan, Yuxuan Zhang, Zezhen Liu, Zhengtao Jiang, Zhenhe Yan, Zheyu Zhang, Zhixiang Wei, Zhuo Chen, Zhuoer Feng, Zijun Yao, Ziwei Chai, Ziyuan Wang, Zuzhou Zhang, Bin Xu, Minlie Huang, Hongning Wang, Juanzi Li, Yuxiao Dong, and Jie Tang. Glm-5: from vibe coding to agentic engineering, 2026.

[11] Hang He, Chuhuai Yue, Chengqi Dong, Chengcheng Wan, Ting Su, Haiying Sun, Jiajun Chai, Xiaohan Wang, and Guojun Yin. Vistahop: Benchmarking multi-hop visual reasoning for visual deepsearch. *arXiv preprint arXiv:2606.03273*, 2026.

[12] Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, et al. Mmdeepresearch-bench: A benchmark for multimodal deep research agents. *arXiv preprint arXiv:2601.12346*, 2026.

[13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[14] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

[15] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, et al. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*, 2025.

[17] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[18] Meredith Ringel Morris, Jaime Teevan, and Katrina Panovich. What do people ask their social networks, and why? a survey study of status message Q&A behavior. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2010.

[19] Qwen Team. On the design of Qwen3.8-Next architecture: Evaluation, efficiency, and training stability. Technical report, Alibaba Group, August 2026.

[20] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026.

[21] Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

[22] Xijia Tao, Teng Yihua, Xinxing Su, Xinyu Fu, Jihao Wu, Chaofan Tao, Ziru Liu, Haoli Bai, Rui Liu, and Lingpeng Kong. Mmsearch-plus: Benchmarking provenance-aware search for multimodal browsing agents. In *International Conference on Learning Representations (ICLR)*, 2026.

[23] Kimi Team. Kimi k2.6: Advancing open-source coding. Moonshot AI Blog, 2026.

[24] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

[25] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(10):2413–2427, 2017.

[26] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[27] Fangda Ye, Yuxin Hu, Pengxiang Zhu, Yibo Li, Ziqi Jin, Yao Xiao, Yibo Wang, Lei Wang, Zhen Zhang, Lu Wang, et al. Miroeval: Benchmarking multimodal deep research agents in process and outcome. *arXiv preprint arXiv:2603.28407*, 2026.

[28] Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, et al. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

[29] Chenghao Zhang, Guanting Dong, Yufan Liu, Tong Zhao, Xiaoxi Li, and Zhicheng Dou. Towards verifiable multimodal deep research: A multi-agent harness for interleaved report generation. *arXiv preprint arXiv:2605.29861*, 2026.