MSR^2: A Benchmark for Multi-Source Retrieval and Reasoning in Visual Question Answering

Kuo-Han Hung* **Hung-Chieh Fang*** **Chao-Wei Huang** **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
{b0990120, b09902106, f07922069}@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

This paper introduces MSR^2, a benchmark for multi-source retrieval and reasoning in visual question answering. Unlike previous knowledge-based visual question answering datasets, MSR^2 focuses on questions involving multiple fine-grained entities, providing a unique opportunity to assess a model's spatial reasoning ability and its capacity to retrieve and aggregate information from various sources for different entities. Through comprehensive evaluation using MSR^2, we gain valuable insights into the capabilities and limitations of state-of-the-art large vision-language models (LVLMs). Our findings reveal that even state-of-the-art LVLMs struggle with questions requiring multi-entities and knowledge-intensive reasoning, highlighting important new directions for future research. Additionally, we demonstrate that enhanced visual entity recognition and knowledge retrieval can significantly improve performance on MSR^2, pinpointing key areas for advancement.[1]

1 Introduction

Knowledge-based visual question answering (KB-VQA) is a challenging visual question answering task that requires integration of external knowledge. It assesses a model's ability to recognize entities within images, interpret spatial relationships between them, and retrieve relevant information from a knowledge corpus to answer questions accurately.

There are several existing KBVQA datasets. Early datasets (Wang et al., 2017; Marino et al., 2019; Jain et al., 2021; Schwenk et al., 2022) typically involves questions requiring commonsense knowledge. This requirement made retrieval necessary for models at that time to answer the questions. However, due to the emergence of large vision language models (LVLMs) (Chen et al., 2023a; Li et al., 2023a; Dai et al., 2023; Achiam et al., 2023), the knowledge required by earlier datasets has become too simple for LVLMs. Recent KBVQA datasets (Mensink et al., 2023; Lin et al., 2023; Chen et al., 2023b) have increased the complexity of questions, making them challenging for LVLMs to answer directly. Nevertheless, due to the difficulty of annotating these datasets, these datasets still focus on single entity, limiting their applicability to more complex, real-world scenarios.

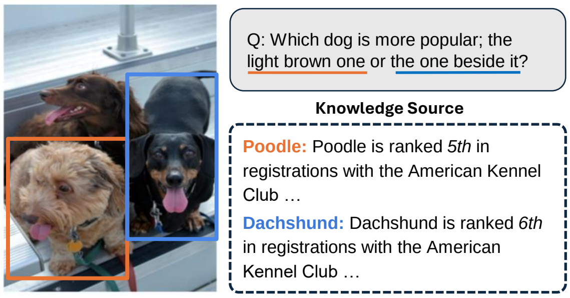


Figure 1: MSR^2 requires an understanding of spatial relationships and the ability to retrieve information from various sources for different entities.

In this work, we explore the question: *Can current LVLMs handle questions involving multiple entities that require information retrieval?* To answer this, we propose a dataset with the following characteristics, as illustrated in Figure 1:

- Questions should reference *multiple* entities within the image, requiring the model to integrate information from diverse sources. For example, identifying the light brown dog requires knowledge about Poodles, while the dog beside it corresponds to a Dachshund.
- Questions should emphasize *spatial* relationships. For example, "the light brown dog and the one next to it" requires the model to understand the arrangement of the dogs.
- Questions should involve *knowledge* that is not based on commonsense, so the model

[1] https://github.com/MiuLab/MSR-VQA

Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing (KnowledgeNLP'25), pages 259–271

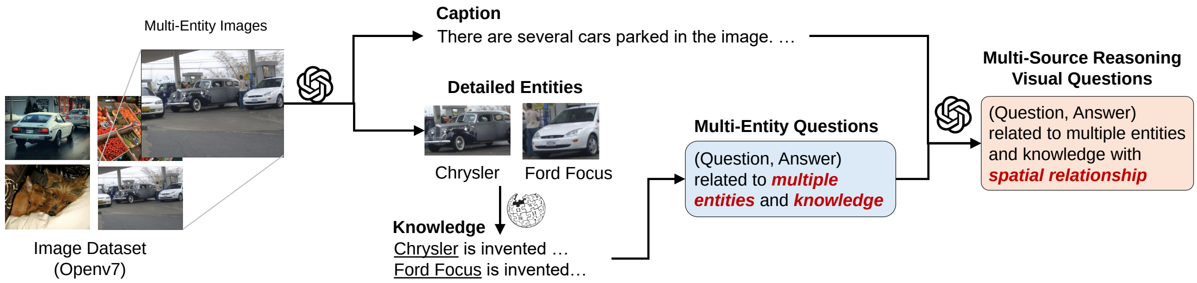


Figure 2: Data generation pipeline for MSR^2.

needs to retrieve external information beyond the image content. For example, the popularity of a dog breed may vary over time.

We evaluate several state-of-the-art LVLMs and pipeline baselines, including an entity tagging model followed by an LLM. Our results reveal that current models struggle in recognizing fine-grained entities and exhibit poor performance in spatial reasoning involving multiple entities. Additionally, we demonstrate that performance significantly improves when entity recognition is more accurate and supported by external knowledge sources. The dataset will be released publicly upon acceptance.

