M^3-VQA: A Benchmark for Multimodal, Multi-Entity, Multi-Hop Visual Question Answering

Jiatong Ma[1,2]*, Longteng Guo[1]*, Yuchen Liu[1,2], Zijia Zhao[1,2], Dongze Hao[1,2], Xuanxu Lin[1,2], Jing Liu[1,2]†

[1] Institute of Automation, Chinese Academy of Sciences,
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences,
majiatong2025@ia.ac.cn, {longteng.guo,jliu}@nlpr.ia.ac.cn

Abstract

We present M^3-VQA, a novel knowledge-based Visual Question Answering (VQA) benchmark, to enhance the evaluation of multimodal large language models (MLLMs) in fine-grained multimodal entity understanding and complex multi-hop reasoning. Unlike existing VQA datasets that focus on coarse-grained categories and simple reasoning over single entities, M^3-VQA introduces diverse multi-entity questions involving multiple distinct entities from both visual and textual sources. It requires models to perform both sequential and parallel multi-hop reasoning across multiple documents, supported by traceable, detailed evidence and a curated multimodal knowledge base. We evaluate 16 leading MLLMs under three settings: without external knowledge, with gold evidence, and with retrieval-augmented input. The poor results reveal significant challenges for MLLMs in knowledge acquisition and reasoning. Models perform poorly without external information but improve markedly when provided with precise evidence. Furthermore, reasoning-aware agentic retrieval surpasses heuristic methods, highlighting the importance of structured reasoning for complex multimodal understanding. M^3-VQA presents a more challenging evaluation for advancing the multimodal reasoning capabilities of MLLMs. Our code and dataset are available at https://github.com/CASIA-IVA-Lab/M3VQA.

1 Introduction

With the rapid advancement of Multimodal Large Language Models (MLLMs), Visual Question Answering (VQA) has emerged as a widely used benchmark for evaluating open-ended multimodal understanding. These models have demonstrated impressive progress in surface-level perception and commonsense reasoning. However, real-world applications often require more than basic comprehension—they demand the ability to perform complex reasoning involving multiple, fine-grained visual entities. For example, a question might require identifying specific people, brands, or animal species within an image and then integrating their attribute and relationship information through multi-entity, multi-hop reasoning to arrive at the correct answer.

Existing knowledge-based VQA benchmarks, such as OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022), have advanced the field by introducing questions that require external knowledge beyond the image content. Nonetheless, these datasets primarily focus on relatively coarse-grained, basic-level categories and entities, and typically emphasize reasoning over single entities. While recent datasets like S3VQA (Jain et al., 2021), ViQuAE (Lerner et al., 2022), KVQA (Shah et al., 2019), EVQA (Mensink et al., 2023), InfoSeek (Chen et al., 2023) and Dyn-VQA (Li et al., 2024b) have attempted to incorporate finer-grained factual knowledge, they mostly address relatively simple scenarios. The questions are often answerable with only one or two reasoning steps and involve a single primary entity, limiting their ability to fully challenge and evaluate the complex multimodal reasoning skills of modern MLLMs.

To address these limitations, we introduce M^3-VQA, a novel and challenging benchmark designed to significantly advance the evaluation of knowledge-based Visual Question Answering. M^3-VQA specifically targets the model's capability in fine-grained, multimodal entity understanding and sophisticated multi-hop reasoning. Our benchmark introduces three key features:

(1) Diverse Multi-Entity Questions: M^3-VQA incorporates a variety of fine-grained named entities—including architectural landmarks, individual person names, brands, and animal species—sourced from both visual scenes and tex-

*Equal Contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 40633–40669
July 2-7, 2026

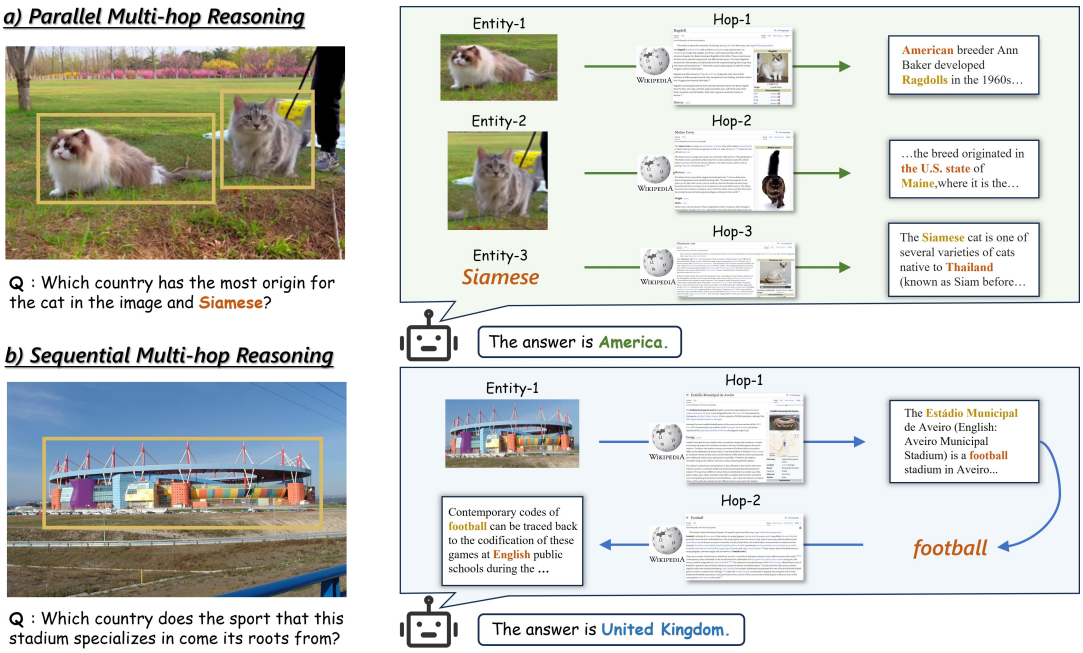

Figure 1: Examples of parallel and sequential multi-hop reasoning in M^3-VQA.

tual information. Each question in the dataset involves multiple distinct entities, closely reflecting realistic multimodal interactions.

(2) Complex Multi-Hop Reasoning: Questions in M^3-VQA require models to retrieve and integrate information from multiple documents or external knowledge sources, promoting extensive cross-modal and cross-document reasoning rather than reliance on single-step inference. Crucially, the dataset encompasses both parallel reasoning tasks—where multiple entities are independently analyzed before synthesizing an answer—and sequential reasoning tasks—where entities form a chain that models must sequentially traverse to reach the solution.

(3) Traceable Supporting Evidence: Each reasoning step in M^3-VQA is explicitly supported by detailed, grounded evidence, allowing transparent traceability of model reasoning processes. Additionally, we provide a carefully curated multimodal knowledge base to support retrieval-augmented evaluation protocol.

This comprehensive design enables rigorous evaluation of MLLMs, both in standalone settings and scenarios enhanced by retrieval mechanisms, effectively pushing the boundaries of real-world multimodal understanding and reasoning.

We evaluate 16 leading MLLMs under three settings: without evidence, with gold evidence, and with retrieval from an external knowledge base. Our experiments reveal the following key findings:

- **MLLMs Perform Poorly without External Knowledge:** When restricted to only the image and question, models consistently underperform, with a maximum accuracy of 32.6%. This reveals a fundamental limitation of current MLLMs in acquiring and applying background knowledge solely from their internal representations.

- **Precise Evidence Significantly Boosts Reasoning Accuracy:** When gold supporting evidence is provided, model performance improves substantially. This indicates that even advanced MLLMs remain heavily reliant on well-structured external information to support complex multi-entity reasoning.

- **Reasoning-Aware Agentic Retrieval Outperforms Heuristic Approach:** While retrieval-augmented approaches boost performance, we find that agentic retrieval—featuring explicit reasoning and iterative planning—outperforms heuristic retrieval. This underscores the importance of structured reasoning strategies in tackling multi-entity, multi-hop questions.

2 Related Work

Real-world VQA often requires external knowledge, leading to growing interest in retrieval for VQA. Early works like KB-VQA (Wang et al.,

	OKVQA	S3VQA	ViQuAE	KVQA	InfoSeek	EVQA	Dyn-VQA	M^3-VQA
Multi-Entity	×	×	×	✓✓	×	×	✓	✓✓
Fine-Grained Entity	×	✓	✓✓	✓✓	✓✓	✓✓	✓✓	✓✓
Entity Diversity	✓✓	✓✓	✓✓	×	✓✓	✓	✓✓	✓✓
Multi-Hop	×	×	×	✓✓	×	✓	✓	✓✓
Reasoning Mode								
↪Sequential	×	×	×	✓	×	✓	✓	✓✓
↪Parallel	×	×	×	✓✓	×	×	✓	✓✓
Controlled Knowledge Base								
↪Traceable Answer	×	×	✓	×	×	✓	×	✓✓
↪Free-Form KB	×	×	✓✓	×	✓✓	✓✓	×	✓✓
Answer Type	Open	Open	Open	Open	Open	Open	Open	Open
#{I, Q, A}	14K	7K	3K	183K	1M	1M	1K	13K

Table 1: Comparison between our M^3-VQA and previous VQA datasets.

2015), FVQA (Wang et al., 2017), and KVQA (Shah et al., 2019) relied on structured knowledge bases, which limited the scope of information. In contrast, we aim to build a knowledge base from free-form text and images (e.g., Wikipedia). KVQA also focuses solely on facial recognition, lacking broad entity coverage. Datasets such as OKVQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022) primarily require commonsense knowledge and rarely involve detailed attributes of fine-grained entities, thus often eliminating the need for retrieval. While S3VQA (Jain et al., 2021) and ViQuAE (Lerner et al., 2022) address fine-grained attributes, S3VQA lacks a knowledge base, and ViQuAE is small (3.6k samples). Recent datasets like EVQA (Mensink et al., 2023) and Infoseek (Chen et al., 2023) do include fine-grained entity attributes and scale to millions of examples, solving the size issue. However, they still focus on single-entity images and questions that can be answered within two retrieval hops, failing to capture the complexity of multi-entity interaction and multi-hop reasoning in real-world scenarios. Dyn-VQA (Li et al., 2024b) recently introduced multi-hop questions but is very small (only 387 samples) and lacks a knowledge base. Our proposed M^3-VQA advances prior work in multiple dimensions: (1) It uniquely integrates multimodal inputs, multi-entity interactions, and multi-hop reasoning, offering a more challenging comprehensive evaluation environment. (2) It includes a dedicated knowledge base and gold evidence annotations to support each answer, enabling precise assessment of reasoning capabilities. See Appendix A for more.

3 M^3-VQA Task Definition

The M^3-VQA task can be formulated as a function $f : (I, Q, K) \rightarrow A$, where I denotes an input image, Q represents a question, K is an external Wikipedia knowledge base, and A is the answer. Specifically, each image I may contain multiple visual entities. Each question Q is expressed in text and may comprise multiple textual entities or sub-questions. The external Wikipedia knowledge base K includes multimodal pages with both text and images. The answer A can take various forms, including strings (words or phrases) and temporal values (e.g., years or dates). For each sample, there exists a reference evidence chain E that serves as a guide for the reasoning process. We assess the difficulty of the task from two key dimensions: **(1) Entity Complexity:** defined as the total number of visual and textual entities involved in Q and I. **(2) Hop Complexity:** measured by the number of nodes in the reference evidence chain E required to answer the question Q.

Specifically, we divided M^3-VQA task into 2 sub-task based on the reasoning pattern of multi-hop evidence chain: **(1) Parallel Multi-hop Reasoning** refers to a multi-hop inference process conducted in a parallel manner, where multiple entities within the image and question are reasoned about simultaneously. The final answer is derived from the integrated understanding of the individual results obtained from reasoning over each entity in parallel. For example, in the first image of Figure 1, we can independently retrieve information about the two cats in the image and the Siamese mentioned in the question to determine their countries of origin, and then identify which country appears most frequently. **(2) Sequential Multi-hop Reasoning** refers to a multi-hop inference process structured in a serial manner, where the input of each subsequent hop is derived from the output of the previous hop. The final result is obtained from the end of this sequential chain. For instance, in the second image of Figure 1, one must first retrieve information about

the stadium to determine which sport it is used for and, subsequently, retrieve information about that sport before determining its country of origin.

4 M^3-VQA Dataset

4.1 Dataset Construction

Image and Entity Annotation. Our dataset construction begins with collecting images containing multiple fine-grained entities, leveraging existing image datasets that already provide detailed entity annotations. Specifically, we utilize fine-grained image datasets such as CelebTo (Zhong et al., 2018), FGVD (Khoba et al., 2022), FlickrLogos-47 (Romberg et al., 2011), LogosInTheWild (Tüzkö et al., 2017), Menu-match (Beijbom et al., 2015), Oktoberfest (Ziller et al., 2019), UEC-FoodPix (Okamoto and Yanai, 2021) and UNIMIB2016 (Ciocca et al., 2016), which include precise labels for person, vehicle, logo, and food within images. These original entity annotations form the foundation for our benchmark, ensuring reliable multi-entity grounding. To further increase dataset diversity, we also incorporate the single-entity, single-hop images from knowledge-based VQA datasets such as EVQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023), and systematically extend their questions to multi-hop settings. Additional images were manually collected by human annotators via web search engines to supplement fine-grained multi-entity scenes.

Entity Linking and Expansion. Starting from these images and their entity annotations, we map each entity label to a corresponding Wikidata identifier (Vrandečić and Krötzsch, 2014) through automatic string matching. For ambiguous or uncertain mappings, human annotators performed manual verification based on image-label pairs to ensure accurate entity linking. Leveraging Wikidata IDs, we retrieved corresponding Wikipedia articles, enabling multimodal grounding between images, structured Wikidata knowledge, and multimodal Wikipedia pages. This linkage facilitates transparent reasoning and retrieval of rich external knowledge for each entity.

Parallel Multi-Hop Question Generation. To generate questions requiring parallel multi-hop reasoning, we exploited Wikidata's structured knowledge triples (subject, relation, object). Human annotators designed natural language question templates for diverse relation types, employing placeholders for high-level visual categories (e.g., "person", "brand"). We used SPARQL queries to extract answers from Wikidata which were processed through the question templates to form final answers. Images were then paired with these question-answer pairs to construct IQA triplets. Subsequently, additional textual entities of the same category were incorporated into questions to increase complexity, with corresponding sub-answers retrieved and aggregated into the final answers. This process resulted in questions that require models to independently reason about multiple entities before synthesizing a response, supporting robust evaluation of parallel multi-hop capabilities. To increase linguistic variety, some template-based questions were paraphrased with LLMs.

