CRAG-MM-Diagnostics: Enabling Stage-Wise Analysis of Knowledge-Intensive VQA

Hanseok Oh[1,3⋆], Parishad BehnamGhader[2,3], Benno Krojer[2,3], Hyunji Lee[4], Paul Liang[5], Siva Reddy[2,3,6], and Verna Dankers[2,3]

[1] New York University, New York NY, USA
[2] McGill University, Montreal, Canada
[3] Mila - Quebec AI Institute, Montreal, Canada
[4] UNC Chapel Hill, Chapel Hill NC, USA
[5] MIT, Cambridge MA, USA
[6] Canada CIFAR AI Chair
hanseok.oh@nyu.edu

McGill-NLP/crag-mm-diagnostics CRAG-MM-Diagnostics

Abstract. *Knowledge-Intensive Visual Question Answering* (KI-VQA) benchmarks evaluate *Vision–Language Models* (VLMs) as multimodal knowledge assistants by requiring external information beyond a provided image to answer questions. KI-VQA involves multiple sub-problems —referring expression understanding, visual grounding, object recognition, knowledge retrieval, and reasoning—yet existing benchmarks typically report only end-task accuracy, obscuring where failures arise. To analyze the *full* KI-VQA pipeline, we introduce `CRAG-MM-Diagnostics`, a diagnostic benchmark with stage-wise data annotations that isolate ① language-based visual grounding, ② object identification, and ③ knowledge retrieval and reasoning. We evaluate fully parametric and retrieval-augmented VLMs, providing fine-grained analyses using newly collected metadata, such as target ROIs, entity names, and visual complexity scores. Our results point to knowledge retrieval and reasoning as the primary bottleneck, but also highlight issues in the other parts of the KI-VQA pipeline, such as the fact that VLMs struggle with target object identification or that image retrievers struggle to integrate textual cues. These findings expose fundamental limitations in current KI-VQA systems and motivate stage-aware evaluation. We, lastly, leverage these findings to propose a grounded bimodal RAG pipeline that integrates a visual grounding module to crop targets before image retrieval, boosting `GPT-5` and `Qwen`'s respective accuracies by 13.3 and 8.5 percentage points.

Keywords: Knowledge-intensive VQA · Retrieval Augmented Generation · Vision-Language Models

1 Introduction

Recent advances in *Vision-Language Models* (VLMs) have accelerated interest in their deployment in real-world applications [2, 4, 25, 32]. A core requirement

⋆ Work conducted during an internship at Mila.

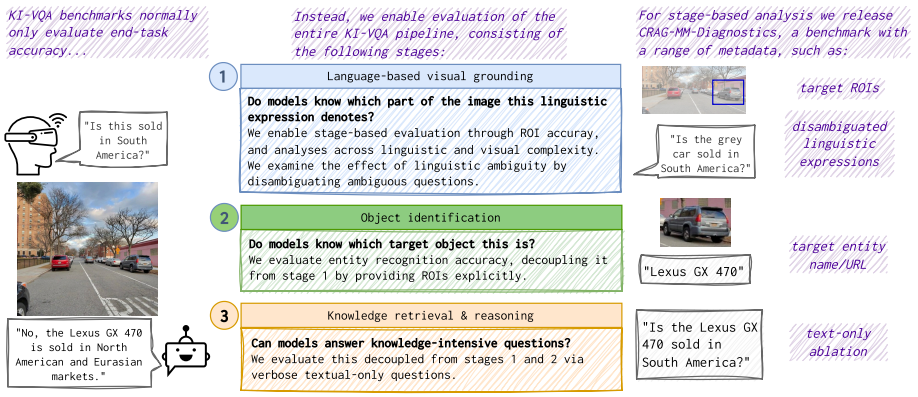


Fig. 1: Illustrative summary of the stage-wise evaluation enabled by `CRAG-MM-Diagnostics`, thanks to new types of metadata we add. The three stages evaluated are (1) language-based visual grounding, (2) object identification, and (3) knowledge extraction & reasoning, and are further detailed in the figure.

for these assistants is the ability to interpret visual scenes and provide informative responses, typically evaluated through *Visual Question Answering* (VQA) benchmarks. However, while standard VQA often focuses on describing visible attributes, there is growing attention for *Knowledge-Intensive* VQA (KI-VQA). In this setting, the question cannot be answered solely from the image but also requires external world knowledge [3, 10, 20, 22, 33]. KI-VQA is a vital step towards a future in which VQA is seamlessly integrated in daily life, in which one can ask complex questions on the go, when walking down the street using smart glasses, or texting photos back and forth with a friend.

However, existing KI-VQA benchmarks often fail to capture real-world visual complexity, instead relying on images with salient, centered objects and minimal clutter [3, 20, 22]. In contrast, real-world scenes contain multiple, often visually similar objects, with less salient and off-center targets (see Figure 1). Challenging benchmarks that do include visually complex images often focus only on end-task evaluation [8, 32], obscuring the root causes of failure (i.e., whether errors stem from flawed visual grounding, incorrect entity identification, or downstream reasoning). For example, when a model fails to answer the question in Figure 1, is the issue target identification or missing knowledge?

In this work, we take a step towards a more fine-grained analysis of KI-VQA, aiming to understand how models perform in intermediate stages of the pipeline. Unlike the majority of KI-VQA studies that only evaluate end-task accuracy, we focus on three intermediate stages: ① **language-based visual grounding**, used to localize the target region of interest in a complex image [26, 35]; ② **object identification**, used to map the localized region to a canonical entity (e.g., a Wikipedia page) [22, 25]; and ③ **knowledge retrieval and reasoning**, which answers the question after possible textual and visual information retrieval [3, 20]. To enable this analysis, we introduce `CRAG-MM-Diagnostics`, a benchmark built

on top of CRAG-MM [32], as described in Section 3. While CRAG-MM provides a foundation for KI-VQA in egocentric, visually complex scenes, it serves primarily as an end-to-end benchmark without the structural markers needed to diagnose internal model failures. CRAG-MM-Diagnostics transforms this resource into a diagnostic framework by introducing stage-specific annotations. Figure 1 illustrates the three stages, and a small subset of the new metadata.

Using the extended benchmark, we conduct a comprehensive analysis for the three stages in Sections 4, 5 and 6, using fully parametric VLMs, models specialized for each stage, and retrieval-augmented models. Our analysis uncovers several fundamental limitations, among which:

- Visual grounding of target objects is more challenging for proprietary models;
- Linguistic ambiguity in QA pairs affects the entire KI-VQA pipeline;
- VLMs, in particular compared to specialized models, struggle to identify target objects that are unpopular;
- Localizing the target object improves image retrieval substantially, especially in ambiguous queries;
- Most KI-VQA errors are due to inadequate knowledge retrieval and reasoning rather than incorrect object identification (e.g., only 20.9% of GPT-5's errors can be resolved by providing the ground-truth target object name).