2 Dataset Construction

We present our data generation pipeline in Figure 2. Below are the detailed steps for constructing the MSR^2 dataset.

Image Source We utilize the Openv7 dataset (Kuznetsova et al., 2020) as our source of images. This dataset originally includes images accompanied by bounding boxes with coarse labels. To align with our objective of analyzing multi-entity images, we apply the following filtering criteria: (1) Each selected image must contain multiple objects with the same coarse, broad label; (2) We focus on a limited set of categories–AIRCRAFT, AIRPLANE, ANIMAL, CAR, CAT, DOG, DOLPHIN, INSECT, MOTORCYCLE, VEGETABLE, MUSICAL INSTRUMENT, SHARK, HORSE, FRUIT, WEAPON, TRUCK, TOOL, and FISH, since most other labels lack the fine-grained categorization necessary for our subsequent analysis.

Entity Finding After filtering the images, our next step is to identify these entities and filter those relevant for VQA generation. For each image retained from the previous step, we employ GPT-4V (Achiam et al., 2023) to generate fine-grained object labels by querying the model with object images cropped from the bounding boxes. Once all entities within an image are tagged, we retain only those images that contain distinct fine-grained labels. In addition, we also apply filtering to check whether the labeled fine-grained object labels match the original coarse label type.

Knowledge Retrieval Next, we perform knowledge retrieval for each entity by querying relevant wiki titles and their corresponding contents. We use BM25 (Robertson and Zaragoza, 2009), a traditional sparse retrieval method, to select the top-k passages. These passages are then filtered using GPT-4 (Achiam et al., 2023), which evaluates their relevance to the entity. As a result, for each entity, we retain the top-k' passages. In our implementation, k and k' is set to 50 and 1, respectively.

Question Generation With the entity names and their corresponding knowledge, we proceed to generate the corresponding questions. We utilize GPT-4 (Achiam et al., 2023) to generate these questions by providing the model with the entity labels and their associated knowledge.

Visual Question Generation In order to incorporate the visual information into the questions, we first generate image captions using GPT-4V (Achiam et al., 2023). Next, we query GPT-4 (Achiam et al., 2023) to replace the entities mentioned in the question-answer pair with the corresponding objects identified in the image captions.

LLM/VLM Filtering To ensure dataset quality, we utilize various GPT-based filtering mechanisms for entity extraction, question generation, and visual question generation.

Human Filtering To ensure the quality of our dataset, we have human evaluators on Amazon MTurk filter out any data that is incorrect or insufficiently natural after generation. Given the complexity of our data, we divide the human evalua-

Dataset	Fine-grained Entity	Knowledge Retrieval	Multiple Entities
FVQA (Wang et al., 2017)	✗	✗	✓
OKVQA (Marino et al., 2019)	✗	✗	✓
S3VQA (Jain et al., 2021)	✗	✓	✗
A-OKVQA (Schwenk et al., 2022)	✗	✗	✓
Encyclopedic VQA (Mensink et al., 2023)	✓	✓	✗
InfoSeek (Chen et al., 2023b)	✓	✓	✗
Ours: MSR^2	✓	✓	✓

Table 1: In comparison to existing knowledge-based VQA datasets, we focus on three primary aspects. (1) Fine-grained Entities: whether the model recognizes specific entities or relies on broad categories; (2) Knowledge Retrieval: whether external knowledge is needed or only image-based information suffices; and (3) Multiple Entities: whether questions involve multiple entities in the image.

tion into two steps: (1) Image Labels Reference: This step checks the correctness of entity labeling and the associated references. (2) Knowledge-Based QA Validation: This step verifies whether the provided knowledge source correctly answers the question and whether the answer itself is accurate. The evaluation user interfaces for the Mechanical Turk workers are shown in Figures 3. Only data that passes both evaluations is included in our final dataset. Originally, our dataset contained 2.8k entries; after human filtering, we retained 1.3k entries.

For further details on the data generation and filtering, please refer to Appendix A.1.

3 MSR^2: Benchmarking Multi-Source Retrieval and Reasoning in Visual Question Answering

3.1 Dataset Statistics

We compare the statistics of our dataset with those of recently proposed datasets that share some similar characteristics with MSR^2, as shown in Table 2. Note that we focus exclusively on the test set, as we aim to evaluate LVLM's zero-shot capabilities. K-VQA (Shah et al., 2019) is a multi-entity dataset that requires understanding relationships between entities to provide answers. However, its entity types are limited to humans, restricting its applicability across different domains. Encyclopedic VQA (Mensink et al., 2023) and InfoSeek (Chen et al., 2023b) are both datasets that require fine-grained entity and knowledge retrieval. However, their questions and images primarily focus on single entities, limiting their effectiveness on testing spatial reasoning.

3.2 Evaluation Metrics

Previous work primarily relied on VQA accuracy (Goyal et al., 2017) as the evaluation metric. However, Mañas et al. (2024) highlighted that VQA accuracy can be overly rigid, often marking correct answers as incorrect due to formatting discrepancies. To address this, they proposed using LLM-based evaluation for reliable accuracy. Building on this approach, we utilize GPT-4 as the evaluator to assess VQA performance. Details of the evaluation prompts are provided in Appendix A.2.