Sequential Multi-Hop Question Generation. For sequential multi-hop reasoning questions, we adopted the concept of "bridging entities" (Yang et al., 2018), where the answer to one sub-question serves as the basis for a subsequent question. We automatically identified bridging entities within answers, filtering out non-informative terms. Leveraging large language models (LLMs), we generated candidate follow-up questions linked to these bridging entities. These candidates were validated by ensuring their answers matched the original grounding information, with invalid or ambiguous questions discarded. Complex multi-hop questions were constructed by chaining multiple sequential reasoning steps.

Knowledge Base and Evidence Annotation. We constructed a multimodal knowledge base by extracting Wikipedia pages—including both textual content and images—from a snapshot comprising approximately 2 million entities, to facilitate retrieval-augmented evaluation. Each sub-question is linked to its corresponding Wikipedia article, where the exact sentence containing the answer was identified and annotated as gold evidence. Additionally, the section indices of supporting sentences were recorded to enable precise traceability of the reasoning process.

Quality Control. To maintain high data quality, we employed a strict annotation protocol involving trained human annotators who underwent screening before participating. Quality control measures included multiple rounds of filtering to exclude low-quality images, irrelevant or overly generic entities, and questions with insufficient length or ambiguous answers. We ensured a balanced distribution across question types and reasoning complexities through strategic sub-sampling. The final

Statistic		Number
(I, Q, A) triplets		13125
Unique I		10565 (80.5%)
Unique Q		7611 (58.0%)
Unique entity		4645
Average length of Q		14.2
Average #hop		3.1
Average #entity		2.8
Average #Evidence		3.1
#Hop	1	1092 (8.3%)
	2	3181 (24.2%)
	3	4546 (34.6%)
	4+	4306 (32.8%)
#Entity	1	2992 (22.8%)
	2	2183 (16.6%)
	3	3644 (27.8%)
	4+	4306 (32.8%)

Table 2: Statistics of M^3-VQA.

M^3-VQA dataset offers a diverse and challenging benchmark tailored to advanced multimodal, multi-entity, multi-hop reasoning in knowledge-based VQA. See Appendix D for more.

4.2 Dataset Statistics

As shown in Tables 2, our dataset contains 13k (I, Q, A) triples, 10,565 unique images, and 7,611 unique questions. The ratio of unique images is 80%, and the ratio of unique questions is 58%, demonstrating the diversity of the dataset. In terms of hop count, the dataset includes 1,092 single-hop questions, 3,181 two-hop questions, 4,546 three-hop questions, and 4,306 questions with four or more hops. In terms of the number of entities, the dataset includes 2,992 single-entity questions, 2,183 two-entity questions, 3,644 three-entity questions, and 4,306 questions with four or more entities. Table 1 also presents a comparative analysis between M^3-VQA and other knowledge-based VQA datasets. The dataset includes numerous visual entities covering a wide range of categories such as person, animal, plant, vehicle, food, logo, building, and landmark. See Appendix E for more.

4.3 Evaluation Metrics

Given the prevalence of multi-answer questions in the dataset, we follow EVQA (Mensink et al., 2023) to use the Intersection over Union (IoU) between the predicted and ground-truth answer sets as the accuracy for each question. The final score is the arithmetic mean across all questions. Following previous work (Goyal et al., 2017; Marino et al., 2019; Chen et al., 2023), we apply exact match using pre-collected answer variations for string-based answers, and we use a relaxed match allowing a one-year margin for time-based questions, considering the estimation nature of historical dates.

5 Experiments

5.1 Evaluation Settings

We benchmark a wide range of models, including advanced closed-source model GPT-4o (Hurst et al., 2024), and open-source models such as Qwen2.5-VL-72B-Instruct (Bai et al., 2025). To compare visual models from different providers, we evaluate 7B/8B versions of Qwen2.5-VL, Qwen2-VL (Wang et al., 2024), InternVL2.5 (Chen et al., 2024), Llava-OneVision (Li et al., 2024a), DeepSeek-VL2 (Wu et al., 2024) and MiniCPM-V-2.6 (Yao et al., 2024). To study the impact of model size, we test the 3B, 7B, 32B, and 72B versions of Qwen2.5-VL, as well as the 4B, 8B, 26B, 38B, and 78B versions of InternVL2.5. We also evaluate pure language models (e.g., LLaMA-3.1 (Grattafiori et al., 2024), Qwen2.5 (Yang et al., 2024)) by replacing image inputs with textual descriptions.

We report results under three different settings: *original results*, *oracle results*, and *KB results*. The *original results* correspond to answers generated directly based on the image and question. The *oracle results* are obtained by providing additional evidence at varying granularities. The *KB results* involve supplying a knowledge base, from which the model retrieves relevant evidence using a retrieval method for reference.

5.2 Original Results

Table 3 presents the original experimental results, in which only the question and the image are provided as input to the MLLMs. We draw the following conclusions: M^3-VQA presents a significant challenge to current MLLMs. Even the best-performing model, Qwen2.5-VL-72B-Instruct, achieves only 32.6% accuracy. Larger model sizes improve performance. Comparing the 3B, 7B, 32B, and 72B versions of Qwen2.5-VL, as well as the 4B, 8B, 26B, 38B, and 78B versions of InternVL2.5, we observe consistent performance gains as parameter count increases. Pure language models perform worse than vision-language models across all settings. This indicates that image descriptions alone are insufficient to capture all visual information, especially with fine-grained en-

Model	Hop				Entity				All
	1	2	3	4+	1	2	3	4+	
Qwen2.5-VL-3B-Instruct (Bai et al., 2025)	37.0	17.4	17.6	16.0	24.5	18.6	17.0	16.0	18.7
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	34.9	20.7	21.6	21.7	24.1	24.3	21.1	21.7	22.5
Qwen2.5-VL-32B-Instruct (Bai et al., 2025)	46.8	24.9	29.8	25.6	30.1	29.4	30.7	25.6	28.7
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	47.9	**29.5**	**34.2**	**29.2**	30.8	**35.5**	**36.2**	**29.2**	**32.6**
Qwen2-VL-7B-Instruct (Wang et al., 2024)	28.7	17.6	20.5	24.3	20.5	20.9	20.1	24.3	21.7
InternVL2.5-4B (Chen et al., 2024)	35.0	13.3	18.0	19.7	18.4	17.0	19.2	19.7	18.8
InternVL2.5-8B (Chen et al., 2024)	26.1	16.9	19.4	17.4	18.5	19.6	19.9	17.4	18.7
InternVL2.5-26B (Chen et al., 2024)	27.4	13.8	15.2	17.0	17.9	15.4	15.3	17.0	16.5
InternVL2.5-38B (Chen et al., 2024)	42.2	24.4	30.6	26.0	28.7	28.9	31.2	26.0	28.6
InternVL2.5-78B (Chen et al., 2024)	43.5	27.2	33.4	29.1	29.8	32.5	34.4	29.1	31.3
LLaVA-OneVision-7B (Li et al., 2024a)	35.0	17.5	21.6	22.5	21.3	20.9	22.7	22.5	22.0
DeepSeek-VL2 (4.5B) (Wu et al., 2024)	40.2	20.3	24.9	22.7	24.6	24.0	26.2	22.7	24.3
MiniCPM-V-2.6 (8B) (Yao et al., 2024)	31.5	19.1	23.1	25.1	19.2	23.1	25.3	25.1	23.5
GPT-4o * (Hurst et al., 2024)	**51.6**	23.3	24.3	27.9	**34.4**	25.9	22.3	27.9	27.5
Question and Image Description as Input									
Llama-3.1-8B-Instruct (*text*) (Grattafiori et al., 2024)	26.2	15.9	21.7	21.5	18.6	19.3	21.9	21.5	20.6
Qwen2.5-7B-Instruct (*text*) (Yang et al., 2024)	24.0	12.9	18.7	17.4	16.7	16.1	18.4	17.4	17.3

Table 3: Performance comparison of various models under Original setting. * indicates that the model may have a lower accuracy due to its tendency to refuse to answer human queries. See Appendix H for more model results.

tities.

To further validate the multimodal nature of M^3-VQA, we also test the model's performance when the image is removed (only the textual question is given, abbr. *Q-Only*). As shown in Table 4, a 8.9% gap between *Original* and *Q-Only* shows that image input is essential. The 15.6% accuracy of *Q-Only* is partly due to correct answers based on text entity properties and our IoU-based metric, which gives partial credit. Models may also guess correctly on factual or counting questions.

5.3 Oracle Results

In the oracle setting, the model is provided with a pre-annotated golden evidence chain. This setting primarily evaluates the model's ability to perform reasoning using multi-hop evidence. We categorize the evidence into three granularities: *Sentence*, which corresponds to the most precise sentence in the linked Wikipedia page which supports the answer; *Section*, referring to the Wikipedia section containing the evidence; and *Entity-Name*, indicating the entity name associated with the evidence.

As shown in Table 4, even under the best-case *Sentence* setting, the top-performing model achieves only 58.7% accuracy, with an average of 49.9%. This highlights the difficulty current models have with multi-hop and multi-entity reasoning. The 10.2% difference between *Entity-Name* and *Original* setting suggests that identifying fine-grained visual entities remains a major hurdle. Adding visual entity recognition shows great promise. Comparing *Sentence* and *Entity-Name* (both having access to entity identity), there's still a 15.2% gap, showing that recalling detailed attributes remains challenging. Retrieval helps significantly here. There's a 4.8% accuracy gap between *Sentence* and *Section*, showing that providing more precise local information boosts performance. Smaller supporting documents lead to more verifiable and explainable answers.

5.4 KB Results

In the knowledge-base (KB) setting, the model has access to a manually constructed knowledge base, in addition to the question and image. This setting allows us to evaluate the model's ability to perform multi-hop reasoning through information retrieval. We design two general retrieval algorithms to assist the tested MLLM in answering questions: *Heuristic Retrieval* and *Agentic Retrieval*.

Heuristic Retrieval. We implement a two-stage heuristic retrieval: *(1) Text Retrieval:* The question is used as a query to retrieve the most relevant text passages from the KB; *(2) Image Retrieval:* The image is used to retrieve similar visual entities from the KB. Once identified, the question is used to search within the corresponding Wikipedia articles for relevant text. The evidence retrieved from both text and image search is concatenated and passed as context to the model.

Agentic Retrieval. We developed agentic retrieval model composed of Planner, Executor, and Solver. The Executor includes tools such as object detection, text retrieval, and image retrieval. The Planner is responsible for devising the problem-solving steps and invoking the relevant tools. It first calls the object detection module within the

Model	Evidence Provided			No Evidence	
	Sentence	Section	Entity-Name	Original	Q-Only
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	49.0	42.7	31.8	22.5	14.4
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	58.4	52.4	**45.1**	**32.6**	16.1
InternVL2.5-8B (Chen et al., 2024)	46.8	42.1	25.7	18.7	11.8
InternVL2.5-78B (Chen et al., 2024)	**58.7**	**55.2**	45.0	31.3	14.7
LLaVA-OneVision-7B (Li et al., 2024a)	47.2	43.6	31.7	22.0	17.3
MiniCPM-V-2.6 (8B) (Yao et al., 2024)	46.9	39.1	33.6	23.5	**18.5**
Llama-3.1-8B-Instruct (*text*) (Grattafiori et al., 2024)	42.3	40.5	29.7	20.6	16.1
Average	49.9	45.1	34.7	24.5	15.6

Table 4: Performance comparison of various models under different settings.

Model	Original	Heuristic Retrieval	Agentic Retrieval
Qwen2.5-VL-7B-Instruct (Bai et al., 2025)	22.5	25.7	31.2
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	**32.6**	**36.6**	**38.9**
InternVL2.5-8B (Chen et al., 2024)	18.7	25.3	29.4
InternVL2.5-78B (Chen et al., 2024)	31.3	33.2	36.4
LLaVA-OneVision-7B (Li et al., 2024a)	22.0	25.8	28.9
MiniCPM-V-2.6 (8B) (Yao et al., 2024)	23.5	24.7	28.0
Llama-3.1-8B-Instruct (*text*) (Grattafiori et al., 2024)	20.6	23.4	28.3

Table 5: Performance comparison of various models under KB setting.

Executor to detect and segment the image. Then, it either uses the segmented image to perform image retrieval or generates single-hop queries for text retrieval. The Executor carries out the actual execution of these tool calls. Finally, the Solver produces the answer based on the retrieved information.

For heuristic retrieval and agentic retrieval, the tested MLLMs are used as the Solver; GPT-4o is used as the Planner; Qwen2.5-VL-7B-Instruct for object detection; BGE-Large-en-v1.5 (Xiao et al., 2024) for text embedding; and CLIP-ViT-Large (Radford et al., 2021) for image embedding. Table 5 presents the experimental results. We draw the following conclusions:

Heuristic retrieval methods suffer from overloaded queries. Using the full question as a single text query is problematic when the question contains multiple entities or sequential sub-questions. Similarly, using the entire image may hinder visual retrieval, especially when multiple entities are present. These one-shot strategies place excessive burden on a single query, often retrieving superficially relevant but unhelpful content. *Agentic retrieval* significantly outperforms all other models, including both open-source and closed-source MLLMs and LLM using heuristic retrieval. We attribute this to two factors: agentic retrieval decomposes complex questions into simpler sub-questions, reducing the burden of single-step retrieval; it rethinks both the content retrieved and the sub-questions, ensuring sufficient information is gathered.

M^3-VQA presents significant challenges to existing models in terms of multi-hop and multi-entity retrieval. A 20% gap between *agentic retrieval* and *Sentence* indicates the challenge of accurate retrieval. Accuracy with retrieved evidence generally falls between the *Original* and *Oracle* settings. This confirms the practical effectiveness of multi-modal, multi-hop, and multi-entity retrieval. The dataset leaves ample room for future research.