By providing a stage-aware analysis, we offer actionable insights for the design of robust KI-VQA systems. In Section 6, we leverage our findings to study a modeling pipeline combining components from all stages to achieve better performance, boosting GPT-5 and Qwen's respective accuracies by 13.3 and 8.5 percentage points. We conclude in Section 7.

2 Related Work

In this section, we discuss the evolution of KI-VQA and recent advances in diagnostic evaluation for VLMs. We provide a more detailed comparison in Section A.

2.1 Knowledge-Intensive Visual Question Answering Benchmarks

Since the introduction of VQA [1], later benchmarks have increased difficulty by requiring external knowledge, leading to the development of *knowledge-intensive* VQA (KI-VQA). *OK-VQA* [20] focuses on commonsense and simple factual knowledge. Subsequent datasets impose more specialized knowledge demands: *EncyclopedicVQA* [22] and *SnapNTell* [28] target long-tail entities (e.g., "*Pinus pinea*"), while *ViQuAE* [12] and *InfoSeek* [3] emphasize named entities and realistic information-seeking scenarios. These benchmarks typically require retrieving external knowledge.

Despite this progress, most KI-VQA benchmarks increase difficulty primarily along the *language axis*—e.g., by requiring complex reasoning [31] or including temporally evolving knowledge [14]—while implicitly assuming object

recognition is trivial or based on salient cues [3,22,28]. In contrast, fine-grained visual benchmarks focus on challenges along the *visual axis*. Datasets such as *V** [35], *GigaGrounding* [17], and *MME-RealWorld* [40] focus on visually challenging scenarios, including crowded, high-resolution, or remote-sensing images where targets are small or non-salient, requiring precise localization. However, these benchmarks are mainly studied outside information-seeking settings.

Such visual difficulty becomes even more critical when retrieval is involved, as external tools must correctly identify subtle targets under noisy conditions—an aspect largely overlooked in KI-VQA. Recent multimodal *Retrieval Augmented Generation* (RAG) benchmarks, such as *MRAG-Bench* [8] and *CRAG-MM* [32], incorporate realistic visual challenges like viewpoint shifts, occlusion, and egocentric perspectives. However, these benchmarks lack fine-grained diagnostics to attribute failures to specific visual factors or to disentangle visual and linguistic errors. We, therefore, present `CRAG-MM-Diagnostics`, which provides fine-grained stage-wise diagnostics for *language-based visual grounding*, *object identification*, and *knowledge retrieval and reasoning*.

2.2 Diagnostic evaluation of VLMs

Recent studies have proposed diagnostic evaluations of intermediate stages to better understand VLMs' capabilities in multimodal tasks. *MMVet* [39] annotates examples from 16 multimodal tasks with six core vision–language skills, which are not specifically curated for the KI-VQA task. *HallusionBench* [6] decomposes hallucinations into language hallucination and visual illusion to analyze failure modes. Similarly, *Prism* [27] disentangles perception from reasoning in general VQA. Despite these advances, diagnostic evaluation for *knowledge-intensive* information seeking tasks remains limited. Existing evaluations are often confined to isolated visual recognition tasks [7] or rely solely on final QA accuracy [3, 20, 22], overlooking realistic KI-VQA scenarios that require precise grounding and contextual target identification [18,38]. The recent KI-VQA benchmark *VisualSimpleQA* [34] partially addresses this gap by comparing multimodal and text-only questions, but still focuses on end-task accuracy and is limited to parametric VLMs.

3 The curation of `CRAG-MM-Diagnostics`

To facilitate fine-grained KI-VQA analysis, we develop a diagnostic benchmark (`CRAG-MM-Diagnostics`) by augmenting and refining `CRAG-MM` [32]. `CRAG-MM` consists of diverse (image, question, answer) triplets spanning 13 domains, including nearly 1.9K single-turn egocentric images that mimic captures from wearable devices. This egocentric setting enables *realistic* KI-VQA evaluation and provides sufficient visual diversity and complexity to analyze performance across different levels of visual saliency—e.g., by varying object sizes or scene crowdedness. However, it only permits end-to-end QA analysis, making it impossible to investigate the ability of different KI-VQA components separately.

Human annotations. To ensure the benchmark's quality, we apply an initial filtering process prior to human annotation, removing samples that are not *knowledge-intensive* or concern dynamically changing knowledge, as detailed in Section A.2. We then manually annotate textual metadata and target object regions using bounding boxes.[7] Each sample in the dataset is annotated once by an author and consecutively double-checked by another. The metadata added through human annotation, illustrated using Figure 1, includes:

1. **Entity name, Wikipedia URL**: the target object's name and corresponding Wikipedia URL (e.g., "Lexus GX 470" in case of Figure 1).
2. **Referring expression type**: the queries' labels that distinguish *salient* cases (e.g., where "this van" suffices as a referral to identify a unique object) from *ambiguous* ones (e.g., the non-specific expression "this" in Figure 1), and ones where there is an *in-image cue* (e.g., "the red car") from a *knowledge-intensive cue* (e.g., recognizing the service associated with a logo to identify that "this restaurant" refers to a specific building). See **??** for examples.
3. **Disambiguated question**: for examples with ambiguous referring expression, we add a disambiguated question. For instance, for Figure 1, this question replaces "this car" with "the grey car on the right".
4. **Text-only question**: for all examples, we add a question which allows models to rely on the textual modality only, by inserting the correct entity into the question, e.g., "Is the Lexus GX 470 sold in South America?" for Figure 1.
5. **Target region**: in addition to the textual metadata mentioned above, the human annotation also involves the selection of a bounding box in which the target object is contained, as previously demonstrated in Figure 1.

After annotation, we removed 95 samples due to missing Wikipedia links for the target entity, mismatched image-QA pairs, questions that do not require visual information, low-quality images that hinder localization (e.g., due to blur), and time-dependent questions (e.g., "last year") identified during the annotation.

Automated annotations. Additionally, we automatically augment the data with metadata capturing an object's popularity, and its visual complexity:

6. **Target popularity**: the Wikipedia page views in 2025, following [19, 28].
7. **Target object size**: measured as the ratio of the target's bounding box area to the image area.
8. **Distance to center**: computed as the distance between the target centroid and the image center.
9. **Scene crowdedness**: approximated by the number of objects detected using the open-set detector Grounding-DINO [16].