3.3 Qualitative Analysis

We show several random examples and quality assessment of our dataset in Figure 4 and Appendix A.3. This dataset offers a broad range of object categories (e.g., cars, airplanes, animals) and scenes (e.g., outdoor shows, hangars, parks), fostering comparative visual reasoning through questions about foreground vs. background objects and attributes like historical significance or function. Its strength lies in filtering overly specialized subcategories while retaining sufficient detail for tasks such as distinguishing car models or dog breeds. However, due to the nature of the dataset, some images show partially occluded or out-of-frame entities, leading to ambiguous tagging and inaccurate identification—especially when key distinguishing features fall outside the frame or are blocked by other objects. This limitation can hinder tasks requiring fine-grained classification or detailed object-specific reasoning. Despite these challenges, the dataset remains a rich multimodal resource for VQA, reference resolution, and spatial reasoning, provided that annotations and bounding

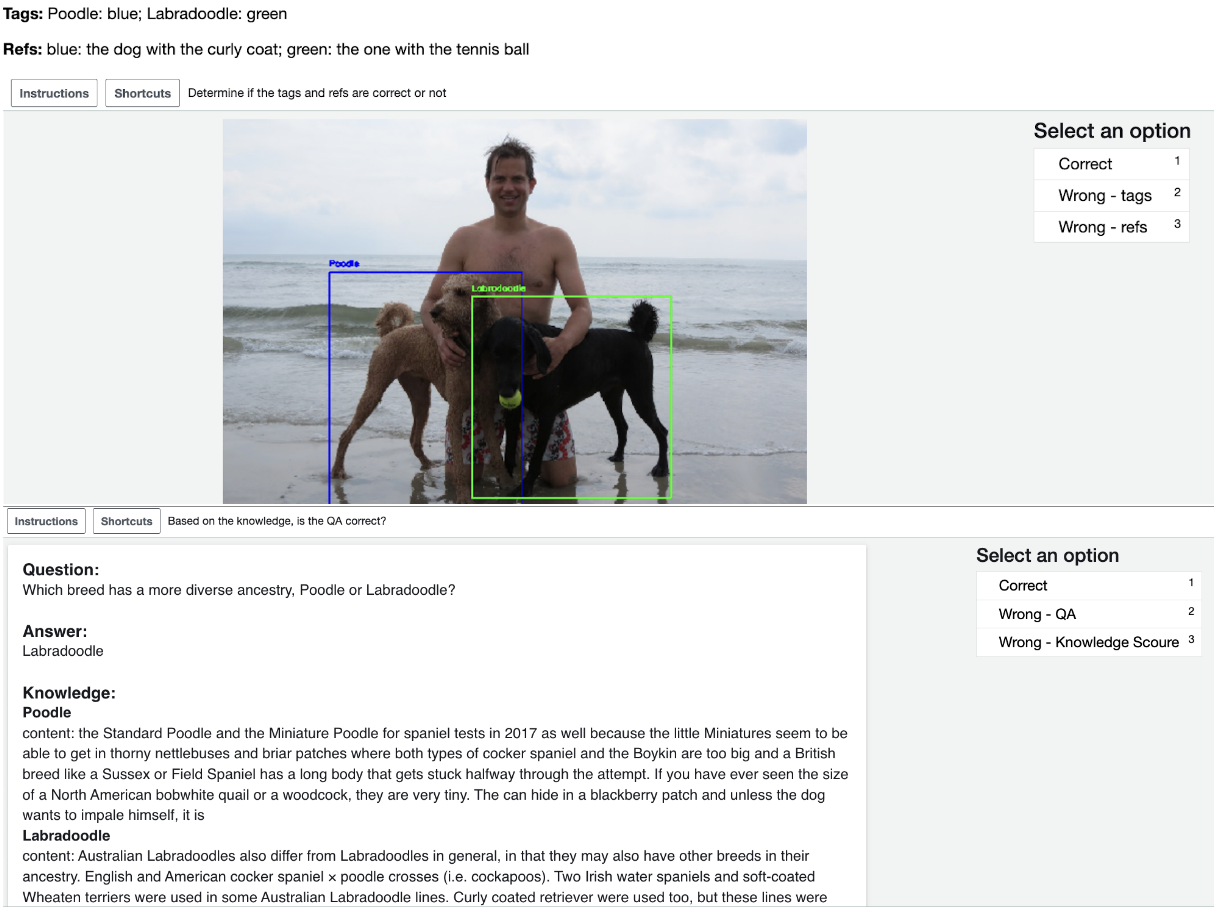


Figure 3: UI of human filtering for Mturk human evaluation. *Top:* Filtering of tags. *Bottom:* Filtering of generated questions and answers based on the provided knowledge.

boxes are carefully maintained and extended metadata is considered to address issues of ambiguity and partial visibility.

4 Experiments

4.1 Tested Models and Settings

We adopt the evaluation method from InfoSeek (Chen et al., 2023b), which includes an end-to-end approach without knowledge retrieval and a pipeline approach with knowledge retrieval.

Large Models without Retrieval We assessed existing LVLMs—BLIP2 (Li et al., 2023a), LLaVA (Liu et al., 2024), and GPT-4V (Achiam et al., 2023)—to evaluate their ability to answer VQA questions without external knowledge sources.

Large Models with Retrieval Following Chen et al. (2023b), we first use CLIP (Radford et al., 2021) to tag the visual entities. Then, an LLM/LVLM (GPT4-V in our case) is employed to answer the question, leveraging knowledge either within its parameters or from an external source.