5.5 Impact of Hop and Entity Counts

To understand how question complexity affects model performance, we analyze accuracy as a function of hop count and entity count, using line charts under the *Oracle setting*, where full reasoning support is provided. As shown in Figure 2, model accuracy monotonically decreases with more reasoning hops or more involved entities. This confirms that increased logical complexity poses greater challenges for current models, even when relevant evidence is explicitly given.

However, this performance drop becomes less pronounced under less supportive conditions, such as the *Original setting*, where no additional external evidence is provided. In these cases, models sometimes perform better on higher-hop or multi-entity questions. This phenomenon is primarily attributed to the structure and content of the questions themselves. Specifically, higher hop or entity complexity tends to introduce more textual entities into the question, and these entity names are explicitly mentioned in the question itself. These elements often relate to general world knowledge

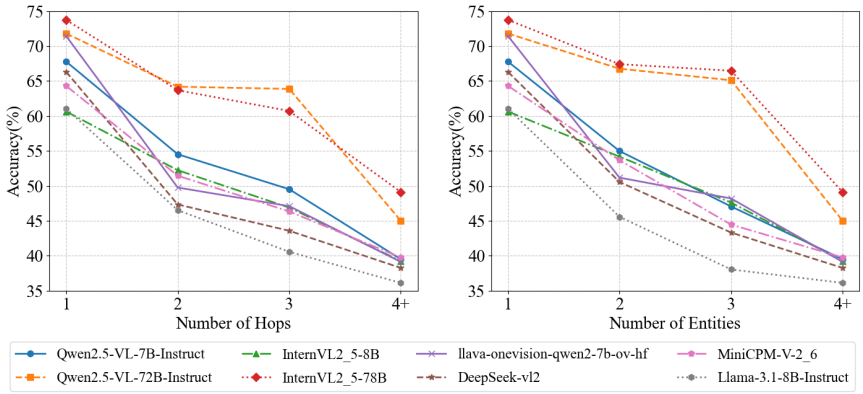

Figure 2: Model accuracy across different hop and entity counts, under the sentence evidence setting.

Question Type	Model	No Evidence	Partial Evidence 1	2	3	Full Evidence
Parallel Reasoning (4 Entities)	Qwen2.5-VL-7B (Bai et al., 2025)	21.7	28.1	34.8	36.7	39.5
	Qwen2.5-VL-72B (Bai et al., 2025)	**29.2**	**34.0**	**38.6**	**39.5**	**45.0**
	InternVL2.5-8B (Chen et al., 2024)	17.4	27.9	33.4	36.6	39.3
	MiniCPM-V-2.6 (Yao et al., 2024)	25.1	31.8	36.1	37.4	39.8
	Llama-3.1-8B-Instruct (*text*) (Grattafiori et al., 2024)	17.4	23.0	26.7	31.7	36.1
Sequential Reasoning (3 Hops)	Qwen2.5-VL-7B (Bai et al., 2025)	25.9	33.8	43.7	-	59.0
	Qwen2.5-VL-72B (Bai et al., 2025)	**26.2**	**35.3**	**45.7**	-	**59.5**
	InternVL2.5-8B (Chen et al., 2024)	17.6	29.0	36.9	-	44.0
	MiniCPM-V-2.6 (Yao et al., 2024)	13.4	29.6	39.5	-	53.9
	Llama-3.1-8B-Instruct (*text*) (Grattafiori et al., 2024)	20.5	31.0	41.2	-	50.7

Table 6: Model accuracy under varying degrees of gold reasoning evidence coverage.

or language-based associations that models can answer without full reasoning or retrieval. In contrast, questions focused on visual entities typically require grounded image understanding, which is more challenging in the absence of evidence.

5.6 Impact of Reasoning Support Coverage

To investigate how partial evidence affects performance on complex reasoning tasks, we design a controlled evaluation where all questions have fixed complexity: either 4-entity parallel reasoning or 3-hop sequential reasoning. For each question, we vary the number of reasoning steps for which gold supporting evidence is provided—from none to full coverage. This setup simulates different levels of difficulty by controlling how much of the reasoning chain is revealed to the model.

As shown in Table 6, model accuracy improves consistently as more evidence is made available. For both reasoning paradigms, even partial support (e.g., 1 or 2 hops/entities) leads to large performance gains over the no-evidence baseline. The most substantial improvement occurs when near-complete or full evidence is given, especially in sequential reasoning tasks where intermediate steps are critical. These results highlight the strong dependency of multi-entity and multi-hop VQA models on the availability of high-quality intermediate knowledge, and motivate future work on retrieval and grounding strategies.

We strongly recommend reading Appendix H for additional experimental results.

6 Conclusion

In this work, we introduce M^3-VQA, a challenging and novel benchmark designed to push the limits of MLLMs in fine-grained entity understanding and complex multi-hop reasoning across multimodal inputs. Unlike prior VQA datasets that focus on simpler, single-entity questions and coarse-grained categories, M^3-VQA presents diverse multi-entity scenarios requiring both sequential and parallel reasoning over multiple documents and modalities. Our comprehensive evaluation of 16 leading MLLMs under varying conditions (without evidence, with gold evidence, and with retrieval-augmented input) reveals substantial gaps in current model capabilities. The poor performance without external information highlights the critical need for effective knowledge integration. Moreover, the significant gains achieved through precise evidence and reasoning-aware agentic retrieval demonstrate the vital role of structured, traceable reasoning and sophisticated retrieval mechanisms

for complex multimodal understanding. We believe M^3-VQA sets a new, rigorous standard for evaluating and advancing the reasoning abilities of future MLLMs, encouraging further research into more robust knowledge acquisition and multi-hop reasoning strategies.

Acknowledgements

This research is supported by Artificial Intelligence-National Science and Technology Major Project (2023ZD0121200), and the National Natural Science Foundation of China (62437001, 62436001, 62531026), the Key Research and Development Program of Jiangsu Province under Grant BE2023016-3, and the Natural Science Foundation of Jiangsu Province under Grant BK20243051.

Limitations

Our dataset has several limitations that should be acknowledged. First, it is currently limited to English-language content, which restricts its applicability to multilingual or cross-lingual research settings. Future work could extend the dataset to include other languages supported by Wikipedia and beyond, facilitating broader cultural and linguistic coverage. Second, the dataset focuses primarily on knowledge derived from Wikipedia. While Wikipedia is a rich and widely-used source, it may not comprehensively represent specialized domains. Future expansions could explore additional domains such as biomedical literature, legal texts, or cultural artifacts. Third, although we employed a combination of automatic generation and manual filtering for question construction, the presence of a small number of errors or ambiguous instances is inevitable. We encourage users to be mindful of this and consider further validation in downstream tasks. Additionally, while we evaluate a wide range of models under multiple settings, our current study does not systematically assess different reasoning strategies, such as chain-of-thought (CoT) (Wei et al., 2022) prompting. Exploring the impact of such reasoning techniques remains an important direction for future work.

Ethical Considerations

Data Sources and Usage Permission

All images used in this dataset are sourced from publicly licensed datasets or from publicly accessible resources explicitly authorized by the rights holders. We strictly adhere to the usage agreements and licensing terms of each data source, and all data is used solely for academic research and non-commercial purposes.

Personal Privacy and Identifiable Information

We acknowledge that images may contain personally identifiable information, particularly facial features and license plates. To address this, we have blurred license plates to prevent vehicle identification and misuse. While the dataset includes some images of public figures, all such images are publicly available on authorized platforms, and their use complies with fair use principles and the terms of the original datasets. We have not collected or published any unauthorized private photos or images.

Bias and Fairness

Although we have made efforts to ensure diversity in image content and question-answer pairs, the dataset may still contain cultural or regional biases inherited from the original image collections. We encourage downstream users to carefully examine and mitigate potential bias impacts when using this dataset for model training or evaluation. Additionally, our work employs pretrained large language models for generating, analyzing, or evaluating VQA content. These models may carry inherent biases from their training data, including linguistic, cultural, or political tendencies. Such issues are beyond our control, and we advise readers to interpret results with caution.

Intended Use and Potential Misuse

This dataset is intended to advance academic research in multimodal AI systems, particularly in the development and evaluation of visual question answering tasks. We explicitly oppose any use of this dataset for privacy infringement or discriminatory technology development. We strongly recommend that users adhere to relevant ethical guidelines and legal regulations during data usage.

Environmental Impact

LLMs were used during our research, which operate on high-performance computing hardware and require significant energy resources, potentially contributing to carbon emissions. We encourage researchers to adopt efficient resource management practices to reduce carbon footprints during model development and evaluation.

References

Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. Feverous: Fact extraction and verification over unstructured and structured information. *arXiv preprint arXiv:2106.05707*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Hannah Bast and Björn Buchhold. 2017. Qlever: A query engine for efficient sparql+ text search. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 647–656.

Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-match: Restaurant-specific food logging from images. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 844–851. IEEE.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer.

Andrea Burns, Krishna Srinivasan, Joshua Ainslie, Geoff Brown, Bryan A Plummer, Kate Saenko, Jianmo Ni, and Mandy Guo. 2023. A suite of generative tasks for multi-level multimodal webpage understanding. *arXiv preprint arXiv:2305.03668*.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. 2022. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*.

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

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, and 1 others. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2016. Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics*, 21(3):588–598.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *arXiv preprint arXiv:2401.08281*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. 2023. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12065–12075.

Drew A Hudson and Christopher D Manning. 2019. 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*, pages 6700–6709.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Aman Jain, Mayank Kothyari, Vishwajeet Kumar, Preethi Jyothi, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2021. Select, substitute, search: A new benchmark for knowledge-augmented visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2491–2498.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. Hover: A dataset for many-hop fact extraction and claim verification. *arXiv preprint arXiv:2011.03088*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Justin Kay and Matt Merrifield. 2021. The fishnet open images database: A dataset for fish detection and fine-grained categorization in fisheries. *arXiv preprint arXiv:2106.09178*.

Prafful Kumar Khoba, Chirag Parikh, CV Jawahar, Ravi Kiran Sarvadevabhatla, and Rohit Saluja. 2022. A fine-grained vehicle detection (fgvd) dataset for unconstrained roads. In *Proceedings of the Thirteenth Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–9.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8082–8090.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Paul Lerner, Olivier Ferret, Camille Guinaudeau, Hervé Le Borgne, Romaric Besançon, José G Moreno, and Jesús Lovón Melgarejo. 2022. 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*, pages 3108–3120.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and 1 others. 2024b. Benchmarking multimodal retrieval augmented generation with dynamic vqa dataset and self-adaptive planning agent. *arXiv preprint arXiv:2411.02937*.

Zhuowan Li, Bhavan Jasani, Peng Tang, and Shabnam Ghadar. 2024c. Synthesize step-by-step: Tools templates and llms as data generators for reasoning-based chart vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13613–13623.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Thomas Mensink, Jasper Uijlings, Lluis Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. 2023. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE.

Kaimu Okamoto and Keiji Yanai. 2021. Uec-foodpix complete: A large-scale food image segmentation dataset. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*, pages 647–659. Springer.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Stefan Romberg, Lluis Garcia Pueyo, Rainer Lienhart, and Roelof Van Zwol. 2011. Scalable logo recognition in real-world images. In *Proceedings of the 1st ACM international conference on multimedia retrieval*, pages 1–8.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8876–8884.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Andras Tüzkö, Christian Herrmann, Daniel Manger, and Jürgen Beyerer. 2017. Open set logo detection and retrieval. *arXiv preprint arXiv:1710.10891*.

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

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

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

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, and 1 others. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Yujie Zhong, Relja Arandjelovic, and Andrew Zisserman. 2018. Compact deep aggregation for set retrieval. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0.

Alexander Ziller, Julius Hansjakob, Vitalii Rusinov, Daniel Zügner, Peter Vogel, and Stephan Günnemann. 2019. Oktoberfest food dataset. *arXiv preprint arXiv:1912.05007*.

A Additional Related Work

In addition to the related work discussed in Section 2, there are several more relevant areas:

A.1 Visual Question Answering

Visual Question Answering refers to answering questions about images. Early VQA datasets such as DAQUAR (Malinowski and Fritz, 2014), GQA (Hudson and Manning, 2019), VQAv1 (Antol et al., 2015), VQAv2 (Goyal et al., 2017), CLEVR (Johnson et al., 2017), FM-IQA (Gao et al., 2015), and Visual Genome (Krishna et al., 2017) mainly focus on reasoning based on image content, without requiring inference over named entities in the image or external knowledge.

A.2 Multi-hop Reasoning

There is a wealth of multi-hop textual datasets in the field of natural language processing (Jiang et al., 2020; Yang et al., 2018; Talmor and Berant, 2018; Joshi et al., 2017; Khot et al., 2020; Dunn et al., 2017; Mihaylov et al., 2018; Aly et al., 2021), covering tasks such as question answering and fact verification. Notably, HOTPOTQA (Yang et al., 2018) introduced the concept of "bridge entities" to construct multi-hop QA datasets. For example, in the question, "When was the singer and songwriter of Radiohead born?", one must first infer that the singer and songwriter of Radiohead is Thom Yorke, and then retrieve his birth date. Here, Thom Yorke serves as the bridge entity. We adopt this concept in constructing our sequential multi-hop questions.

A.3 Entity Recognition

Fine-grained visual entity recognition is a key challenge in M^3-VQA. Accurate reasoning and correct answers rely on correctly identifying the names of entities within images. Early work on fine-grained recognition focused on narrow domains such as plant identification (Van Horn et al., 2018), flowers (Nilsback and Zisserman, 2008), bird species (Wah et al., 2011), dogs (Khosla et al., 2011), cats (Parkhi et al., 2012), fish species (Kay and Merrifield, 2021), food (Bossard et al., 2014), cars (Krause et al., 2013), or aircraft types (Maji et al., 2013)—typically relying on deep learning models for classification. Recently, OVEN (Hu et al., 2023) unified labels from these datasets under Wikipedia's 6 million entities, proposing the task of Open-domain Visual Entity Recognition (OVEN), which requires models to link images to Wikipedia entities based on textual queries. These queries can be viewed as simple VQA tasks without detailed attribute reasoning. Unfortunately, OVEN lacks images involving multiple entities, which led us to build our own dataset from scratch.