Using the three visual complexity dimensions (metadata 7–9), we define a summary metric, **visual saliency**, to quantify a target's prominence in an image (see Section A.4). After filtering and annotation, `CRAG-MM-Diagnostics` contains 1,149 samples. Dataset statistics are shown in **??**; example QA pairs, further metadata details, the annotation protocol and the annotation tool are described in Section A.

[7] Some metadata annotations were model-assisted via preliminary annotations, as detailed in Section A.4.

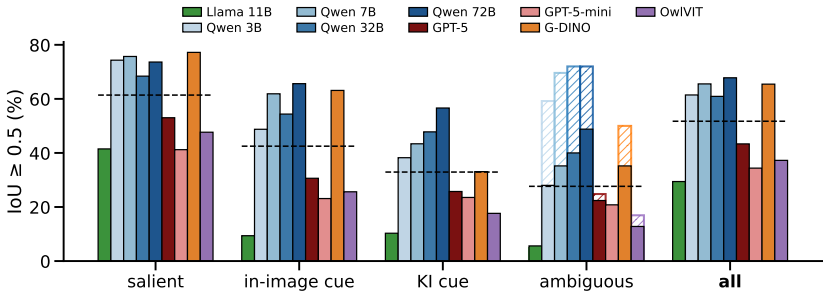


Fig. 2: Grounding accuracy per referring expression type (horizontal lines: mean; hatched: disambiguated query). It shows that: (1) Drops on non-salient and ambiguous targets reveal high sensitivity to expression type. (2) Gains from disambiguation suggest many failures arise from query ambiguity rather than visual processing deficits.

4 Stage (1) Language-based Visual Grounding

We use `CRAG-MM-Diagnostics` to systematically evaluate the KI-VQA stages (as shown in Figure 1), beginning with the system's ability to localize and identify the subject of a referring expression within a visual scene. We assess this **language-based visual grounding** ability by measuring how accurately models predict the bounding box of a target object across varying linguistic and visual variations. We first describe the experimental setup.

4.1 Experimental Preliminaries

Models. Our evaluation spans both generalized VLMs (which are capable of end-to-end KI-VQA) and specialized models designed specifically for this step of the pipeline (i.e., localization). We examine the performance of both, to inform building KI-VQA solutions that integrate specialized models, which we will perform later on, in Section 6 (when performing stage (3) evaluation).

Among the **generalized models** that jointly encode images and text, we include open-source architectures `Llama-3.2-11B` [23] and the `Qwen2.5-VL` family (ranging from 3B to 72B parameters) [2], alongside proprietary state-of-the-art models `GPT-5` and `GPT-5-mini`. The **specialized models** focus explicitly on alignment between language and visual regions: `Grounding-DINO` [16] and `OWL-ViT` [24] perform open-vocabulary object detection by aligning text embeddings with image features for precise localization. Detailed prompts and details for this task are included in Section C; the prompts provide models with the question and image, and require outputting precise bounding box coordinates.

Evaluation metric. Following established related work on visual grounding evaluation [17,36], we measure *Intersection over Union* (IoU) between the predicted and ground-truth bounding boxes of the target object. A prediction is classified as correct if the IoU exceeds a threshold of 0.5.

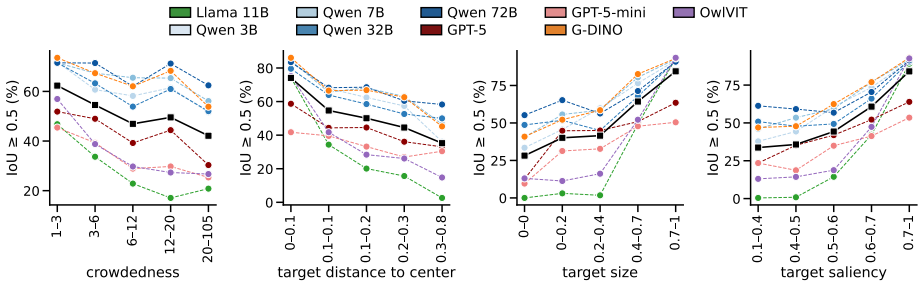


Fig. 3: Visual grounding scores along four dimensions of visual complexity (black squares indicate model mean). These plots show that low target saliency is the primary bottleneck, acutely compounding performance drops caused by scene density, off-center placement, and small target size.

4.2 Analyses and Findings

Figure 2 details the IoU per model, per referring expression type, and across all examples through the '**all**' bar. The `Qwen2.5-VL` family demonstrates robust grounding across all sizes, peaking at 67.8% accuracy. `Llama-3.2` and the `GPT-5` variants, however, have markedly lower performance. The fact that even the smallest `Qwen` is quite good at this task is likely because the model family was explicitly trained to understand bounding boxes [2]; the `GPT-5` results, in particular, underscore that spatial understanding of images does not naturally emerge, even in SOTA models. Even if `GPT-5` *can* perform KI-VQA in general, the fact that it struggles with explicit grounding in this first stage limits the explainability and interpretability of the model. When contrasting the generalized and specialized models, it is noteworthy that `Grounding-DINO` performs remarkably well; despite its modest 0.2B parameters, it rivals the `Qwen` models in accuracy, presenting a highly efficient alternative for modular KI-VQA pipelines that include grounding.

Referring expressions highly influence performance. Overall performance obscures the impact of the *type* of referring expression. The 'referring expression' metadata we include in `CRAG-MM-Diagnostics` distinguishes salient cases from more complex ones, which constitute 36.6% of the dataset. Section A.3 provides concrete examples of the different types of expressions. Figure 2 shows that while salient targets are the most straightforward to ground, ambiguous expressions represent a severe challenge. This gap is especially pronounced for smaller models (e.g., `Qwen` 3B and 7B), highlighting the role of model capacity in handling linguistic nuance and multimodal integration. Moreover, results demonstrate that although `Grounding-DINO` performs similarly to the `Qwen` family overall, it falls short of the larger models in samples with complex referring expressions (i.e., in the case of a 'knowledge-intensive cue' and 'ambiguous' examples).

Ambiguity is a bottleneck for visual grounding. To isolate the effects of ambiguity, we compared performance on samples with ambiguous queries against disambiguated ones (e.g., changing "this" to "the grey car on the right" in Figure 1). Disambiguation resulted in a substantial performance surge, a mean 44.0% improvement (see hatched bars in Figure 2). This indicates that a lack of linguistic specificity is a primary bottleneck. The gains are largest for `Qwen` models, whereas disambiguated queries do not improve performance for `Llama` and `GPT-5-mini`, underscoring that their subpar performance is a more fundamental issue with these models' spatial reasoning abilities rather than merely the result of ambiguity. The presence of ambiguous queries in `CRAG-MM` is likely due to the data's egocentric nature. If a QA pair does not capture the user's gaze or intent, evaluation using the original queries may underestimate models' true KI-VQA performance.