We also include oracle toplines in our ablation studies to evaluate the model's performance in identifying fine-grained entities, spatial reasoning, and knowledge coverage. Two methods are used to incorporate entities: (1) entities are provided without being mapped to the question, and (2) entities are provided and mapped to the question. This setup allows us to evaluate the model's spatial reasoning, specifically whether it can accurately map entities to their corresponding references in the question.

4.2 Evaluation Results

As shown in Table 3, existing LVLMs perform poorly on MSR[2], achieving only a 10% improvement over the random baseline. Furthermore, pipeline methods, which first identify entities and then use an LVLM to answer, demonstrate even worse performance. We further discuss the results from the following aspects:

Existing models fail to identify fine-grained entities. The oracle baselines demonstrate an improvement of 15.9% when entity recognition is

Dataset	# {Q, I}	Avg # Ent. per I	# Ent. Type	Rationale
K-VQA (Shah et al., 2019)	183k	> 1	1	✗
Encyclopedic VQA (Mensink et al., 2023)	5.7k	1	2.1k	✗
InfoSeek$_{Human}$ (Chen et al., 2023b)	8.9k	1	527	✗
Ours: MSR^2	1.3k	2.25	53	✓

Table 2: Dataset Statistics. Q: Questions; I: Images; Ent.: Entities. The test set is used for comparison.

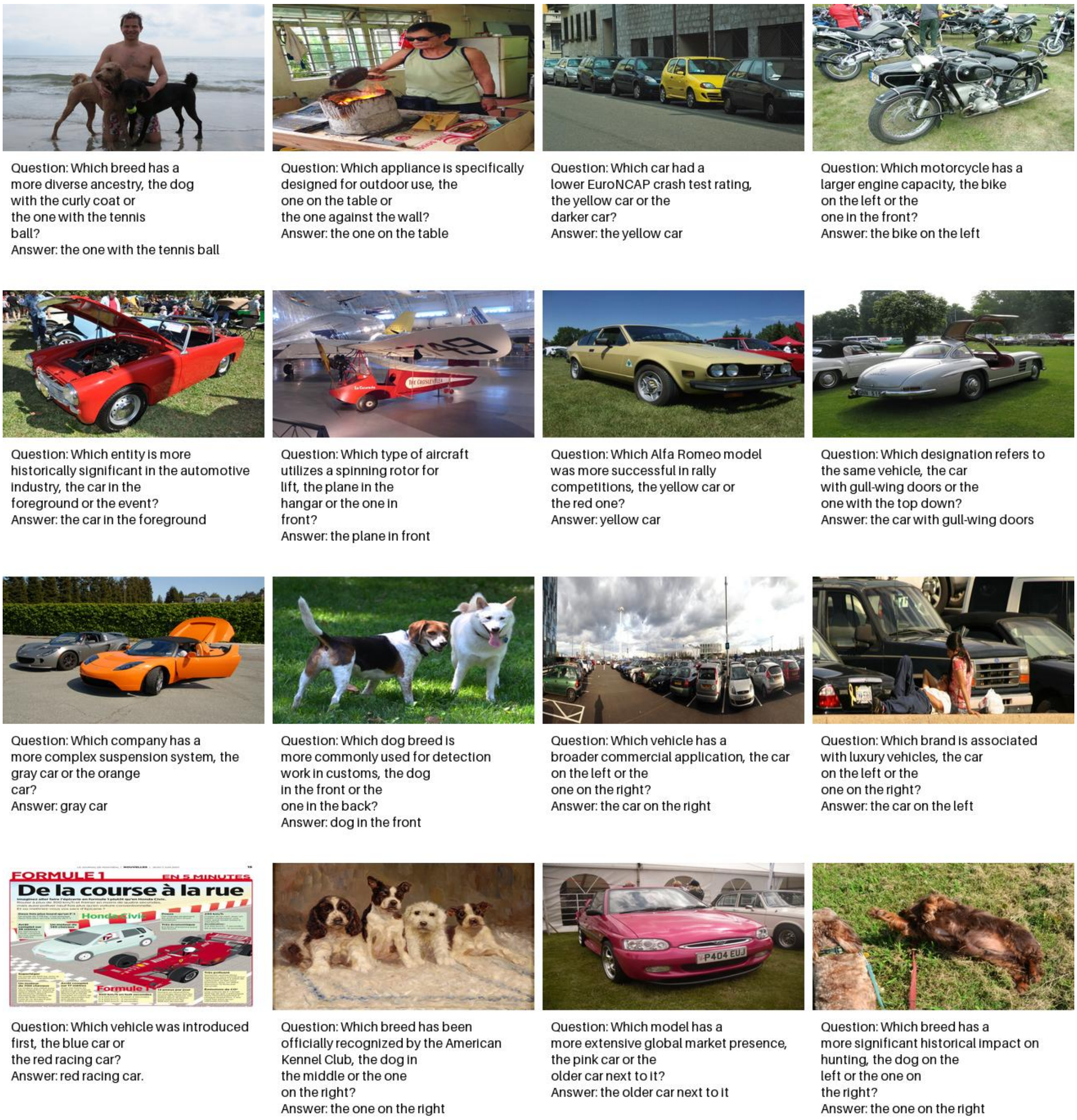


Figure 4: Random examples VQA question of MSR^2.

accurate. This highlights the limitations of LVLMs in identifying fine-grained entities.