A.4 Automatic Question Generation

Some prior work (Changpinyo et al., 2022; Mensink et al., 2023; Chen et al., 2023; Shah et al., 2019; Li et al., 2024c) has explored the automatic generation of single-hop VQA datasets. These approaches can be broadly categorized into three types. The first (Chen et al., 2023; Shah et al., 2019) uses question templates combined with SPARQL queries to Wikidata for retrieving answers. This method ensures high accuracy but suffers from limited diversity. The second approach (Changpinyo et al., 2022; Mensink et al., 2023) employs trained models to generate questions from free text, offering some diversity, though answers often depend heavily on the specific text. The third (Li et al., 2024c) uses large language models (LLMs), which can automatically generate questions. However, in our tests, neither trained models nor LLMs were able to generate high-accuracy multi-hop datasets. Therefore, we used a Wikidata-based method to construct multi-hop, multi-entity datasets. We further employed LLMs for question rewriting, combined with strict filtering and sampling to enhance question diversity.

B Proof of Theorem

For each (I, Q, A) triplet, we define: $E_I = \{e_1^I, e_2^I, \ldots\}$ as the set of visual entities in the image; $E_Q = \{e_1^Q, e_2^Q, \ldots\}$ as the set of textual entities in the question. Complexity definitions are as follows:

Image Parallel Complexity Defined as the number of visual entities in image I, denoted as $IP = |E_I|$.

Text Parallel Complexity Defined as the number of explicitly mentioned textual entities in question Q, denoted as $TP = |E_Q|$.

Sequential Complexity The question Q can be represented as a sequence $Q = (q_1, q_2, \ldots, q_s)$, where each q_i for $i = 2, 3, \ldots, s$ depends on the answer to q_{i-1}. The maximum length s is defined as the sequential complexity: $S = \max(s)$.

Entity Complexity The total number of visual and textual entities, defined as $P = IP + TP$.

Hop The number of Wikipedia pages needed to answer question Q.

Theorem $Hop = IP + TP + S - 1$.

$Proof$: Each Wikipedia page corresponds to one entity. Considering the maximum number s, the questions $q_2, \ldots, q_s$ each introduce a new entity (i.e., the answer to the previous question), while the entity in q_1 can only be from existing ones, i.e.,

$E_{q_1} = E_I \cup E_Q$. Therefore:

$$\begin{aligned} Hop &= |E_I \cup E_Q \cup E_S| \\ &= |E_I \cup E_Q \cup E_{q_2} \cup \ldots \cup E_{q_s}| \\ &= |E_I| + |E_Q| + |E_{q_2}| + \ldots + |E_{q_s}| \\ &= IP + TP + (s-1) \\ &= IP + TP + S - 1 \end{aligned}$$

C Extending M^3-VQA to Other Tasks

The M^3-VQA dataset is not only suitable for visual question answering (VQA) tasks, but it also naturally supports or can be extended to the following tasks:

C.1 Multimodal Retrieval

Based on M^3-VQA, multimodal retrieval involves multi-hop and multi-entity retrieval, termed as **M3RAG** (Multimodal Multi-hop Multi-entity Retrieval-Augmented Generation). This is also a novel task that has not been previously explored. The M3RAG task can be represented as a function $f : (I, Q, K) \rightarrow G$, where:

- I is an image. Each image $I \in \mathcal{I}$ is represented by pixels and may contain multiple visual entities;
- Q is a question. Each question $Q \in \mathcal{Q}$ is represented by text and may contain multiple textual entities and sub-questions;
- K is an external Wikipedia knowledge base, consisting of free-form text and images;
- G is the gold evidence, with each $G \in \mathcal{G}$ corresponding to a subset of pages in K.

We define IP, TP, and S similarly to M^3-VQA. Hop is defined as the total number of knowledge pages that the gold evidence G belongs to. In our dataset, we have:

$$Hop = IP + TP + S - 1$$

C.2 Multi-hop Reasoning

When evidence retrieval is not required and gold evidence is directly provided, M^3-VQA reduces to the **M3REASONING** task. In this task, the model bypasses the retrieval challenge and focuses on performing multi-step reasoning based on the given knowledge to answer the question. The M3REASONING task is defined as a function $f : (I, Q, G) \rightarrow A$, where:

- I is an image. Each image $I \in \mathcal{I}$ is represented by pixels and may contain multiple visual entities;
- Q is a question. Each question $Q \in \mathcal{Q}$ is represented by text and may contain multiple textual entities and sub-questions;
- G is the gold evidence, with each $G \in \mathcal{G}$ containing several evidence items;
- A is the answer, with each $A \in \mathcal{A}$ being a string, time, number, etc.

We define IP, TP, and S similarly to M^3-VQA. Hop corresponds to the number of reasoning steps. In our dataset, we have:

$$Hop = IP + TP + S - 1$$

C.3 Visual Entity Recognition

By discarding the textual questions in M^3-VQA, we can instead construct a textual query Q, such as "What kind of lion is on the left side of the picture?". Additionally, we construct an entity label set E based on Wikipedia. This leads to the task of **Open-domain Multiple Visual Entity Recognition (OMVER)**, which is also a previously unexplored task. The OMVER task is defined as a function $f : (I, Q, K) \rightarrow A$, where:

- I is an image. Each image $I \in \mathcal{I}$ is represented by pixels and may contain multiple visual entities;
- Q is a query. Each query $Q \in \mathcal{Q}$ is represented by text and indicates the intention of identifying elements in image I;
- K is a knowledge base, represented as $K = \{(e, p(e), t(e)) \mid e \in E\}$, a set of triples. Each triple contains an entity $t(e)$ (i.e., the entity's name, description, etc.) and a (possibly empty) set of associated images $p(e)$;
- A is the answer space, where each $A \in \mathcal{A}$ is an entity from E.

C.4 Summary

Note that these three tasks are in fact sub-tasks of M^3-VQA, but they can also be studied independently. On the other hand, improving M^3-VQA often requires advances in visual entity recognition, multimodal retrieval, and multi-hop reasoning.

Progress in these three tasks will significantly promote the development of M^3-VQA. Furthermore, merging visual entity recognition and multimodal retrieval into a single unified step may be a promising future direction.

D Dataset Construction Details

D.1 Image Sources

Manual Collection To gather more fine-grained, multi-entity images, we used search engines like Google and Baidu. Specifically, annotators crafted queries likely to yield multi-entity images—for example, "Ragdoll cat and Maine Coon." They then searched for images and manually reviewed and filtered the results.

CelebTo (Zhong et al., 2018). This dataset contains images with multiple labeled celebrities in each picture. It can be used as a benchmark for evaluating the retrieval of a set of identities. The dataset consists of 194,000 images, totaling 546,000 faces, covering 2,622 labeled celebrities. 59% of the faces correspond to these 2,622 celebrities, while the remaining faces are considered "unknown" individuals. The images in this dataset are sourced from Google Image Search and validated through manual labeling.

FGVD (Khoba et al., 2022). It is a dataset for fine-grained vehicle detection, with images collected from mobile cameras mounted on cars. The FGVD dataset is challenging because the vehicles appear in complex traffic scenes with a diversity of types, scales, poses, occlusions, and lighting conditions within and across categories. The dataset contains 5,502 scene images and covers 210 unique fine-grained labels, including various vehicle types, and is organized into a three-layer hierarchical structure. While previous classification datasets have included manufacturer information for different vehicle models, FGVD introduces new category labels for classifying two-wheeled vehicles, auto rickshaws (three-wheelers), and trucks.

FlickrLogos-47 (Romberg et al., 2011). The dataset contains photos showcasing brand logos and is designed for evaluating logo detection and recognition systems in real-world images. This dataset is constructed from the same images as the FlickrLogos-32 dataset but has been re-annotated to fix missing labels and add more categories.

LogosInTheWild (Tüzkö et al., 2017). This dataset consists of web images and their corresponding logo annotations, collected via Google Image Search.

Menu-match (Beijbom et al., 2015). The dataset includes meal images from three restaurants: an Asian restaurant offering a buffet-style service where customers can choose 1 to 3 side dishes served with brown or white rice; an Italian restaurant serving a variety of pizzas, lasagna, pasta, and accompanying breadsticks or salads; and a soup restaurant providing 10 types of soups with a choice of 5 types of bread. The dataset contains 646 images, annotated with 1,386 food items, covering 41 categories.

Oktoberfest (Ziller et al., 2019). It is a real, diverse, and challenging dataset for object detection in images. The data was collected from a beer tent scene in Germany and includes 15 different categories of food and drink items.

UEC-FoodPix (Okamoto and Yanai, 2021). It is a food image dataset with segmentation masks, containing 9,000 training images and 1,000 test images. The segmentation masks are enhanced according to food categories. In UECFoodPix, the mask images are generated using bounding boxes and the GrabCut method. The mask images have pixel-level labels for 103 food categories, which are stored only in the red (R) channel of the images.

UNIMIB2016 (Ciocca et al., 2016). This dataset is used for food recognition and segmentation tasks. It contains 1,027 images of trays with multiple food items, covering 73 food categories. All images have been manually labeled using polygons for each food instance, with corresponding food labels.

Infoseek (Chen et al., 2023). This is a visual question answering dataset focused on information retrieval-type questions that cannot be answered through common sense knowledge. It collects high-quality visual information retrieval question-answer pairs through multi-stage manual annotation. Additionally, it constructs a large-scale automatically collected dataset by combining existing visual entity recognition datasets with Wikidata, providing over one million examples for model fine-tuning and validation.

EVQA (Mensink et al., 2023). This is a large-scale visual question answering (VQA) dataset fo-

cused on visual questions about fine-grained categories and detailed attributes of instances. The dataset contains 221,000 unique question-answer pairs, each paired with (up to) 5 images, forming 1 million VQA samples. It provides a controlled knowledge base derived from Wikipedia and offers supporting evidence for each answer.

D.2 Entity Linking

Exact Matching Stage We used QLever (Bast and Buchhold, 2017), a highly efficient SPARQL engine capable of handling large-scale knowledge graphs like the full Wikidata (Vrandečić and Krötzsch, 2014) dump (dated August 22, 2022). We queried the Wikidata QIDs using the original label texts. If the query returned a unique result, we accepted it.

Manual Matching Stage If the query yielded multiple results or if poor linking was detected, annotators manually resolved the match. We excluded entities that could not be found in Wikidata or lacked an English Wikipedia page.

D.3 Knowledge Base

To align with Wikidata, we selected the English Wikipedia (Burns et al., 2023) snapshot closest in time (August 13, 2022), containing approximately 2 million articles. We retained the pages of all entities included in M^3-VQA. For each entity, we sampled about 5 hard negative examples. Specifically, we used all-mpnet-base-v2 to compute embeddings for Wikipedia article summaries and retrieved the top 20 most similar entities using FAISS, randomly selecting 5 as hard negatives. The candidate entity set included all M^3-VQA entities and their hard negatives. From each entity's Wikipedia page, we selected at least 3 sections—including the one containing golden evidence—to form the text corpus. The first image of each entity page was used to build the image corpus. These two corpora together formed the knowledge base. Importantly, M^3-VQA requires retrieval from the full constructed knowledge base, not just from the positive and hard negative examples. This strategy avoids searching over all 2 million articles (reducing computational cost and evaluation time) while maintaining retrieval difficulty through the inclusion of challenging negative samples.

D.4 Quality Control

Annotator Selection To ensure annotation quality, we provided annotators with thorough training. They attended in-person tutorial sessions covering annotation guidelines and common mistakes. Then, they took a qualification test; those who failed received personalized feedback and were retested. Only those who passed could participate in the main task. All annotators were at least undergraduate students. We explicitly explained to them how the collected data would be utilized. This project is voluntary, and all collaborators were informed of the authorship mechanism before joining.

Filtering We applied strict filtering at multiple stages:

- Low-quality images were removed.
- Entities not found in Wikidata or Wikipedia were discarded.
- Questions lacking corresponding entity attributes in Wikidata were eliminated.
- Sequential multi-hop questions were verified with a large language model (LLM): if the new question's answer didn't match the original, or if it was too short/long, it was discarded.
- For IQA (Image-Question-Answer) pairs, since Wikidata and Wikipedia are independently crowd-sourced, some answers from Wikidata might not appear in Wikipedia. We removed any question whose answer couldn't be found in the relevant Wikipedia page.
- Annotators also filtered out poorly structured questions.

In total, we retained approximately 550,000 IQA pairs.

Sampling To ensure balance across different types and complexities, and to maintain diversity in questions, entities, and images, we performed subsampling on the original dataset. Details are in Appendix E.

Quality Evaluation To further assess quality, we evaluated annotation accuracy. Experts reviewed 200 stratified samples from M^3-VQA. For each sample, they determined the answer based on the image, question, Wikipedia page, and evidence, and compared it with the provided annotation. Results showed that 93% of expert answers matched

	S	P	TP	IP	hop
0	-	-	6929	-	-
1	11225	2992	1909	6992	1092
2	998	2183	2287	2827	3181
3	902	3644	1000	2355	4546
4+	-	4306	1000	951	4306

Table 7: Complexity of M^3-VQA.

TP \ IP	1	2	3	4	5+
0	1092	1183	2000	500	254
1	1000	644	171	73	21
2	1000	1000	184	75	28
3	1000	-	-	-	-
4	1000	-	-	-	-

Table 8: Distribution of IP and TP under $S = 1$

the annotations. For comparison, A-OKVQA has 86% correctness, INFOSEEK 95%, and EVQA 86%. Thus, M^3-VQA demonstrates high annotation quality.

E Additional Dataset Statistics

E.1 Complexity

Table 7 and Figure 3 presents the specific counts of questions with varying levels of complexity. In terms of hop count, the dataset includes 1,092 single-hop questions, 3,181 two-hop questions, 4,546 three-hop questions, and 4,306 questions with four or more hops. In terms of the number of entities, the dataset includes 2,992 single-entity questions, 2,183 two-entity questions, 3,644 three-entity questions, and 4,306 questions with four or more entities.

E.2 Detailed Complexity of M^3-VQA

In our dataset, there are 998 questions with $S = 2, IP = 1, TP = 0$, and 902 questions with $S = 3, IP = 1, TP = 0$. All other questions have $S = 1$. Under the condition of $S = 1$, the distribution of IP and TP in the dataset is shown in Table 8.