Grounding degrades as visual complexity grows. Language-based visual grounding depends not only on the referring expression but also on image complexity and target saliency [17, 35]. Figure 3 demonstrates this for the four visual complexity axes previously introduced in Section 3. Performance decreases with increased crowdedness and distance—with *point biserial* (pb) correlations of $r_{pb} = -.117$ and $-.254$, respectively, averaged over models—and improves with larger size and higher saliency—$r_{pb} = .422$ and .391.

Intermediate takeaways based on visual grounding (stage ①) evaluation:

1. VLMs lack spatial awareness without being explicitly trained for this, as `Llama` and `GPT-5(-mini)` struggle a lot more than `Qwen` when predicting bounding boxes;
2. Complex referring expressions and ambiguity negatively affect visual grounding;
3. In line with prior work [17,35] various factors of visual complexity degrade grounding performance;
4. `Grounding-DINO` best balances performance with efficiency.

5 Stage ② Object Identification

Next, we study the models' ability to perform **target object identification**; a vital step towards KI-VQA since the vast majority of `CRAG-MM-Diagnostics` questions concern specific named entities. Here, we first elaborate on the experimental setup prior to diving into our findings. We evaluate models by providing the original image and question, alongside task-specific instructions (detailed in Section D), and require the models to output the name of the target entity.

5.1 Experimental Preliminaries

Models. Consistent with our previous analysis, we evaluate two model classes. Firstly, we consider **generalized models** (see Section 4.1), which rely solely on

internal parametric knowledge. Evaluating them on object identification, therefore, isolates a step that these models normally perform implicitly during KI-VQA. In contrast, **specialized models** retrieve images and metadata from external knowledge sources. They implicitly perform object identification when at least one retrieved item corresponds to the target. Although this does not yield a direct comparison with generalized models, it allows us to examine complementary strengths of parametric vs. retrieval-based identification. The specialized models are the unimodal retriever `CLIP-ViT-Large-Patch14-336` [29] and the multimodal retrievers `VLM2Vec-V2.0` [21] and `Qwen-3-VL-Embedding-2B` [13]. We use the `CRAG-MM` image knowledge graph [32] as the retrieval corpus and recompute embeddings for each model.

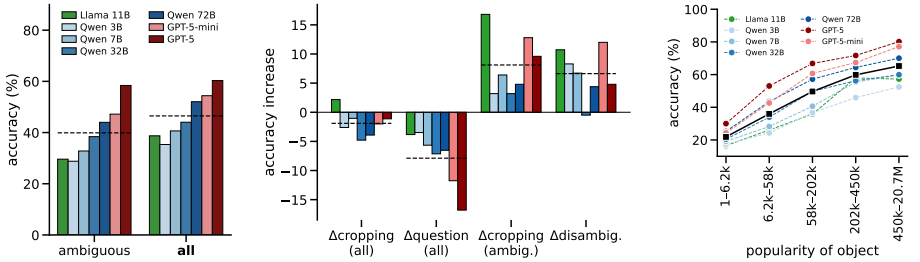
Evaluation metric. For generalized models, we measure entity prediction accuracy using exact string match. For cases without an exact match, we utilize LLM-as-a-judge (`GPT-4o-mini`) to determine semantic equivalence. For specialized retrieval models, we report `recall@10`. Specifically, we perform normalized partial string matching between the ground-truth entity name and entity names associated with the top-10 retrieved images following related work [5,30]. We do not apply the LLM-as-a-judge approach here, as verifying semantic equivalence for all top-k results from retrieval sources is very computationally expensive.

5.2 Analyses and Findings

Before diving into fine-grained analyses, Figure 4a reveals the overall performance and performance for examples with ambiguous referring expressions. Although proprietary models struggled with visual grounding in Section 4, they now achieve the strongest object identification performance. Grounding being a prerequisite for object identification suggests that while these models lack the specialized 'language' of coordinate-based grounding, they possess a strong implicit grounding capability that facilitates identification. Yet, even `GPT-5` struggles to predict the correct entity name for approximately 40% of the examples, which will have implications for the end-task KI-VQA accuracy down the line.

VLMs benefit from multimodal cues. Next, we analyze the contribution of each modality to object identification for the generalized models (Figure 4b):

- Δ cropping: Providing a crop of the target region instead of the full image results in a slight net decrease in accuracy for most models, though it predictably aids ambiguous stimuli. This suggests that models might benefit from contextual cues beyond the target region for performing accurate object identification.
- Δ question: Replacing the original query with a simplified prompt (e.g., "What is this object?") reduces accuracy by 7.9 percentage points. This confirms that models leverage both visual and textual cues to refine their visual identification. An example of a useful textual cue is, for instance, if the question refers to an "SUV", greatly restricting the set of possible target objects.



(a) OI performance **(b)** Effect of modifying the question/image **(c)** Accuracy vs popularity

Fig. 4: Performance breakdown of generalized models on the stage ② object identification task. (b) Accuracy changes under input manipulations: Δ 'cropping' marks the accuracy change after cropping the target object; Δ 'question' marks the change after simplifying the question; Δ 'disambig.' marks the change after using the disambiguated questions for the ambiguous stimuli. (c) Accuracy across target object popularity bins.

- Δ disambiguated: Using disambiguated questions provides performance gains nearly on par with those of cropping, suggesting that ambiguous examples cause an underestimation of a model's latent KI-VQA capabilities.

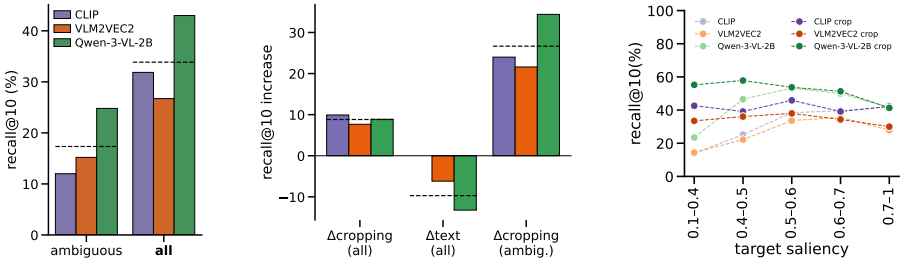
VLMs are affected by objects' popularity. Next, we utilize other types of metadata to further understand what does or does not affect object identification performance when the generalized VLMs perform the task. Unlike the grounding task in Stage ①, object identification is less sensitive to visual complexity; in fact, saliency is (only weakly) negatively related to accuracy ($r_{pb} = -.168$) (described in Section D.1). A clearer trend, however, exists for the impact of popularity as depicted in Figure 4c (with an average Spearman's ρ of 0.314): VLMs are better at identifying entities with high Wikipedia page-view counts, consistent with established literature on long-tail visual knowledge distribution [7, 22, 28].