Since the image contains multiple entities, pipeline methods using CLIP to compute embeddings for the image and match them to the closest entity embedding may be too coarse, potentially missing the details of individual entities.

Model	Accuracy
Without KB	
Random	50.00
BLIP2 (Li et al., 2023a)	54.45
LLaVA (Liu et al., 2024)	53.05
GPT4-V (Achiam et al., 2023)	62.47
With KB	
CLIP → GPT4-V (parameter)	51.73
CLIP → GPT4-V (wiki)	57.86
Oracle ent. → GPT4-V (parameter)	63.96
Oracle ent. → GPT4-V (wiki)	69.35
Oracle → GPT4-V (parameter)	76.97
Oracle → GPT4-V (wiki)	81.44

Table 3: Main results on MSR^2 (%). The "Oracle ent." toplines provide the entity without mapping it to the question, whereas the "Oracle" toplines include both the entity and its mapping to the question.

LVLMs are poor at spatial reasoning. We compare the performance of 'Oracle ent.' to 'Oracle' to evaluate the spatial reasoning ability of LVLM. The results show that providing entities improves performance by 6.9% compared to GPT4-V, where no entities are given. However, there is an 12.1% performance gap between the 'Oracle' toplines (where entities are mapped to the question) and 'Oracle ent.', indicating that LVLM struggles with correctly mapping entities back to the questions.

External knowledge can further boost performance. The 'Oracle → GPT-4 (parameter)' approach shows a significant improvement over existing baselines, demonstrating that a large number of questions can be effectively answered using the knowledge encoded within the model's parameters. Additionally, integrating external knowledge from Wikipedia further boosts performance by 4.47%, highlighting the importance of the external knowledge.

4.3 Qualitative Study

In Figure 5, we study two different types of errors. The top image illustrates that answering more precise questions (e.g., identifying a specific span) requires verifying information across multiple sources. The bottom image reveals a failure in entity mapping, where the model struggles to link the correct entity to the question despite possessing accurate knowledge.

5 Related Work

Visual Question Answering. Visual Question Answering (VQA) is a long-standing problem where models must answer questions based on a given image. There have been numerous benchmark datasets proposed for the VQA task, including VQAv1 (Antol et al., 2015), VQAv2 (Goyal et al., 2017), DAQUAR (Malinowski and Fritz, 2014), FMIQA (Gao et al., 2015) and Visual Madlibs (Yu et al., 2015).

Knowledge-based VQA. Knowledge-based VQA is a type of VQA where the questions require external knowledge beyond the image content. Some of the well-known benchmarks include OKVQA (Marino et al., 2019), which focuses on open-ended questions that require world knowledge, FVQA (Wang et al., 2017), where answers are derived from facts stored in a structured knowledge base, S3VQA (Jain et al., 2021), which involves selective question answering using different knowledge sources, and A-OKVQA (Schwenk et al., 2022), which emphasizes the reasoning ability of models. Several solutions have been proposed to solve these problems (Salemi et al., 2023; Lin and Byrne, 2022; Lin et al., 2022; Ma et al., 2024; Si et al., 2023; Hu et al., 2023; Lin et al., 2024; Hu et al., 2024). Most models use a two-stage approach: retrieving relevant knowledge then answering the question. However, much of the knowledge required by these benchmarks is commonsense, which recent (LVLMs) can handle directly through their embedded knowledge.

Knowledge-based VQA with LVLMs. Prior research has demonstrated LVLMs' capabilities for knowledge-intensive VQA. Li et al. (2023b) provides a comprehensive evaluation of GPT-4V on its knowledge-based VQA capabilities. PROOFREAD (Zhou et al., 2023) and Prophet (Shao et al., 2023) leverage LLMs as implicit knowledge engines, paired with visual-language models to filter irrelevant or noisy information and improve accuracy. Uehara et al. (2024) introduces a chain-of-thought framework with question-asking capabilities, leveraging LLM-generated datasets for knowledge-based VQA. Xu et al. (2024) proposes a novel reverse image retrieval approach for improving multimodal LLM performance on knowledge-intensive VQA tasks. Cao and Jiang (2024) uses LLMs as a knowledge source which generates relevant knowledge snippets for the downstream VQA task. To evaluate LVLMs on knowledge-intensive

Q: Which vehicle had a longer production span, the blue car on the right or the white cars to its left?

Oracle ent. -> GPT-4V (parameter)

- **Answer:** The blue car on the right (Borgward Isabella).
- **Reasoning:** The Borgward Isabella (blue car) was produced from 1954 to 1962, while the Ford Consul Cortina I (white car) was produced from 1962 to 1966

Oracle ent. -> GPT-4V (wiki)

- **Answer:** The white cars on the left.
- **Reasoning:** The Ford Consul Cortina (white car) had a longer production span, from 1962 to 1970, compared to the Borgward Isabella (blue car), which

---

Q: Which compact car was introduced later, the car in red or the car in black?

Oracle ent. -> GPT-4V (wiki)

- **Answer:** The car in red.
- **Reasoning:** The Chevrolet Cobalt was introduced later, in 2004 for the 2005 model year. The Honda Civic was introduced earlier, with its second generation debuting in June 1979 for the 1980 model year.