E.3 Question Type

We conducted heuristic identification for each question type in the dataset. To identify question types, we performed hierarchical parsing on each question. Each question sentence was split by spaces (interpreted as words or phrases), and up to the first three levels were used to build a hierarchical structure. We then counted the frequency of each path combination at each level and set a threshold of 50 to filter out low-frequency paths. As shown in Figure 4, the dataset covers a diverse range of questions centered around locations, entities, events, comparisons, and numerical information.

E.4 Data Type

To gain a more comprehensive understanding of the dataset's semantic composition, we conducted a categorical analysis of the subjects referenced by the questions associated with each image. The data were manually annotated and categorized into nine distinct semantic classes: person, animal, building, vehicle, plant, food, logo, landmark, and other. The distribution across these categories reflects a broad coverage of real-world entities, which enhances the model's generalization capability in diverse scenarios. As shown in Figure 5 and Figure 6, the most frequent categories are person (2,406 instances), animal (2,115), and building (1,388), followed by vehicle (1,312) and plant (1,122). Less common yet still representative categories include food (772), logo (652), and landmark (574). The "other" category, consisting of samples that are difficult to classify or semantically ambiguous, contains 224 instances. This diverse category distribution supports comprehensive and robust model evaluation across a wide range of semantic domains.

F Experimental Details

F.1 Benchmarking Protocols

We introduce six benchmarking protocols to assess model performance under varying levels of information accessibility. The table 9 and table 10 compare these settings.

F.2 Agentic Retrieval

Heuristic retrieval methods suffer from overloaded queries. Using the full question as a single text query is problematic when the question contains multiple entities or sequential sub-questions. Similarly, using the entire image may hinder visual retrieval, especially when multiple entities are present. These one-shot strategies place excessive burden on a single query, often retrieving superficially relevant but unhelpful content.

To handle complex multi-entity, multi-hop questions in the KB setting, we implement Agentic Retrieval, a multi-agent retrieval-augmented gen-

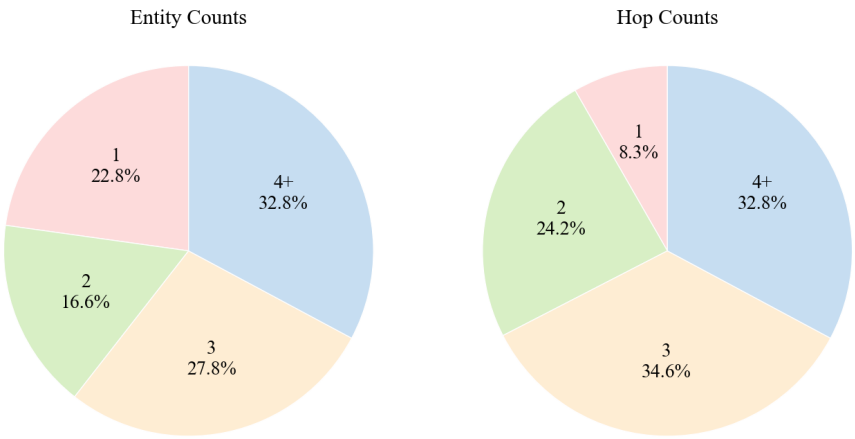

Figure 3: Complexity of M^3-VQA.

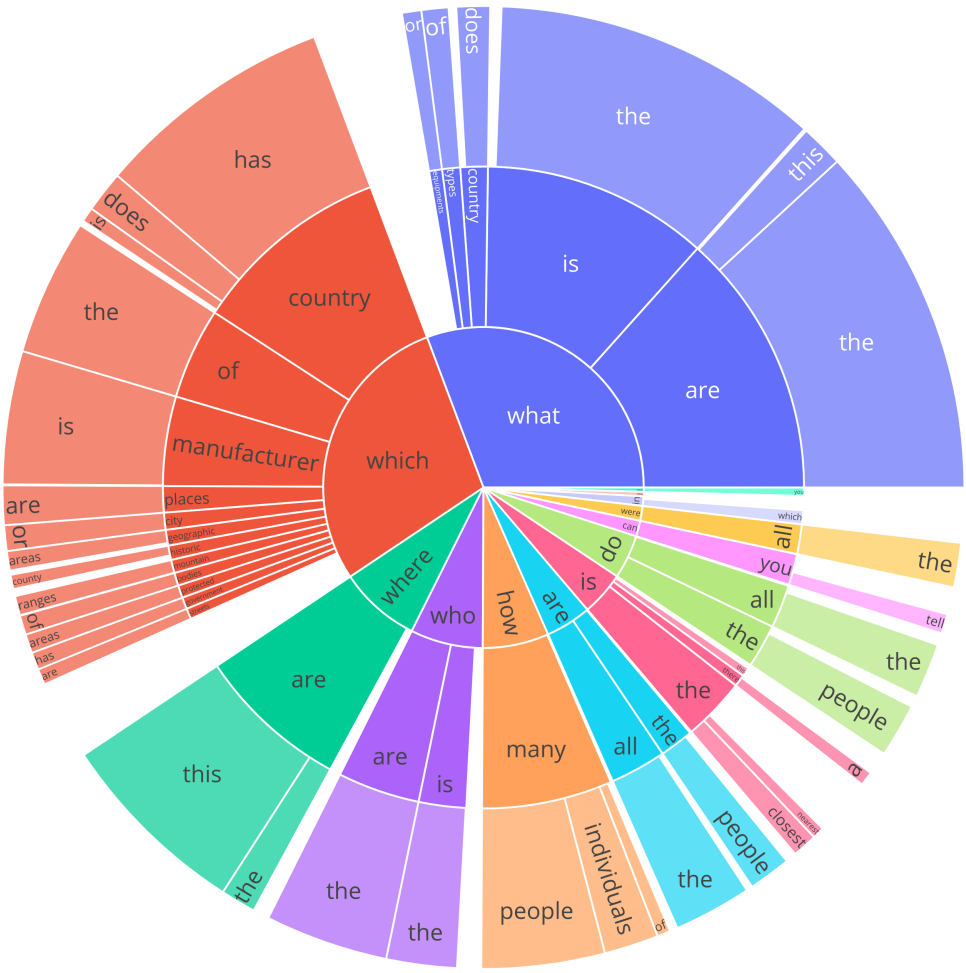

Figure 4: Types of questions covered in M^3-VQA.

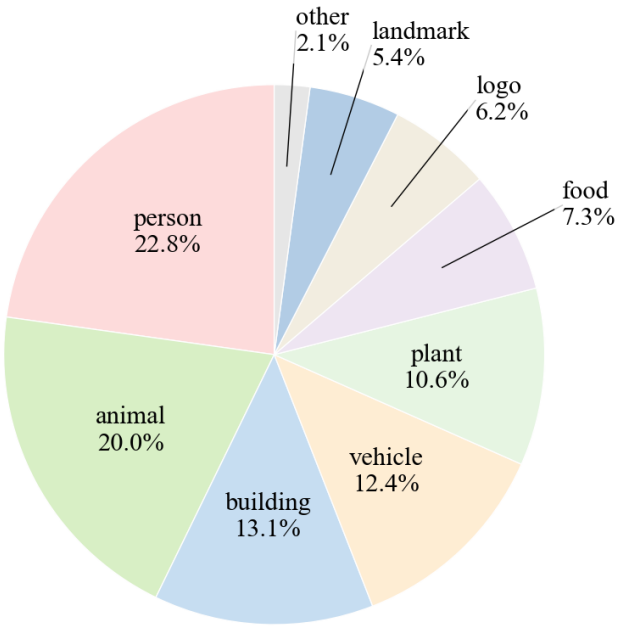

Figure 5: Types of data in M^3-VQA .

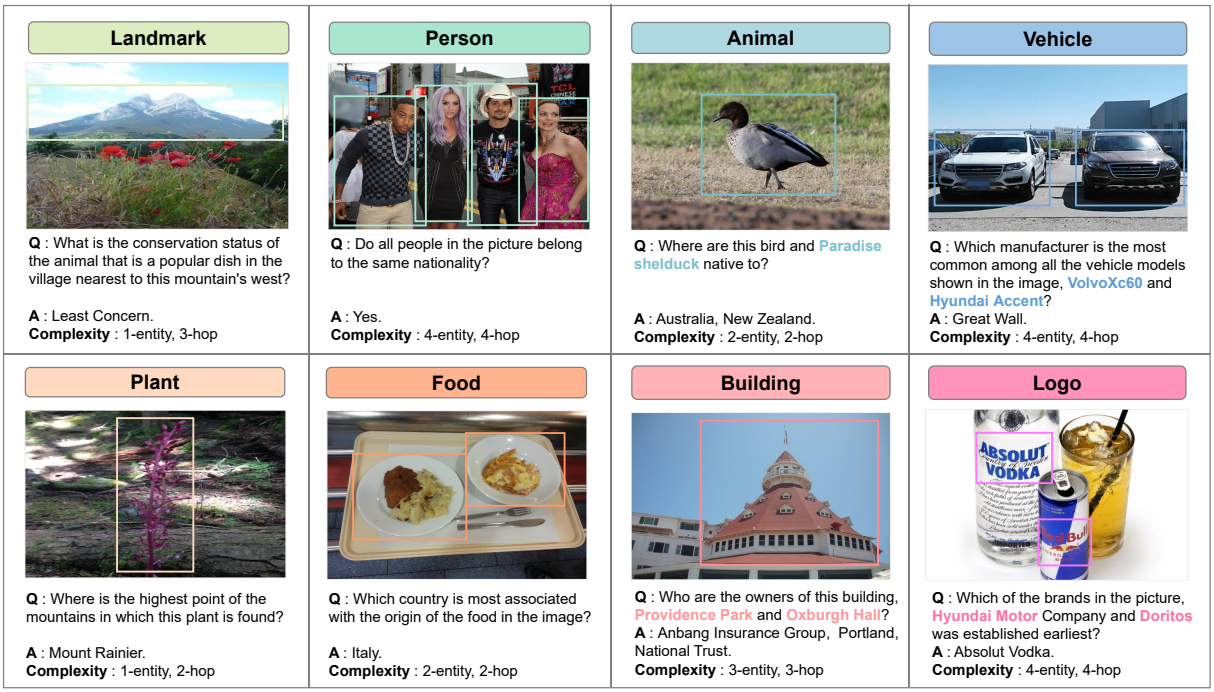

Figure 6: Examples of M^3-VQA questions across diverse fine-grained entity types.

	Sentence	Section	Entity-Name	KB	Original	Q-Only
I	✓	✓	✓	✓	✓	✗
Q	✓	✓	✓	✓	✓	✓
Visual entity name	✓	✓	✓	Retrieved	✗	✗
Evidence	Sentence	Section	✗	Retrieved	✗	✗

Table 9: Comparison of different benchmarking protocols.

Setting	Input	Methods	Example Models	Provided Evidence	Knowledge Base
Original	{I, Q}	End-to-end	Qwen2.5-VL-72B-Instruct	No	-
Oracle	{I, Q, E}	End-to-end	Qwen2.5-VL-72B-Instruct	Yes	-
KB	{I, Q, K}	Pipeline	Heuristic Retrieval, Agentic Retrieval	No	Wikipedia

Table 10: Comparison of different evaluation settings.

eration framework. The core idea is to emulate how humans break down and solve complex problems in steps. The framework consists of three components: Planner, Executor, and Solver. The Executor includes tools such as object detection, text retrieval, and image retrieval. The Planner devises a strategy and calls the appropriate tools. The Solver generates the final answer based on the retrieved information.

Planner The Planner is the core module, responsible for decomposing the problem and planning the next steps based on feedback from the Executor. It first calls the object detection module to locate relevant objects in the image, then segments the image accordingly. It may then perform image retrieval on the segments or issue single-hop queries for text retrieval. Once sufficient information is gathered, it passes the results to the Solver.

Executor The Executor is responsible for executing tool calls. It includes not only the text and image retrieval modules from heuristic retrieval, but also an object detection module to identify object coordinates in the image based on specific criteria.

Solver The Solver compiles the outputs from retrieval tools, removing tool execution traces and retaining only the retrieved content. It then attempts to answer the original question based on this information.

The entire process is fully automated and runs iteratively until the system determines that sufficient knowledge has been gathered to generate the final answer.

F.3 Text Retrieval

In both the heuristic retrieval and agentic retrieval models, we use text retrieval. We employ bge-large-en-v1.5 (Xiao et al., 2024) to compute embeddings for the text, which is an advanced and commonly used text embedding model. For knowledge base chapters, we directly compute embeddings and index them using FAISS (Douze et al., 2024). For queries, we append instructions and compute embeddings, then use FAISS's cosine similarity to retrieve the top 10 closest chapters. The instruction is: "Represent this sentence for searching relevant passages:"

F.4 Image Retrieval

For image retrieval, we use clip-vit-large (Radford et al., 2021) to compute embeddings. We compute embeddings separately for the article summary and its image, then sum the two embeddings to form the entity embedding, which is indexed using FAISS (Douze et al., 2024). Based on our preliminary experiments, the combined embedding of both the summary and the image outperforms using either image or summary alone, as well as combining entity names and image embeddings. More effective embedding calculation methods will be explored in future research. For image queries, we directly compute the image embedding, then use FAISS's cosine similarity to retrieve the top 10 closest entities. We then use the question as a query in the text retrieval module to search for the 10 most similar chapters from the 10 corresponding articles as retrieval results.

F.5 Object Detection

In the agentic retrieval model, we incorporate an object detection module. We chose Qwen2.5-VL-7B-Instruct (Bai et al., 2025), which uses a variety of rectangular box and point-based techniques for general object localization, enabling hierarchical localization and standardized JSON format output. The prompt we use is: "Outline the position of query and output all the bbox coordinates in JSON format."

F.6 Backbone Models

The backbone models we selected have the following characteristics: state-of-the-art; both open-source and closed-source; from different brands; with varying parameter sizes; multimodal and pure language models. The specific model introductions are as follows.

Qwen2.5-VL-3B, 7B, 72B-Instruct Released in January 2025. This model (Bai et al., 2025) makes

significant progress in world perception, acting as a visual agent, understanding long videos, capturing events, visual localization, and structured outputs.

Qwen2.5-VL-32B-Instruct Released in March 2025. Based on the Qwen2.5-VL series, this model (Bai et al., 2025) has been optimized using reinforcement learning for responses that better align with human subjective preferences, enhancing mathematical reasoning, fine-grained image understanding, and reasoning capabilities.