Image retrievers benefit from visual isolation and struggle with linguistic comprehension. Specialized retrievers present a different performance profile. Figure 5a shows recall@10 for all examples versus ambiguous cases using image-only queries.[8] These results confirm that object identification is a challenging task for retrievers, while `CLIP`'s performance stands out as relatively strong, when taking into account the size gap to the other models (427M vs 2B). Notably, ambiguous cases suffer from poor performance even without textual queries, likely because these instances are more visually complex.[9] Figure 5b demonstrates performance changes when varying the input to the specialized models:

- Δ cropping: Unlike generalized models, specialized retrievers are consistently and positively impacted by cropping, especially for ambiguous, visually complex stimuli.

[8] Detailed performance of different input variants and performance trends across different top-k values are described in Section D.2.

[9] Mean visual saliency is 0.38 for ambiguous cases vs. 0.58 for non-ambiguous ones.



(a) OI performance **(b)** Effect of modifying the question/image **(c)** Recall vs saliency

Fig. 5: Stage ② object identification results for specialized models. (b) Accuracy change under input manipulations: Δ 'cropping' marks the change after cropping the target object; Δ 'text' adds the question for the multimodal models. (c) Recall@10 across visual saliency bins.

- Δ text: Surprisingly, providing bimodal retrievers (`VLM2Vec2`, `Qwen-3-VL`) with the original question does not improve performance, unlike in generalized models. This likely reflects limited linguistic grounding ability, as current multimodal retrievers are not explicitly optimized to align textual cues with precise region-level retrieval.

Cropping helps specialized models when targets are non-salient. Lastly, we revisit the role of visual complexity and popularity, and analyze how they affect the specialized models. Visual complexity and popularity only weakly affect the performance of specialized models (Spearman's $\rho = 0.05$ for popularity, $r_{pb} = 0.13$ for visual saliency) (described in Section D.2), yet further inspection of the relation between recall and visual saliency reveals a more subtle trend, as depicted in Figure 5c: the more non-salient a target object is, the more cropping helps.

Intermediate takeaways based on object identification (stage ②):

1. All VLMs struggle with object identification, a prerequisite for successful KI-VQA. `CLIP` appears relatively strong considering its small size;
2. Whereas VLMs do best when receiving rich multimodal input (full image and question), specialized image retrievers perform best when only receiving a cropped target object image;
3. Visual complexity plays less of a role than in stage ①, whereas target popularity acts as a critical bottleneck—affecting VLMs far more acutely than image retrievers.

6 Stage ③ Knowledge Retrieval and Reasoning

The final KI-VQA stage requires synthesizing an answer by **reasoning over knowledge**. We assess this stage through end-task performance, while providing

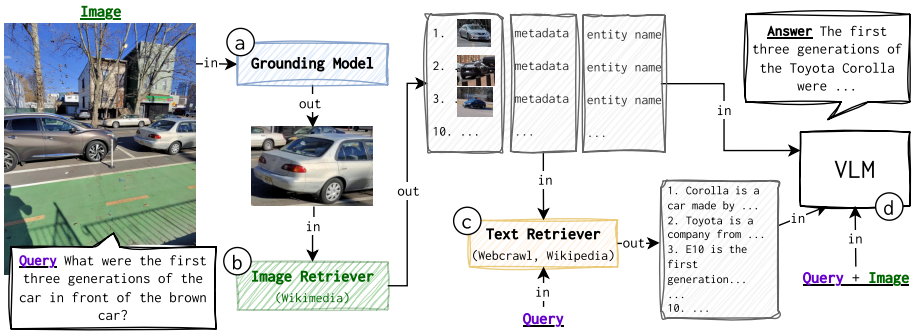


Fig. 6: Grounded bimodal RAG: In stage ③, we study a modeling pipeline that combines VLMs with RAG and visual grounding. Adding the grounding module is meant to make the image retriever more effective.

an analysis that isolates it from previous stages using 'text-only query' metadata. Below, we detail the experimental setup before elaborating on our findings.

6.1 Experimental Preliminaries

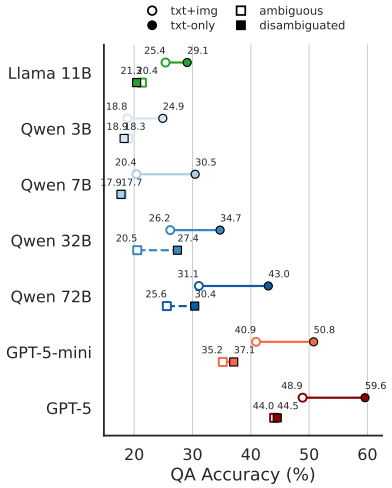
Models and RAG pipeline. We evaluate the same suite of **generalized** VLMs as in Section 4.1. These models are first tested in a zero-shot *baseline configuration*, relying solely on their parametric knowledge to answer the KI-VQA questions.

We then equip these VLMs with **RAG**, utilizing text and image retrievers as established in the original `CRAG-MM` framework [32]. Given an input image, the image retriever returns similar images and structured metadata from a knowledge graph using `CLIP-ViT-Large-Patch14-336` [29]. The text retriever similarly returns webpages given an input text, using documents embedded by `BGE-large-en-v1.5` [37]. We employ the original `CRAG-MM` text and image indices.

Standard RAG can be limited by imperfect image retrieval (as observed in Section 5). Informed by Sections 4 and 5, we evaluate a grounded bimodal RAG configuration (illustrated in Figure 6) that integrates an upstream visual grounding module to improve retrieval similar to region-aware retrieval [9,11,15].[10] This pipeline follows these steps:

ⓐ *Language-based visual grounding*: `Grounding-DINO` generates a precise crop of the target region to eliminate visual noise.
ⓑ *Image retrieval*: The retriever picks images and metadata for cropped target.
ⓒ *Text retrieval*: The text retriever fetches passages using the query and the metadata from the visual retrieval step.
ⓓ *Multimodal reasoning*: The VLM generates the final answer by synthesizing the query, the original image, and the retrieved context.