Oracle -> GPT-4V (wiki)

- **Answer:** The car in black was introduced later.
- **Reasoning:** The Chevrolet Cobalt was introduced in 2004 as a 2005 model, while the Honda Civic was first introduced in 1972, making the Cobalt the later introduction among

Figure 5: Qualitative analysis of two different error types: (1) **Insufficient Multi-Source Knowledge**: The top image illustrates that answering more precise questions requires integrating knowledge from multiple sources. **Entity Mapping Failure**: The bottom image demonstrates the model's inability to correctly map the relevant entity to the question, despite having accurate knowledge.

and information-seeking questions that go beyond visual content, we need more rigorous benchmarks with detailed knowledge. Recent benchmarks like Infoseek (Chen et al., 2023b) and Encyclopedic VQA (Mensink et al., 2023) assess models' capabilities in fine-grained object recognition and answering rare questions about those objects. Building on these efforts, we introduce a new benchmark with multi-entity, knowledge-intensive, and spatial reasoning questions.

6 Conclusion

We introduce MSR^2, a VQA dataset focused on KBVQA questions involving multiple entities, requiring both multi-retrieval and spatial reasoning. Our experiments demonstrate that MSR^2 presents a substantial challenge for standard LVLMs. However, incorporating an oracle retrieval component significantly enhances performance. We anticipate that MSR^2 will inspire future research into more generalized retrieval-augmented LVLMs.

Limitations

MSR^2 is limited to English; future research could extend it to a multilingual setting. Additionally, the image sources employed in our study lack sufficient diversity—particularly regarding images containing multiple objects within the same broad category.

This limitation may affect the quality and diversity of the generated dataset. Future work should explore more varied and representative image datasets that include multiple instances of different objects within the same category to improve the robustness and generalizability of the approach.

Acknowledgements

We thank the reviewers for their insightful comments. This work was financially supported by the National Science and Technology Council (NSTC) in Taiwan, under Grants 111-2222-E-002-013-MY3 and 112-2223-E002-012-MY5. We thank the National Center for High-performance Computing (NCHC) of National Applied Research Laboratories (NARLabs) in Taiwan for providing computational and storage resources.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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.

Rui Cao and Jing Jiang. 2024. Knowledge generation for zero-shot knowledge-based VQA. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 533–549, St. Julian's, Malta. Association for Computational Linguistics.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme Ruiz, Andreas Peter Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023a. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. 2023b. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore. Association for Computational Linguistics.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided image captioning for vqa with gpt-3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2963–2975.

Yushi Hu, Otilia Stretcu, Chun-Ta Lu, Krishnamurthy Viswanathan, Kenji Hata, Enming Luo, Ranjay Krishna, and Ariel Fuxman. 2024. Visual program distillation: Distilling tools and programmatic reasoning into vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9590–9601.

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.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

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

Yunxin Li, Longyue Wang, Baotian Hu, Xinyu Chen, Wanqi Zhong, Chenyang Lyu, Wei Wang, and Min Zhang. 2023b. A comprehensive evaluation of gpt-4v on knowledge-intensive visual question answering. *arXiv preprint arXiv:2311.07536*.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *arXiv preprint arXiv:2210.03809*.

Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. 2024. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*.

Weizhe Lin, Zhilin Wang, and Bill Byrne. 2023. FVQA 2.0: Introducing adversarial samples into fact-based visual question answering. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 149–157, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering. *Advances in Neural Information Processing Systems*, 35:10560–10571.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ziyu Ma, Shutao Li, Bin Sun, Jianfei Cai, Zuxiang Long, and Fuyan Ma. 2024. Gerea: Question-aware prompt captions for knowledge-based visual question answering. *arXiv preprint arXiv:2402.02503*.

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.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

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.

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

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 110–120, New York, NY, USA. Association for Computing Machinery.

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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8876–8884.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14974–14983.

Qingyi Si, Yuchen Mo, Zheng Lin, Huishan Ji, and Weiping Wang. 2023. Combo of thinking and observing for outside-knowledge vqa. *arXiv preprint arXiv:2305.06407*.

Kohei Uehara, Nabarun Goswami, Hanqin Wang, Toshiaki Baba, Kohtaro Tanaka, Tomohiro Hashimoto, Kai Wang, Rei Ito, Takagi Naoya, Ryo Umagami, et al. 2024. Advancing large multi-modal models with explicit chain-of-reasoning and visual question generation. *arXiv preprint arXiv:2401.10005*.

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.

Jialiang Xu, Michael Moor, and Jure Leskovec. 2024. Reverse image retrieval cues parametric memory in multimodal llms. *arXiv preprint arXiv:2405.18740*.

Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual madlibs: Fill in the blank image generation and question answering. *arXiv preprint arXiv:1506.00278*.

Yang Zhou, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2023. Prompting vision language model with knowledge from large language model for knowledge-based vqa. *arXiv preprint arXiv:2308.15851*.

A Appendix

A.1 Details for data generation

The following section are the prompts for different stage of our generation pipeline.