Qwen2-VL Released in August 2024. This model (Wang et al., 2024) is Capable of understanding images with various resolutions and aspect ratios, processing long videos over 20 minutes, and acting as a visual agent for mobile devices and robots, supporting multiple languages.

LLaVA-OneVision-7B Released in August 2024. LLaVA-OneVision (Li et al., 2024a) is an open-source multimodal LLM that trains Qwen2 on multimodal instruction-following data generated by GPT. It is the first model that can break the performance limits of open LMMs across three significant computing environments: single image, multi-image, and video scenarios. Notably, LLaVA-OneVision's design enables strong transfer learning across different modalities/scenarios, showcasing powerful video understanding and cross-scenario capabilities by transferring tasks from images to videos.

InternVL2.5 Released in December 2024. This (Chen et al., 2024) is an advanced multimodal large language model (MLLM) series built on InternVL2_0, significantly enhanced in training and testing strategies as well as data quality, while retaining its core "ViT-MLP-LLM" architecture.

MiniCPM-V-2.6 Released in August 2024. The most powerful model in the MiniCPM-V series, with 8B parameters, it (Yao et al., 2024) outperforms GPT-4V in single-image, multi-image, and video understanding, excelling in single-image understanding compared to GPT-4o mini, Gemini 1.5 Pro, and Claude 3.5 Sonnet, and is the first to support real-time video understanding on an iPad.

DeepSeek-VL2 Released in December 2024. This (Wu et al., 2024) is an advanced large expert mixture (MoE) visual language model with 4.5B active parameters, significantly improved over its predecessor, DeepSeek-VL. DeepSeek-VL2 demonstrates outstanding performance in various tasks, including visual question answering, optical character recognition, document/table/chart understanding, and visual grounding.

Llama-3.1-Instruct Released in July 2024. Llama 3.1 (Grattafiori et al., 2024) is an autoregressive language model using an optimized transformer architecture. The adjusted version uses supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for utility and safety.

Qwen2.5-Instruct Released in September 2024. Qwen2.5 (Yang et al., 2024) is the latest series in the Qwen large language model family, significantly improving encoding and mathematical abilities. It also shows marked improvement in instruction-following, generating long texts, understanding structured data, and generating structured outputs. The model is more resilient to system prompt diversity and enhances chatbot role-playing and condition setting.

GPT-4o GPT-4o (Hurst et al., 2024) is the fourth-generation multimodal large language model released by OpenAI in 2024, further optimized from GPT-4. It not only understands and generates natural language but also processes image, audio, and video inputs, enabling cross-modal information integration. GPT-4o features significant improvements in computational efficiency and inference speed, supporting real-time interaction. It is suitable for a wide range of applications, including conversational assistants, creative design, intelligent education, and visual search. The model is trained on massive datasets comprising text, images, and multimedia content, and leverages advanced self-supervised learning and reinforcement learning strategies. As a result, GPT-4o achieves industry-leading performance in accuracy, coherence, and safety.

F.7 Image Descriptions

Since pure language models cannot directly process images, we use image descriptions as a substitute input for the pure language models. Specifically, we use Qwen2.5-VL-7B-Instruct (Bai et al., 2025) to generate detailed descriptions of the images.

G Evaluation Metrics

Following previous work (Chen et al., 2023), we use answer aliases from Wikidata as multiple reference answers for string-based questions. Following

previous work (Goyal et al., 2017; Marino et al., 2019; Chen et al., 2023), we allow a one-year error margin for temporal questions, as historical events often have approximate dates. Since some questions may have multiple valid answers, we use Intersection over Union (IoU) accuracy, following previous work (Mensink et al., 2023). This is more accurate than exact match, as exact match overlooks partial correctness and cannot differentiate between completely wrong and partially correct answers. The arithmetic mean of all question accuracies is taken as the final score.

G.1 Maximum Matching

Since we provide aliases for some answers, we cannot directly compute the IoU between two sets. Our evaluation metric is essentially a maximum bipartite matching problem. For example, suppose the predicted answers are $[a1, b4, a3, d1]$, and the candidate answers are $[[a1, a2, a3], [b1, b2], [c1]]$. Each sublist in the candidate answers represents an answer and its aliases. We treat the predicted answers as node set X, with each item as a node in X. The candidate answers form node set Y, where each sublist is a node in Y. If an item in the predicted answers belongs to a sublist in Y, we draw an edge between the corresponding nodes. This forms a bipartite graph. The size of the maximum matching in this graph is denoted as NUM. Then, the IoU accuracy can be calculated as: $accuracy = NUM/(|X| + |Y| - NUM)$. In our example, $NUM = 1$ (note that both $a1$ and $a3$ connect to $[a1, a2, a3]$, but only one match is allowed). So, $accuracy = 1/(4 + 3 - 1)$. The optimal algorithm for maximum matching is the Hopcroft-Karp algorithm, with a time complexity of $O(\sqrt{V}, E)$.

G.2 Efficient Evaluation

Since some answers have many aliases, calculating IoU accuracy can be time-consuming for certain questions. In our practice, it is rare for two predicted answers to belong to the same alias group in the candidate answers (e.g., both $a1$ and $a3$ appearing simultaneously). Therefore, we provide an alternative method for calculating IoU accuracy. Specifically, we flatten the candidate answers into a one-dimensional list and count how many predicted items exist in this flattened list, denoted as NUM'. Then, IoU accuracy is: $accuracy = NUM'/(|X| + |Y| - NUM')$. In our tests, this method differs from the maximum matching-based approach by less than 0.2%, while significantly reducing evaluation time — completing evaluation on 13K items in just a few seconds.

G.3 Other Common Evaluation Metrics

INFOSEEK (Chen et al., 2023) uses Wikidata to obtain multiple reference answers. For temporal and numerical questions, it adopts lenient accuracy and computes the harmonic mean across dataset splits as the overall accuracy. We use a similar method but extend it to support multiple answers. DynVQA (Li et al., 2024b) tokenizes both generated and gold answers into token lists, then calculates the ratio of generated tokens found in the gold token list as the accuracy. However, this method has a drawback: longer or more verbose answers may receive higher scores even if additional content is incorrect. In contrast, the IoU metric penalizes extra incorrect answers — guessing more answers may lead to a lower score. EVQA (Mensink et al., 2023) employs BERT-based matching for evaluation, which is computationally expensive, especially for our larger dataset (13K items) and numerous candidate answers. Some works also use GPT-based evaluation, but due to the number of candidate answers in our setting, this approach is costly and less stable.

G.4 Quality of the IoU Metric

To verify that our IoU metric reliably reflects the model's actual performance, we introduce human-based accuracy. Specifically, we sample 200 predicted answers from our experiments and ask an expert to evaluate their correctness. We compare these expert evaluations with the scores given by the IoU metric. We find that in 95% of cases, the IoU-based score agrees with human judgment. This demonstrates the effectiveness of the IoU evaluation metric. At the same time, as an automated metric, it offers significantly lower computational cost and better scalability.

H Additional Experiment

H.1 Additional Experimental Results

We benchmark a wide range of models. This includes advanced closed-source models like GPT-4o (Hurst et al., 2024), and open-source models such as Qwen2.5-VL-72B-Instruct (Bai et al., 2025). To compare visual models from different providers, we evaluate 7B/8B versions of Qwen2.5-VL, Qwen2-VL (Wang et al., 2024), InternVL2.5

Setting	Model	Hop				Entity				All
		1	2	3	4+	1	2	3	4+	
Sentence	Qwen2.5-VL-3B-Instruct	71.30	48.47	41.11	32.55	55.99	48.25	40.10	32.55	42.60
	Qwen2.5-VL-7B-Instruct	67.77	54.52	49.52	39.52	60.51	54.98	47.06	39.52	48.97
	Qwen2.5-VL-32B-Instruct	73.42	57.58	56.66	42.03	63.09	59.15	55.72	42.03	53.48
	Qwen2.5-VL-72B-Instruct	71.78	64.18	63.86	45.01	63.46	66.75	65.10	45.01	58.41
	Qwen2-VL-7B-Instruct	57.70	46.64	39.75	34.58	51.66	47.13	36.92	34.58	41.21
	InternVL2.5-1B	31.50	22.06	19.89	17.01	21.40	24.59	21.22	17.01	20.44
	InternVL2.5-2B	49.77	37.55	32.87	30.02	43.73	36.92	30.66	30.02	34.47
	InternVL2.5-4B	67.88	47.34	40.86	35.04	51.40	50.05	40.46	35.04	42.77
	InternVL2.5-8B	60.63	52.24	46.89	39.26	51.36	54.18	47.62	39.26	46.82
	InternVL2.5-26B	58.31	54.92	48.44	41.32	52.06	56.01	49.56	41.32	48.50
	InternVL2.5-38B	77.24	64.78	62.85	44.56	64.16	67.15	65.18	44.56	58.51
	InternVL2.5-78B	73.73	63.67	60.71	49.12	56.73	67.40	66.46	49.12	58.71
	LLaVA-OneVision-7B	71.43	49.76	47.10	39.13	54.47	51.18	48.20	39.13	47.15
	DeepSeek-VL2	66.31	47.34	43.58	38.27	51.08	50.56	43.33	38.27	44.64
	MiniCPM-V-2.6	64.30	51.45	46.34	39.77	55.27	53.67	44.47	39.77	46.92
	GPT-4o	73.86	61.40	63.06	48.04	59.63	64.23	66.97	48.04	58.63
	Llama-3.1-8B-Instruct	61.05	46.53	40.57	36.13	53.79	45.58	38.05	36.13	42.26
	Qwen2.5-7B-Instruct	62.02	47.62	44.60	34.67	55.31	47.45	41.94	34.67	43.52
Section	Qwen2.5-VL-3B-Instruct	69.08	43.23	38.40	30.99	50.82	44.64	37.87	30.99	39.69
	Qwen2.5-VL-7B-Instruct	53.55	47.36	44.65	34.32	50.15	49.51	42.26	34.32	42.66
	Qwen2.5-VL-32B-Instruct	70.86	52.17	53.69	39.28	56.83	55.87	53.78	39.28	50.02
	Qwen2.5-VL-72B-Instruct	67.03	56.45	56.44	41.27	53.79	60.97	59.10	41.27	52.35
	Qwen2-VL-7B-Instruct	55.86	43.08	37.75	32.77	47.68	45.00	35.35	32.77	38.92
	InternVL2.5-1B	31.57	21.48	18.24	12.28	20.91	24.49	19.13	12.28	18.18
	InternVL2.5-2B	48.62	36.24	29.32	27.31	37.37	36.99	29.93	27.31	31.94
	InternVL2.5-4B	66.01	41.32	37.66	32.91	47.17	44.31	37.53	32.91	39.34
	InternVL2.5-8B	58.61	46.03	42.30	34.85	47.23	48.47	42.71	34.85	42.12
	InternVL2.5-26B	55.25	49.61	46.21	39.23	46.42	52.13	48.18	39.23	45.50
	InternVL2.5-38B	73.21	59.30	57.28	41.60	56.66	62.96	60.92	41.60	53.95
	InternVL2.5-78B	72.12	59.84	56.93	45.54	55.03	63.92	61.39	45.54	55.16
	LLaVA-OneVision-7B	69.87	46.10	42.82	35.96	50.15	47.93	44.73	35.96	43.62
	DeepSeek-VL2	63.04	41.31	38.49	31.17	47.09	44.88	37.40	31.17	38.81
	MiniCPM-V-2.6	60.77	39.93	40.14	31.91	47.75	40.85	39.48	31.91	39.11
	GPT-4o	71.03	54.57	56.07	44.22	53.99	59.22	59.07	44.22	53.07
	Llama-3.1-8B-Instruct	58.36	41.87	39.65	35.82	48.60	41.86	38.53	35.82	40.49
	Qwen2.5-7B-Instruct	59.38	41.90	39.02	31.26	51.16	41.68	36.09	31.26	38.87
Name	Qwen2.5-VL-3B-Instruct	49.40	25.91	22.98	20.20	29.57	30.30	23.67	20.20	24.98
	Qwen2.5-VL-7B-Instruct	41.26	35.62	34.80	23.49	28.35	43.13	37.77	23.49	31.83
	Qwen2.5-VL-32B-Instruct	62.99	38.25	42.18	34.36	38.11	46.50	45.73	34.36	40.39
	Qwen2.5-VL-72B-Instruct	65.20	42.52	48.37	38.29	39.55	52.82	52.90	38.29	45.05
	Qwen2-VL-7B-Instruct	40.33	28.85	29.99	28.23	26.50	34.82	32.08	28.23	30.00
	InternVL2.5-1B	36.08	20.28	22.02	13.90	18.76	26.35	24.78	13.90	20.10
	InternVL2.5-2B	34.48	19.54	25.21	20.92	20.95	24.86	26.75	20.92	23.20
	InternVL2.5-4B	51.81	27.14	29.90	25.34	28.83	33.46	32.81	25.34	29.56
	InternVL2.5-8B	41.50	26.53	25.39	21.44	25.00	32.55	27.24	21.44	25.71
	InternVL2.5-26B	46.45	34.75	35.51	28.97	28.43	43.00	39.45	28.97	34.09
	InternVL2.5-38B	65.26	43.15	46.81	37.33	39.54	54.07	50.77	37.33	44.35
	InternVL2.5-78B	63.26	44.21	47.04	38.87	38.92	55.15	51.23	38.87	45.02
	LLaVA-OneVision-7B	49.68	30.79	33.21	26.20	28.23	38.93	36.68	26.20	31.69
	DeepSeek-VL2	50.67	29.96	31.45	24.97	28.79	37.24	34.62	24.97	30.56
	MiniCPM-V-2.6	53.61	32.28	32.66	30.49	30.85	39.52	35.99	30.49	33.60
	GPT-4o	69.69	48.18	54.59	45.29	42.02	60.65	60.08	45.29	51.21
	Llama-3.1-8B-Instruct	51.01	28.81	29.67	25.09	30.31	35.10	31.51	25.10	29.73
	Qwen2.5-7B-Instruct	45.45	26.70	28.77	20.59	28.20	33.28	29.72	20.59	26.97

Table 11: Performance comparison of various models under different settings.