[10] Direct empirical comparison to related frameworks, such as [9, 11, 15], is precluded by unavailable public codebases or incomplete implementation details; however, they represent compelling directions for future benchmarking.



(a) Generalized models

	Model			*All*		*Subset*	
	ground.	img ret.	txt ret.	txt+img	txt-only	ambig.	dis.
Qwen 32B	✗	✗	✗	26.2	34.7	20.5	27.4
	✗	✓	✗	27.2	-	22.7	27.5
	✗	✗	✓	31.0	57.2	25.9	**33.6**
	✗	✓	✓	34.4	-	24.5	29.6
	✓	✓	✓	**34.7**	-	**28.0**	29.9
	GT	✓	✓	36.1	-	32.8	33.6
GPT-5	✗	✗	✗	48.9	59.6	44.0	44.5
	✗	✓	✗	61.5	-	50.9	53.9
	✗	✗	✓	56.7	73.3	53.1	**57.1**
	✗	✓	✓	61.5	-	46.4	55.2
	✓	✓	✓	**62.2**	-	**53.9**	56.5
	GT	✓	✓	62.2	-	55.7	58.4

(b) Ablations on grounded RAG components

Fig. 7: Comparison of (a) generalized models and (b) RAG pipeline ablations on KI-VQA accuracy. *All* reports performance over all instances; *Subset* evaluates ambiguous (*ambig.*) vs. disambiguated (*dis.*) queries. Columns *ground.*, *img ret.*, and *txt ret.* denote the activation of grounding, image retrieval, and text retrieval modules. *txt-only* evaluates questions with gold entity names replacing image, while *txt+img* uses the standard VQA setup. Shaded rows indicate use of ground-truth (GT) target regions.

Evaluation metric. Following [32], we evaluate QA accuracy using `gpt-4o-mini` as a binary classifier given the question, ground-truth, and prediction (prompt in Section F).[11] Reported scores are averaged over three runs of the judge.

6.2 Analyses and Findings

The end-task accuracies in Figure 7 underscore the difficulty of out-of-the-box KI-VQA, even for proprietary models (see 'txt+img' in Figure 7a). Although ambiguous expressions ('ambiguous' marker) degrade performance, the modest 1.8-point gain from disambiguation suggests ambiguity is not the primary bottleneck.

Retrieval/reasoning is the primary KI-VQA bottleneck. Using the 'text-only questions' from `CRAG-MM-Diagnostics`—where target entity names are explicitly mentioned—decouples retrieval and reasoning from stage ① and ② localization errors. Figure 7a ('txt-only' marker) shows that while this improves results by 8.7 points on average, the majority of errors persist. Notably, `GPT-5` improves most, with 20.9% of errors disappearing. This underscores that parametric knowledge

[11] [32] reports a 99.1% human agreement rate for this method.

is insufficient for most KI-VQA tasks, particularly for smaller models with limited capacity.

Unimodal RAG yields modest gains. Figure 7b demonstrates the impact of coupling VLMs with image and text retrievers. Image retrieval provides a consistent 6.8-point average boost for `Qwen` and `GPT-5`. Additionally, text retrieval alone adds 6.3 points; however, its effectiveness is limited by uninformative queries (e.g., "Is this sold in South America?" in Figure 1). By using 'text-only' questions with explicit entity names (replacing "white vehicle" with "Tesla Model Y" in "How does the range of the white vehicle compare to that of the hyundai ioniq 5?"), performance jumps by 16.6 and 26.2 points for `GPT-5` and `Qwen`, respectively. These results emphasize that accurate query reformulation and object identification are critical for maximizing retrieval gains.

Bimodal and grounded RAG yield top scores. Combining bimodal retrieval with visual grounding achieves our highest performance (Figure 7b), confirming that 'cleaning' visual inputs affects the KI-VQA positively. Using ground-truth regions (`GT`, highlighted in gray) provides an additional 1.4-point gain for `Qwen`, yet performance still trails the text-only retrieval baseline. Since image retrieval is implicitly intended to serve object identification, this reinforces that it does not succeed in that task, even when given a ground-truth target region. The error analysis in Section G shows that 25% of failures stem from poor object identification even with `GPT-5`. This highlights the need for more robust recognizers handling fine-grained taxonomies and effective retrieval to bridge parametric knowledge gaps.

Intermediate takeaways from the knowledge retrieval and reasoning stage:

1. Even proprietary models like `GPT-5` see substantial gains when target entity names are explicitly provided, indicating that internal knowledge cannot substitute for effective retrieval in KI-VQA;
2. Integrating a visual grounding module to crop target objects before retrieval consistently yields the highest performance, demonstrating that 'cleaning' the visual input is vital for accurate multimodal RAG;
3. Even with ground-truth target regions, performance lags behind text-only retrieval baselines, revealing that current retrieval pipelines still fail at fine-grained object identification and reasoning.

7 Conclusion

We introduced `CRAG-MM-Diagnostics`, a diagnostic benchmark designed for the stage-wise evaluation of KI-VQA under realistic visual conditions. By augmenting `CRAG-MM` with fine-grained annotations—including bounding boxes, referring expression types, disambiguated or text-only queries, and visual saliency metadata—we conducted a systematic analysis of the KI-VQA pipeline. Our

analysis spanned fully parametric VLMs, specialized grounding and retrieval models, and modular retrieval-augmented pipelines.

Our findings surface many patterns that explain why KI-VQA is complex, such as the visual complexity of the scene, the popularity of the target object, and the complexity of the referring expression directly correlating with performance in the various stages of the pipeline. We also uncovered failure modes (such as the surprising find that `GPT-5` struggles much more with explicit grounding than `Qwen`) and determined that for some stages, a small, specialized model can outperform larger, generalized VLMs. We end with the following three concrete lessons for the design of future KI-VQA systems:

1. **Linguistic ambiguity is a structural problem, not a minor nuisance.** Disambiguating referring expressions yields consistent gains across all stages of the pipeline. Future benchmarks should audit and annotate query ambiguity, and models should explicitly address it as ignoring ambiguity will lead to consistent underestimation of models' reasoning capacity. We advocate for models that explicitly detect and resolve ambiguity—potentially through iterative query reformulation—prior to knowledge retrieval.
2. **Object identification should inform text retrieval.** Text retrieval with original, image-dependent queries is often uninformative when the textual question alone does not identify the target (e.g., "Is this car sold in South America?"). Using the target entity name in the retrieval query—derived from either model prediction or early-stage identification—yields the largest performance gains observed in our study. This 'identify first, then retrieve' ordering is the most impactful design principle we uncover.
3. **Knowledge retrieval and reasoning is the dominant bottleneck.** Critically, even when models are provided the ground-truth entity name, the majority of errors persist, indicating that limited parametric knowledge and imperfect reasoning—not visual grounding or object identification—account for most KI-VQA failures. This suggests that the field should invest more in improving retrieval quality and multi-hop reasoning over retrieved evidence.