Entity Finding The following are the prompts for entity finding.

```
Given the object, you have to generate one question to gain a more detailed class of the object. The answer of the question should be the detailed class of the object.

Examples:
{examples}
Object: {label}
Question:
```

Listing 1: Prompt for entity finding query generation

```
{generated_query} Answer with a  noun.
```

Listing 2: Prompt for entity finding

```
Decide whether the statement is true.
Examples:
Question: Panther is a type/class of Giraffe,
Answer: False
{more examples}
Question: {tag} is a type/class of {label}
Answer:
```

Listing 3: Prompt for entity filtering - subclass

```
Given a tag list, decide whether the tag list contains multiple different entities.
Examples:
Entities:['volkswagen t1', 'audi a4']
Answer: True
{more examples}
Entities:{tags}
Answer:
```

Listing 4: Prompt for entity filtering - different tags

Question Generation The following are the prompts for question generation.

```
You are a knowledge-based question answer generator. Given the objects and knowledge of each objects, generate a question and answer with rationale and a short answer.
Rules:
1. Answer should be a word, not a sentence.
2. Only ask one short question.
3. Question should be generated based on the object and knowledge.
4. Question should be related to at least two objects and the object must be in the Object List.
5. Question should be hard, do not ask common question that can be easily answered without knowledge source.
6. **All the options in the question and answer should be in the Objects List, question should contain the choices. i.e. _____, A or B?. Both A and B should in the Object List**
7. Do not output Objects List and Knowledge, only output Question, Rationale and Answer.
Format: {...}
Examples: {examples}
Objects List: {objects_list}
Knowledge: {knowledge}
```

Listing 5: Prompt for QA generation

```
Decide whether the QA question follow this criteria.
1. All the entities in the question are in the object list, it can be a slightly calling difference
2. The question contains more than one entities. If the provided question and object list satisfy the criteria above, output True Otherwise output False. Do not output any other information other than True or False.
Question: {question}
Object List: {objects_list}
```

Listing 6: Prompt for QA filtering

Visual Question Generation The following are the prompts for visual question generation.

```
There are {tags} in the image.
Describe their (1) appearance (2) place it located (3) other objects/people that are related to this object in the image.
Do not describe objects that are not related to the provided object list.
Write the response in a short passsage.
```

Listing 7: Prompt for image captioning

```
You are a VQA rewriter. Given a QA question and an image caption, rewrite the part after the comma in the question to create a more natural and human-like visual question answering format.

Rules:
1. Rewrite the entities in both the answer and the part of the question after the comma, using the visual information provided in the image.
```

```
2. The part of the question before the
    ↪ comma should remain unchanged.
3. Rewrite with simpler words and fewer
    ↪  object details.

Format: {...}
Examples: {examples}
Caption: {caption}
Question: {question}
Answer: {answer}
```

Listing 8: Prompt for VQA generation

A.2 Details for evaluation

The following are the prompts for model evaluation.

```
Given a question, a prediction, and an
    ↪ answer, evaluate whether the
    ↪ prediction aligned with the
    ↪ answer based on the question.
    ↪ Answer with Yes or No.

Question: {question}
Prediction: {prediction}
Answer: {answer}
```

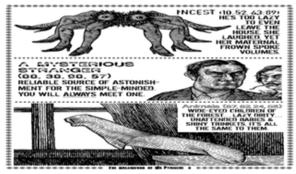
Listing 9: Prompt for model answer evaluation

A.3 Example data of MSR^2

Figure 4, 6 and 7 contain some random example data of MSR^2.

Question: Which dog breed is more commonly kept as a companion, the dog with the shaggy coat or the one with the smooth coat?
Answer: the one with the smooth coat

Question: Which ingredient is more commonly used in baking, the greens on the counter or celery?
Answer: the greens on the counter



Question: Which animal is known for forming social groups, the one at the top or the one at the bottom?
Answer: the one at the bottom

Question: Which vehicle is historically recognized as a pioneer of the muscle car era, the car in the front or the one in the back?
Answer: the car in the front

Question: Which breed is known for its semi-feral populations, the horse or the one near the feeder?
Answer: the one near the feeder

Question: Which vehicle is associated with a broader range of models, the car in front or the one next to it?
Answer: the one next to it

Question: Which fruit is known for its creamy texture and has been praised by Mark Twain, the green fruit in the corner or the round fruit next to it?
Answer: the green fruit in the corner

Question: Which vehicle has a broader international racing presenc car on the left or the one on the right?
Answer: the one on the right

Question: Which model has a more significant focus on performance features, the shiny car on the left or the one on the right?
Answer: the one on the right

Question: Which type of residence typically has a building superint the tall building or the houses next to it?
Answer: tall building

Question: Which aircraft was introduced first, the yellow-winged pla the high-winged one?
Answer: yellow-winged plane

Question: Which vehicle was associated with a motor racing competi in New Zealand and Australia, the car shown or the one not visible?
Answer: the car shown

Question: Which cat breed is known for its white spotting gene, the cat lying down or the one sitting up?
Answer: the one sitting up

Question: Which vehicle has a more significant historical impact in motorsport, the car on the right or the one on the left?
Answer: the car on the right

Question: Which name is more commonly used in scientific literatu the eel in the image or the other one?
Answer: the eel in the image

Question: Which breed is more commonly kept as a companion dog, the dog on the left or the dog on the right?
Answer: the dog on the left

Question: Which vegetable has a longer history of cultivation, the one in the bag or the one in bunches?
Answer: the one in bunches