Setting	Model	Hop				Entity				All
		1	2	3	4+	1	2	3	4+	
KB	*Heuristic Retrieval*									
	Qwen2.5-VL-3B-Instruct	38.47	18.13	21.82	25.37	23.84	23.63	22.66	25.37	23.48
	Qwen2.5-VL-7B-Instruct	41.62	22.33	23.38	26.54	26.07	24.40	25.11	26.54	25.68
	Qwen2.5-VL-32B-Instruct	48.49	29.41	32.83	31.13	31.20	34.60	34.83	31.13	32.75
	Qwen2.5-VL-72B-Instruct	51.65	33.67	36.49	34.99	33.51	39.69	39.08	34.99	36.57
	Qwen2-VL-7B-Instruct	31.27	19.21	21.21	25.14	20.96	21.64	22.42	25.14	22.85
	InternVL2.5-1B	27.03	9.49	13.49	12.47	13.71	10.10	15.90	12.47	13.31
	InternVL2.5-2B	21.69	14.78	18.21	20.64	16.65	16.12	18.81	20.64	18.47
	InternVL2.5-4B	32.56	17.70	20.58	23.55	21.80	19.54	21.27	23.55	21.85
	InternVL2.5-8B	30.46	20.33	24.61	28.49	22.28	22.87	25.87	28.49	25.33
	InternVL2.5-26B	31.81	24.36	29.00	34.45	25.23	26.91	30.15	34.45	29.90
	InternVL2.5-38B	45.34	28.50	32.70	32.40	31.42	31.79	34.40	32.40	32.63
	InternVL2.5-78B	44.70	28.82	33.80	32.77	29.82	33.32	36.27	32.77	33.16
	LLaVA-OneVision-7B	37.62	19.75	24.85	28.37	26.80	27.03	24.91	28.37	25.83
	MiniCPM-V-2.6	35.72	19.47	24.98	25.44	24.99	23.88	25.84	25.44	24.69
	GPT-4o	54.93	30.31	33.17	33.80	35.28	34.52	34.64	33.80	34.49
	Llama-3.1-8B-Instruct	30.73	18.91	20.13	28.23	19.10	23.56	22.22	28.23	23.37
	Qwen2.5-7B-Instruct	33.32	18.66	21.53	26.26	22.17	23.92	22.41	26.26	23.37
	Agentic Retrieval									
	Qwen2.5-VL-3B-Instruct	42.72	20.86	24.25	27.25	26.67	22.66	25.80	27.25	25.95
	Qwen2.5-VL-7B-Instruct	41.44	28.25	30.69	31.45	27.52	32.13	33.52	31.45	31.24
	Qwen2.5-VL-32B-Instruct	52.72	34.50	36.49	36.12	33.98	40.42	39.32	36.12	37.24
	Qwen2.5-VL-72B-Instruct	54.49	36.89	39.83	35.48	35.68	43.07	43.11	35.48	38.91
	Qwen2-VL-7B-Instruct	36.94	23.34	25.96	27.37	23.68	27.19	28.11	27.37	26.70
	InternVL2.5-1B	32.92	11.37	15.39	12.17	15.78	12.76	18.40	12.17	14.82
	InternVL2.5-2B	28.34	17.31	19.29	21.47	18.67	19.74	20.52	21.47	20.28
	InternVL2.5-4B	37.24	18.03	20.63	23.84	21.52	21.01	22.40	23.84	22.44
	InternVL2.5-8B	39.53	26.15	27.86	30.97	28.32	28.45	29.13	30.97	29.44
	InternVL2.5-26B	36.72	26.72	31.85	34.46	26.50	30.88	33.81	34.46	31.87
	InternVL2.5-38B	49.83	31.87	36.15	33.67	33.04	36.91	38.61	33.67	35.44
	InternVL2.5-78B	19.97	32.52	36.90	35.20	31.38	39.04	40.25	35.20	36.37
	LLaVA-OneVision-7B	42.73	24.07	28.35	29.38	28.28	27.32	29.61	29.38	28.85
	MiniCPM-V-2.6	42.84	23.08	27.90	28.02	27.11	25.92	30.00	28.02	28.01
	GPT-4o	55.36	35.73	39.66	36.06	35.34	42.77	42.61	36.06	38.83
	Llama-3.1-8B-Instruct	38.17	23.93	25.92	31.43	24.78	26.74	28.29	31.43	28.26
	Qwen2.5-7B-Instruct	36.26	20.10	22.11	27.45	22.55	22.59	23.95	27.45	24.55

Table 12: Performance comparison of various models under different settings.

Setting	Model	Hop				Entity				All
		1	2	3	4+	1	2	3	4+	
Original	Qwen2.5-VL-3B-Instruct	36.95	17.44	17.62	16.01	24.53	18.55	17.04	16.01	18.66
	Qwen2.5-VL-7B-Instruct	34.87	20.66	21.63	21.74	24.14	24.30	21.08	21.74	22.53
	Qwen2.5-VL-32B-Instruct	46.84	24.94	29.84	25.55	30.12	29.44	30.67	25.55	28.66
	Qwen2.5-VL-72B-Instruct	47.94	29.48	34.19	29.20	30.81	35.51	36.16	29.20	32.55
	Qwen2-VL-7B-Instruct	28.68	17.55	20.45	24.34	20.49	20.87	20.11	24.34	21.71
	InternVL2.5-1B	18.09	2.63	4.31	2.76	7.20	3.60	5.02	2.76	4.54
	InternVL2.5-2B	17.02	5.04	8.71	8.41	9.88	6.53	8.33	8.41	8.41
	InternVL2.5-4B	35.01	13.25	17.97	19.73	18.37	16.96	19.23	19.73	18.82
	InternVL2.5-8B	26.14	16.86	19.44	17.36	18.45	19.60	19.91	17.36	18.69
	InternVL2.5-26B	27.40	13.76	15.24	17.01	17.92	15.42	15.31	17.01	16.48
	InternVL2.5-38B	42.20	24.43	30.55	26.02	28.69	28.87	31.23	26.02	28.55
	InternVL2.5-78B	43.45	27.15	33.35	29.05	29.76	32.54	34.41	29.05	31.28
	LLaVA-OneVision-7B	34.98	17.53	21.56	22.54	21.26	20.87	22.72	22.54	22.02
	DeepSeek-VL2	40.20	20.31	24.86	22.68	24.62	23.96	26.21	22.68	24.32
	MiniCPM-V-2.6	31.46	19.13	23.05	25.05	19.17	23.05	25.31	25.05	23.45
	GPT-4o	51.60	23.30	24.25	27.93	34.40	25.87	22.31	27.93	27.50
	Llama-3.1-8B-Instruct	26.16	15.94	21.66	21.54	18.62	19.27	21.95	21.54	20.61
	Qwen2.5-7B-Instruct	23.99	12.85	18.69	17.35	16.70	16.09	18.35	17.35	17.27
Q-Only	Qwen2.5-VL-3B-Instruct	8.70	11.22	15.15	16.84	8.11	12.59	17.12	16.84	14.22
	Qwen2.5-VL-7B-Instruct	4.99	9.45	14.00	20.84	4.56	12.19	16.16	20.84	14.39
	Qwen2.5-VL-32B-Instruct	9.89	9.36	17.73	20.99	8.44	11.17	19.64	20.99	16.12
	Qwen2.5-VL-72B-Instruct	9.16	10.23	14.38	23.90	5.92	13.69	16.54	23.90	16.06
	Qwen2-VL-7B-Instruct	9.80	10.74	17.07	20.23	12.95	12.54	15.54	20.23	15.99
	InternVL2.5-1B	7.76	7.34	11.23	8.36	4.93	9.83	12.82	8.36	9.06
	InternVL2.5-2B	6.27	6.18	12.18	11.49	4.19	7.98	14.26	11.49	10.01
	InternVL2.5-4B	2.66	6.76	11.42	14.96	2.73	8.62	13.53	14.96	10.72
	InternVL2.5-8B	2.75	8.08	12.33	16.36	2.48	10.86	14.70	16.36	11.82
	InternVL2.5-26B	4.17	9.92	13.80	21.14	4.32	12.72	15.97	21.14	14.47
	InternVL2.5-38B	1.01	7.23	10.44	18.06	1.08	9.85	12.86	18.06	11.38
	InternVL2.5-78B	1.01	9.39	14.74	22.14	4.69	11.94	15.89	22.14	14.73
	LLaVA-OneVision-7B	10.58	11.41	19.81	20.77	10.48	14.32	20.66	20.77	17.32
	DeepSeek-VL2	9.57	9.99	14.23	17.90	6.91	12.59	16.13	17.90	14.02
	MiniCPM-V-2.6	10.71	14.02	19.55	22.72	13.69	16.24	18.85	22.72	18.51
	GPT-4o	4.67	13.16	19.11	28.76	7.35	17.39	20.28	28.76	19.63
	Llama-3.1-8B-Instruct	9.89	9.36	17.73	20.99	8.44	11.17	19.64	20.99	16.12
	Qwen2.5-7B-Instruct	2.83	7.70	12.51	17.44	3.86	10.40	13.79	17.44	12.16

Table 13: Performance comparison of various models under different settings.

(Chen et al., 2024), Llava-OneVision (Li et al., 2024a), DeepSeek-VL2 (Wu et al., 2024) and MiniCPM-V-2.6 (Yao et al., 2024). To study the impact of model size, we test the 3B, 7B, 32B, and 72B versions of Qwen2.5-VL, as well as the 1B, 2B, 4B, 8B, 26B, 38B, and 78B versions of InternVL2.5. We also evaluate pure language models (e.g., LLaMA-3.1 (Grattafiori et al., 2024), Qwen2.5 (Yang et al., 2024)) by replacing image inputs with textual descriptions. Due to the high computational cost, each experiment was conducted only once.

We provide the performance of additional models in Tables 11, 12 and 13. Please note that due to the context length limitations of certain models, results under the KB setting may be missing, and their performance under the Section setting may also be affected (e.g., DeepSeek-VL2).

H.2 Attribution Analysis of Retrieval Models

Based on the attribution annotations in our dataset, we can assess whether retrieval-augmented models successfully retrieve the correct information. We provide more detailed retrieval accuracy results for the heuristic retrieval and agentic retrieval in Tables 14 and 15. “Text" refers to text-based retrieval, “Image" refers to image-based retrieval, and “All represents the combined results of both. “Page" indicates the accuracy of retrieving the correct entity page, while “Section" denotes the accuracy of retrieving the correct section containing the evidence. “Top-k" refers to the number of hops in the retrieved evidence. “all" in Hop refers to all hop levels.

As seen in both tables, the accuracy of question-based text retrieval for 1 hop queries is extremely low, primarily because 1 hop questions do not contain additional textual entities. The highest text retrieval accuracy is observed for 4+ hop queries, largely because higher hop questions tend to include more textual entities. On the other hand, although the number of visual entities increases with the number of hops, the accuracy of image retrieval decreases, reflecting the inherent difficulty of retrieving information involving multiple visual entities.

Comparing the results in Table 16, we observe that agentic retrieval retrieves the correct knowledge base pages approximately 6% more often than heuristic retrieval. At the more fine-grained level of knowledge base sections, its retrieval accuracy is also about 6% higher. This indicates that decomposing complex queries into single hop queries can effectively improve retrieval precision. More specifically, we observe that agentic retrieval achieves 1% higher accuracy in text retrieval and 5% higher accuracy in image retrieval compared to the heuristic approach. Therefore, the segmentation of multi-entity images contributes more significantly to performance improvement than the segmentation of text queries.

Additionally, we find that agentic retrieval performs worse than heuristic retrieval in image retrieval accuracy for 1 hop queries. This is understandable, as single entity images do not require segmentation, and the segmentation of agentic retrieval introduces additional noise in these cases. The improvements in the image retrieval accuracy are more pronounced at 3 hop and 4+ hop levels compared to 2 hop, further demonstrating that entity segmentation is more effective when dealing with images containing more visual entities.

H.3 The Impact of Evidence Number

In multi-hop, multi-entity reasoning tasks, the model must rely on multiple interrelated pieces of knowledge to complete intermediate reasoning chains and gradually construct a logical path toward the final answer. To further analyze the role of evidence in such complex tasks, we designed a controlled experiment: we provided the model with varying numbers of gold evidence for each question and observed how the accuracy changed. The results for 4-hop and 4-entity cases have already been discussed in the main text. Table 17 presents more detailed results.

I Case Study

We provide a case study of agentic retrieval in Figure 7. The input image contains two cats on the grass. The input question is: *“Which country has the most origin for the cat in the image and Siamese?”* This means that we need to (1) identify the breeds of the two cats in the image, (2) retrieve the origin countries for both cats and the Siamese breed, and (3) determine which country appears most frequently among those origins. Since there are two visual entities in the image and one textual entity in the question, we have: $IP = 2, TP = 1, P = 3, S = 1, hop = 3$.

In agentic retrieval, the Planner first invokes the Object Detection module in the executor to segment the two cats from the image. Then, it calls the Image Retrieval module twice to obtain informa-

Hop	Top-k	Text		Image		All	
		Page	Section	Page	Section	Page	Section
all	3	23.11	20.25	19.41	9.93	41.88	28.96
	5	26.36	23.76	23.49	13.97	48.79	36.23
	7	28.04	25.56	26.45	17.37	52.98	41.20
	10	**29.85**	**27.64**	**29.56**	**21.27**	**57.21**	**46.92**
1 hop	3	0.92	0.64	37.64	25.00	38.00	25.09
	5	1.28	0.73	46.25	32.88	46.79	33.06
	7	2.01	1.47	50.92	37.73	51.65	38.28
	10	**2.38**	**1.74**	**55.22**	**44.41**	**56.95**	**45.05**
2 hop	3	20.47	19.10	26.94	13.94	46.15	30.97
	5	22.12	20.98	32.33	19.05	52.42	37.35
	7	23.20	21.95	36.22	23.64	56.63	42.49
	10	**24.36**	**23.10**	**40.29**	**29.02**	**60.69**	**48.48**
3 hop	3	18.99	16.08	18.07	6.72	36.52	21.96
	5	20.30	18.67	21.46	10.32	41.58	27.97
	7	21.69	19.44	24.08	13.50	44.70	31.80
	10	**23.08**	**21.12**	**26.88**	**17.34**	**48.57**	**37.23**
4 hop	3	35.06	30.47	10.64	6.52	45.37	35.85
	5	41.63	37.02	13.31	9.30	54.24	44.94
	7	44.93	40.79	15.54	11.65	59.37	50.91
	10	**48.02**	**44.46**	**17.96**	**13.82**	**64.09**	**56.47**

Table 14: Retrieval accuracy of the heuristic retrieval model under different hop and top-k settings.