By providing a foundation for *stage-aware* KI-VQA evaluation, `CRAG-MM-Diagnostics` encourages a shift from 'black-box' testing to principled, modular assessment. We hope this benchmark serves as a catalyst for developing multimodal knowledge assistants that can truly navigate the intersection of visual perception and world knowledge.

Acknowledgements

We thank the reviewers for their valuable suggestions. We thank our colleagues from McGillNLP and Mila, especially Marius Mosbach, Vaibhav Adlakha, Rabiul Awal, and Aishwarya Agrawal. PB is supported by the RBC Borealis AI Global Fellowship Award and the ServiceNow-Mitacs Accelerate program. SR is supported by the Canada CIFAR AI Chair, the NSERC Discovery Grant and the Sloan Fellowship. The project is partly funded by the IVADO R3AI program. VD acknowledges the support of the IVADO Postdoctoral Research Funding.

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). p. 2425–2433. ICCV '15, IEEE Computer Society, USA (2015). https://doi.org/10.1109/ICCV.2015.279, https://doi.org/10.1109/ICCV.2015.279 3
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint (2025), https://arxiv.org/abs/2502.13923 1, 6, 7
3. Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 14948–14968. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.925, https://aclanthology.org/2023.emnlp-main.925/ 2, 3, 4
4. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., Lu, J., Anderson, T., Bransom, E., Ehsani, K., Ngo, H., Chen, Y., Patel, A., Yatskar, M., Callison-Burch, C., Head, A., Hendrix, R., Bastani, F., VanderBilt, E., Lambert, N., Chou, Y., Chheda, A., Sparks, J., Skjonsberg, S., Schmitz, M., Sarnat, A., Bischoff, B., Walsh, P., Newell, C., Wolters, P., Gupta, T., Zeng, K.H., Borchardt, J., Groeneveld, D., Nam, C., Lebrecht, S., Wittlif, C., Schoenick, C., Michel, O., Krishna, R., Weihs, L., Smith, N.A., Hajishirzi, H., Girshick, R., Farhadi, A., Kembhavi, A.: Molmo and PixMo: Open weights and open data for state-of-the-art vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 91–104 (June 2025), https://openaccess.thecvf.com/content/CVPR2025/papers/Deitke_Molmo_and_PixMo_Open_Weights_and_Open_Data_for_State-of-the-Art_CVPR_2025_paper.pdf 1
5. Ding, Y., Ren, K., Huang, J., Luo, S., Han, S.C.: MMVQA: A comprehensive dataset for investigating multipage multimodal information retrieval in pdf-based visual question answering. In: Larson, K. (ed.) Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24. pp. 6243–6251. International Joint Conferences on Artificial Intelligence Organization (8 2024). https://doi.org/10.24963/ijcai.2024/690, https://doi.org/10.24963/ijcai.2024/690, main Track 9
6. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14375–14385 (June 2024), https://openaccess.thecvf.com/content/CVPR2024/papers/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.pdf 4
7. Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., Chang, M.W.: Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In: Proceedings of the IEEE/CVF International Con-

ference on Computer Vision (ICCV). pp. 12065–12075 (October 2023), https://openaccess.thecvf.com/content/ICCV2023/papers/Hu_Open-domain_Visual_Entity_Recognition_Towards_Recognizing_Millions_of_Wikipedia_Entities_ICCV_2023_paper.pdf 4, 10

8. Hu, W., Gu, J.C., Dou, Z.Y., Fayyaz, M., Lu, P., Chang, K.W., Peng, N.V.: MRAG-Bench: Vision-centric evaluation for retrieval-augmented multimodal models. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) International Conference on Learning Representations. vol. 2025, pp. 95558–95581 (2025), https://proceedings.iclr.cc/paper_files/paper/2025/file/ee46288ab2aaf5c6e53aebebe719712c-Paper-Conference.pdf 2, 4
9. Jian, P., Yu, D., Zhang, J.: Large language models know what is key visual entity: An LLM-assisted multimodal retrieval for VQA. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 10939–10956. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.613, https://aclanthology.org/2024.emnlp-main.613/ 12
10. Kabir, R., Haque, N., Islam, M.S., et al.: A comprehensive survey on visual question answering datasets and algorithms. arXiv preprint (2024), https://api.semanticscholar.org/CorpusID:274131862 2
11. Kim, J., Tao, R., Sharma, S., Wang, J., Sun, K., Lin, Z., Moon, S., Mathias, L., Kumar, A., Ji, H., et al.: Pixel-grounded retrieval for knowledgeable large multimodal models. arXiv preprint (2026), https://arxiv.org/abs/2601.19060 12
12. Lerner, P., Ferret, O., Guinaudeau, C., Le Borgne, H., Besançon, R., Moreno, J.G., Lovón Melgarejo, J.: ViQuAE, a dataset for knowledge-based visual question answering about named entities. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 3108–3120. SIGIR '22, Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3477495.3531753, https://doi.org/10.1145/3477495.3531753 3
13. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint (2026), https://arxiv.org/abs/2601.04720 9
14. Li, Y., Li, Y., Wang, X., Jiang, Y., Zhang, Z., Zheng, X., Wang, H., Zheng, H.T., Huang, F., Zhou, J., Yu, P.S.: Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=VvDEuyVXkG 3
15. Lin, L., Xie, Y., Chen, D., Xu, Y., Zhu, C., Yuan, L.: REVIVE: Regional visual representation matters in knowledge-based visual question answering. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=wwyiEyK-G5D 12
16. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L.: Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVII. p. 38–55. Springer-Verlag, Berlin, Heidelberg (2024). https://doi.org/10.1007/978-3-031-72970-6_3, https://doi.org/10.1007/978-3-031-72970-6_3 5, 6