Question: Which cat breed is mentioned in the Tamra Maew, the white cat or the orange and white one?
Answer: the white cat

Question: Which model was first imported into the U.K, the car in the front or the one next to it?
Answer: the car in the front

Question: Which car model is associated with a V8 engine, the blue car or the pink car?
Answer: the pink car

Question: Which breed is primarily known for its role as a sled dog, the dog on the left or the dog on the right?
Answer: dog on the right

Question: Which vehicle is primarily used for emergency situations, one with equipment or the one with a big front grille?
Answer: the one with equipment

Question: Which fruit is known for its creamy texture, the fruit on the left or the one in the middle?
Answer: the one in the middle

Question: Which type of residence is recognized for its historic significance in the US, the house on the corner or the ones along the street?
Answer: the house on the corner

Question: Which vehicle is designed primarily for leisure and passe comfort, the red sports car or the grey van?
Answer: the grey van

Question: Which fish is known for having a luminescent organ for attracting prey, the fish in the image or another type?
Answer: the fish in the image

Question: Which breed is known for having a rare genetic mutation causing white spotting, the big dog or the small dog?
Answer: big dog

Question: Which vehicle features a V8 engine, the red car or the maroon car?
Answer: maroon car

Figure 6: Random examples VQA question of MSR^2 - group2

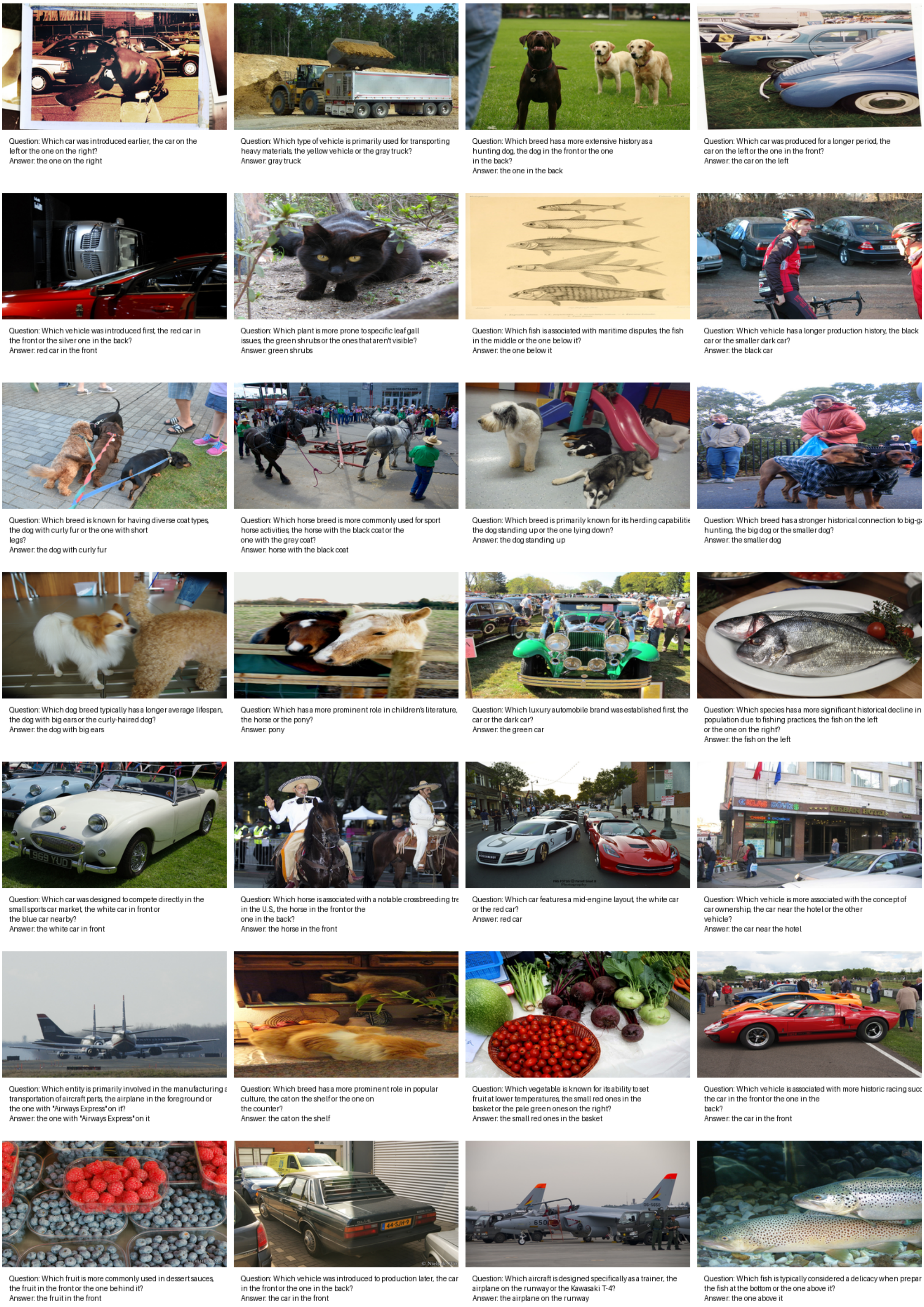


Figure 7: Random examples VQA question of MSR^2 - group3