Hop	Text		Image		All	
	Page	Section	Page	Section	Page	Section
all	**30.30**	**28.73**	**35.48**	**26.64**	**63.35**	**53.13**
1 hop	6.23	5.86	52.56	39.84	54.03	42.77
2 hop	23.22	22.28	45.27	34.25	65.23	53.62
3 hop	23.19	21.91	34.07	24.19	55.65	44.05
4 hop	49.15	46.51	25.40	20.26	72.47	64.98

Table 15: Retrieval accuracy of agentic retrieval under different hop.

Model	Text		Image		All	
	Page	Section	Page	Section	Page	Section
Heuristic Retrieval	29.85	27.64	29.56	21.27	57.21	46.92
Agentic Retrieval	30.30	28.73	35.48	26.64	63.35	53.13

Table 16: Retrieval accuracy of the heuristic retrieval and agentic retrieval.

Setting	Model	Hop				Entity				All
		1	2	3	4+	1	2	3	4+	
Gold@0	Qwen2.5-VL-7B-Instruct	34.87	20.66	21.63	21.74	24.14	24.30	21.08	21.74	22.53
	Qwen2.5-VL-72B-Instruct	47.94	29.48	34.19	29.20	30.81	35.51	36.16	29.20	32.55
	InternVL2.5-8B	26.14	16.86	19.44	17.36	18.45	19.60	19.91	17.36	18.69
	MiniCPM-V-2.6	31.46	19.13	23.05	25.05	19.17	23.05	25.31	25.05	23.45
	Llama-3.1-8B-Instruct	26.16	15.94	21.66	21.54	18.62	19.27	21.95	21.54	20.61
Gold@1	Qwen2.5-VL-7B-Instruct	67.77	32.43	33.69	28.13	39.01	38.23	35.73	28.13	34.40
	Qwen2.5-VL-72B-Instruct	71.78	40.44	42.10	33.97	43.02	48.04	45.20	33.97	41.49
	InternVL2.5-8B	60.63	32.41	33.34	27.92	33.54	39.92	36.61	27.92	33.61
	MiniCPM-V-2.6	64.30	33.16	32.60	31.82	34.87	40.60	35.94	31.82	35.12
	Llama-3.1-8B-Instruct	61.05	27.63	26.48	22.95	34.35	32.35	27.88	22.95	28.48
Gold@2	Qwen2.5-VL-7B-Instruct	67.77	54.52	39.18	34.75	49.28	54.98	43.53	34.75	43.86
	Qwen2.5-VL-72B-Instruct	71.78	64.18	51.24	38.64	54.37	66.75	56.77	38.64	51.93
	InternVL2.5-8B	60.63	52.24	37.93	33.43	44.53	54.18	42.05	33.43	41.81
	MiniCPM-V-2.6	64.30	51.45	37.94	36.11	44.94	53.67	42.53	36.11	42.84
	Llama-3.1-8B-Instruct	61.05	46.53	33.28	26.69	46.43	45.58	35.05	26.69	36.71
Gold@3	Qwen2.5-VL-7B-Instruct	67.77	54.52	49.52	36.65	60.50	54.98	47.06	36.65	47.97
	Qwen2.5-VL-72B-Instruct	71.78	64.18	63.86	39.50	63.57	66.75	65.10	39.50	56.63
	InternVL2.5-8B	60.63	52.24	46.89	36.64	51.34	54.18	47.62	36.64	45.84
	MiniCPM-V-2.6	64.30	51.45	46.34	37.44	55.24	53.67	44.47	37.44	46.29
	Llama-3.1-8B-Instruct	61.05	46.53	40.57	31.78	53.93	45.58	38.05	31.78	40.95
Gold@All	Qwen2.5-VL-7B-Instruct	67.77	54.52	49.52	39.52	60.51	54.98	47.06	39.52	48.97
	Qwen2.5-VL-72B-Instruct	71.78	64.18	63.86	45.01	63.46	66.75	65.10	45.01	58.41
	InternVL2.5-8B	60.63	52.24	46.89	39.26	51.36	54.18	47.62	39.26	46.82
	MiniCPM-V-2.6	64.30	51.45	46.34	39.77	55.27	53.67	44.47	39.77	46.92
	Llama-3.1-8B-Instruct	61.05	46.53	40.57	36.13	53.79	45.58	38.05	36.13	42.26

Table 17: Performance comparison of various models under different evidence number.

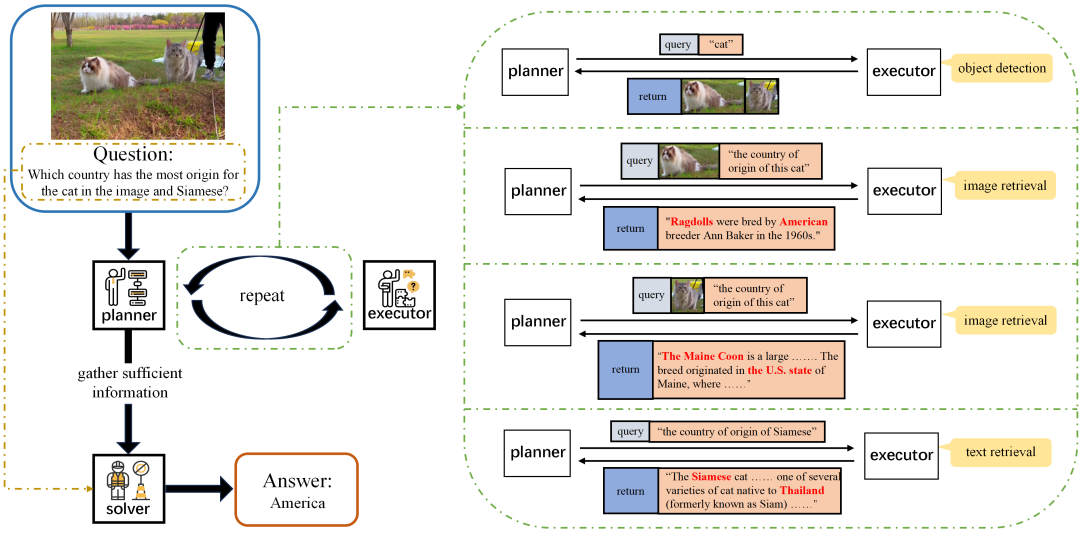

Figure 7: Schematic diagram and case study of the agentic retrieval model.

tion about the origins of the two cats. Afterwards, it uses the Text Retrieval module to obtain information about the origin of the Siamese breed. Before each call to the executor, the planner must assess whether sufficient information has already been acquired. If not, it continues the loop by calling the executor again. Once the planner deems the information sufficient, it organizes the retrieved data into an evidence list and passes it to the Solver. The solver then reasons over the image, the question, and the evidence to derive the final answer: *America*.

J Compute Resources

Experiments were conducted on an internal computing cluster equipped with NVIDIA A800 GPUs. The total computational cost was approximately 2000 GPU hours, primarily used for dataset construction and model inference. To accelerate inference and improve throughput, we deployed the models using the vLLM (Kwon et al., 2023) framework, which enables efficient execution of large language models while optimizing GPU memory usage. No large-scale model training was performed during the course of this study.

K Dataset Examples

We provide several dataset examples for readers to review intuitively.

Example 1

Question: Which manufacturer is the most common among all the vehicle models shown in the image, Volvo Xc60 and Hyundai Accent?
Question Type: IP2TP2S1 **Question Hop:** 4 **Entity Num:** 4
Answers: Great Wall.
Image Entity Names: Haval H8, Haval H8
Evidence:

- The Haval H8 is a mid-size SUV produced by Haval, a sub-brand of Great Wall Motor.
- The Volvo XC60 is a compact luxury crossover SUV manufactured and marketed by Swedish automaker Volvo Cars since 2008.
- The Hyundai Accent, or Hyundai Verna is a subcompact car produced by Hyundai.

Evidence URLs:

- https://en.wikipedia.org/wiki/Haval_H8
- https://en.wikipedia.org/wiki/Volvo_XC60
- https://en.wikipedia.org/wiki/Hyundai_Accent

Example 2

Question: Which of the brands in the picture, Hyundai Motor Company and Doritos was established earliest?
Question Type: IP2TP2S1 **Question Hop:** 4 **Entity Num:** 4
Answers: Absolut Vodka.
Image Entity Names: Absolut Vodka, Red Bull
Evidence:

- Absolut was established in 1879 by Lars Olsson Smith and is produced in Åhus, Sweden.
- Since its launch in 1987, more than 100 billion cans of Red Bull have been sold worldwide, including 9.8 billion in 2021.
- Hyundai Motor Company was founded in 1967.
- Doritos is an American brand of flavored tortilla chips produced since 1964 by Frito-Lay, a wholly owned subsidiary of PepsiCo.

Evidence URLs:

- https://en.wikipedia.org/wiki/Absolut_Vodka
- https://en.wikipedia.org/wiki/Red_Bull
- https://en.wikipedia.org/wiki/Hyundai_Motor_Company
- https://en.wikipedia.org/wiki/Doritos

Example 3

Question: What is the conservation status of the animal that is a popular dish in the village nearest to this mountain's west?
Question Type: IP1TP0S3 **Question Hop:** 3 **Entity Num:** 1
Answers: Least Concern.
Image Entity Names: Pedraforca
Evidence:

- Located within the Cadí-Moixeró Natural Park, Pedraforca has been declared a Natural Site of National Interest by the Generalitat de Catalunya. The closest villages to Pedraforca are Gósol to the west and Saldes to the east. Pedraforca marks the boundary between the two municipalities, as well as between the provinces of Barcelona and Lleida.
- Pèsol negre, a local variety of 'black pea'; Blat de moro escairat, or 'peeled corn', often cooked in a pork broth; Patates emmascarades, or "Masked Potatoes", mashed potatoes cooked with blood or black pudding; All i oli with pork, eaten during the swine-harvest in fall; Veal with wild mushrooms; Wild boar
- It has been assessed as least concern on the IUCN Red List due to its wide range, high numbers, and adaptability to a diversity of habitats.

Evidence URLs:

- https://en.wikipedia.org/wiki/Pedraforca
- https://en.wikipedia.org/wiki/G%C3%B3sol
- https://en.wikipedia.org/wiki/Wild_boar

Example 4

Question: Where is the highest point of the mountains in which this plant is found?
Question Type: IP1TP0S2 **Question Hop:** 2 **Entity Num:** 1
Answers: Mount Rainier.
Image Entity Names: Corallorhiza mertensiana
Evidence:

- Corallorrhiza mertensiana is found in the Cascades from Alaska to California, and the Rocky Mountains from Alberta to Wyoming.
- The highest peak in the range is Mount Rainier in Washington at 14,411 feet (4,392 m).

Evidence URLs:

- https://en.wikipedia.org/wiki/Corallorhiza_mertensiana
- https://en.wikipedia.org/wiki/Cascade_Range

Example 5

Question: Did all the people in the photo come from the same country of birth?
Question Type: IP3TP0S1 **Question Hop:** 3 **Entity Num:** 3
Answers: Yes.
Image Entity Names: Jesse Plemons, Jean Smart, Jeffrey Donovan
Evidence:

- Jesse Plemons is an American actor.
- Jean Elizabeth Smart (born September 13, 1951) is an American actress.
- Jeffrey Donovan (born May 11, 1968) is an American actor.

Evidence URLs:

- https://en.wikipedia.org/wiki/Jesse_Plemons
- https://en.wikipedia.org/wiki/Jean_Smart
- https://en.wikipedia.org/wiki/Jeffrey_Donovan

Example 6

Question: Which country is most associated with the origin of the food in the image?
Question Type: IP2TP0S1 **Question Hop:** 2 **Entity Num:** 2
Answers: Italy.
Image Entity Names: Cotoletta, lasagna
Evidence:

- Cotoletta alla Bolognese is a traditional dish of Bologna.
- Lasagne are a type of pasta, possibly one of the oldest types, made of very wide, flat sheets.

Evidence URLs:

- https://en.wikipedia.org/wiki/Cotoletta
- https://en.wikipedia.org/wiki/Lasagne

Example 7

Question: Where are this bird and Paradise shelduck native to?
Question Type: IP1TP1S1 **Question Hop:** 2 **Entity Num:** 2
Answers: Australia, New Zealand.
Image Entity Names: Australian wood duck
Evidence:

- The Australian wood duck, maned duck or maned goose (*Chenonetta jubata*) is a dabbling duck found throughout much of Australia.
- The paradise shelduck (*Tadorna variegata*), also known as the paradise duck, or pūtangitangi in Māori, is a species of shelduck, a group of goose-like ducks, which is endemic to New Zealand.

Evidence URLs:

- https://en.wikipedia.org/wiki/Australian_wood_duck
- https://en.wikipedia.org/wiki/Paradise_shelduck

Example 8

Question: Who are the owners of this building, Providence Park and Oxburgh Hall?
Question Type: IP1TP2S1 **Question Hop:** 3 **Entity Num:** 3
Answers: Anbang Insurance Group, Portland, National Trust.
Image Entity Names: Hotel del Coronado
Evidence:

- In March 2016, Blackstone sold Strategic Hotels & Resorts to Anbang Insurance Group, a Beijing-based Chinese insurance company, in a $6.5 billion deal involving multiple resorts.
- Providence Park (formerly Jeld-Wen Field; PGE Park; Civic Stadium; originally Multnomah Stadium; and from 1893 until the stadium was built, Multnomah Field) is an outdoor soccer venue located in the Goose Hollow neighborhood of Portland, Oregon.
- The Bedingfelds gained the manor of Oxborough through marriage in the early 15th century, and the family has lived at the hall since its construction, although ownership passed to the National Trust in 1952.

Evidence URLs:

- https://en.wikipedia.org/wiki/Hotel_del_Coronado
- https://en.wikipedia.org/wiki/Providence_Park
- https://en.wikipedia.org/wiki/Oxburgh_Hall