17. Ma, T., Bai, B., Lin, H., Wang, H., Wang, Y., Luo, L., Fang, L.: When visual grounding meets gigapixel-level large-scale scenes: Benchmark and approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22119–22128 (June 2024), https://openaccess.thecvf.com/content/CVPR2024/papers/Ma_When_Visual_Grounding_Meets_Gigapixel-level_Large-scale_Scenes_Benchmark_and_Approach_CVPR_2024_paper.pdf 4, 6, 8
18. Ma, X., Ding, Z., Luo, Z., Chen, C., Guo, Z., Wong, D.F., Feng, X., Sun, M.: DeepPerception: Advancing r1-like cognitive visual perception in MLLMs for knowledge-intensive visual grounding. arXiv preprint (2025), https://arxiv.org/abs/2503.12797 4
19. Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D., Hajishirzi, H.: When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9802–9822. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.546, https://aclanthology.org/2023.acl-long.546/ 5
20. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: OK-VQA: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019), https://openaccess.thecvf.com/content_CVPR_2019/papers/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.pdf 2, 3, 4
21. Meng, R., Jiang, Z., Liu, Y., Su, M., Yang, X., Fu, Y., Qin, C., Thirukovalluru, R., Zhang, X., Chen, Z., Xu, R., Xiong, C., Zhou, Y., Chen, W., Yavuz, S.: VLM2vec-v2: Advancing multimodal embedding for videos, images, and visual documents. Transactions on Machine Learning Research (2026), https://openreview.net/forum?id=TpU38jbKIJ 9
22. Mensink, T., Uijlings, J., Castrejon, L., Goel, A., Cadar, F., Zhou, H., Sha, F., Araujo, A., Ferrari, V.: Encyclopedic VQA: Visual questions about detailed properties of fine-grained categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 3113–3124 (October 2023), https://openaccess.thecvf.com/content/ICCV2023/papers/Mensink_Encyclopedic_VQA_Visual_Questions_About_Detailed_Properties_of_Fine-Grained_Categories_ICCV_2023_paper.pdf 2, 3, 4, 10
23. Meta, A.: Llama 3.2: Revolutionizing edge AI and vision with open, customizable models (2024), https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ 6
24. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection. In: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X. p. 728–755. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-20080-9_42, https://doi.org/10.1007/978-3-031-20080-9_42 6
25. Narayan, K., Xu, Y., Cao, T., Nerella, K., Patel, V.M., Shiee, N., Grasch, P., Jia, C., Yang, Y., Gan, Z.: DeepMMSearch-R1: Empowering multimodal llms in multimodal web search. arXiv preprint (2025), https://arxiv.org/abs/2510.12801 1, 2

26. Qiang, C., Wei, Z., Han, X., Wang, Z., Li, S., Lan, X., Jiao, J., Han, Z.: VER-Bench: Evaluating MLLMs on reasoning with fine-grained visual evidence. In: Proceedings of the 33rd ACM International Conference on Multimedia. p. 12698–12705. MM '25, Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3746027.3758208, https://doi.org/10.1145/3746027.3758208 2
27. Qiao, Y., Duan, H., Fang, X., Yang, J., Chen, L., Zhang, S., Wang, J., Lin, D., Chen, K.: Prism: A framework for decoupling and assessing the capabilities of VLMs. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=qLnXPVvwLx 4
28. Qiu, J., Madotto, A., Lin, Z., Crook, P.A., Xu, Y.E., Damavandi, B., Dong, X.L., Faloutsos, C., Li, L., Moon, S.: SnapNTell: Enhancing entity-centric visual question answering with retrieval augmented multimodal LLM. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 247–266. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.14, https://aclanthology.org/2024.findings-emnlp.14/ 3, 4, 5, 10
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning, ICML 2021. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (2021), http://proceedings.mlr.press/v139/radford21a.html 9, 12
30. Singh, P., Dhawan, S., Agarwal, S., Thakur, N.: Implementation of an efficient fuzzy logic based information retrieval system. arXiv preprint (2015), https://arxiv.org/abs/1503.03957 9
31. Tran, D.T., Tran, T.K., Hauswirth, M., Le Phuoc, D.: ReasonVQA: A multi-hop reasoning benchmark with structural knowledge for visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18793–18803 (October 2025), https://openaccess.thecvf.com/content/ICCV2025/papers/Tran_ReasonVQA_A_Multi-hop_Reasoning_Benchmark_with_Structural_Knowledge_for_Visual_ICCV_2025_paper.pdf 3
32. Wang, J., Yang, X., Sun, K., Suresh, P., Sharma, S., Czyzewski, A., Andersen, D., Appini, S., Banerjee, A., Choudhary, S., et al.: CRAG-MM: Multi-modal multi-turn comprehensive rag benchmark. arXiv preprint (2025), https://arxiv.org/abs/2510.26160 1, 2, 3, 4, 9, 12, 13
33. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Henge, A.: Explicit knowledge-based reasoning for visual question answering. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence. p. 1290–1296. IJCAI'17, AAAI Press (2017), https://dl.acm.org/doi/10.5555/3171642.3171825 2
34. Wang, Y., Zhao, Y., Chen, X., Guo, S., Liu, L., Li, H., Xiao, Y., Zhang, J., Li, Q., Xu, K.: VisualSimpleQA: A benchmark for decoupled evaluation of large vision-language models in fact-seeking question answering. arXiv preprint (2025), https://arxiv.org/abs/2503.06492 4
35. Wu, P., Xie, S.: V?: Guided visual search as a core mechanism in multimodal LLMs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13084–13094 (June 2024), https://openaccess.thecvf.com/content/CVPR2024/papers/Wu_V_Guided_Visual_Search_as_a_Core_Mechanism_in_Multimodal_CVPR_2024_paper.pdf 2, 4, 8

36. Xiao, L., Yang, X., Lan, X., Wang, Y., Xu, C.: Toward visual grounding: A survey. IEEE Transactions on Pattern Analysis & Machine Intelligence **48**(03), 2749–2771 (Mar 2026). https://doi.org/10.1109/TPAMI.2025.3630635, https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3630635 6
37. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.Y.: C-pack: Packed resources for general Chinese embeddings. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 641–649. SIGIR '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3626772.3657878, https://doi.org/10.1145/3626772.3657878 12
38. Xu, Y., Zhu, L., Yang, Y.: MC-Bench: A benchmark for multi-context visual grounding in the era of MLLMs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 17675–17687 (October 2025), https://openaccess.thecvf.com/content/ICCV2025/papers/Xu_MC-Bench_A_Benchmark_for_Multi-Context_Visual_Grounding_in_the_Era_ICCV_2025_paper.pdf 4
39. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: MM-Vet: evaluating large multimodal models for integrated capabilities. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024), https://dl.acm.org/doi/10.5555/3692070.3694451 4
40. Zhang, Y., Zhang, H., Tian, H., Fu, C., Zhang, S., Wu, J., Li, F., Wang, K., Wen, Q., Zhang, Z., Wang, L., Jin, R.: MME-realworld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In: The Thirteenth International Conference on Learning Representations (2025), https://openreview.net/forum?id=k5VHHgsRbi 4