Exploring Visual Multiple-Choice Question Answering with Pre-trained Vision-Language Models

Gia-Nghia Tran[1,4], Duc-Tuan Luu[1,2,3,4], and Dang-Van Thin[1,4]

[1] University of Information Technology, VNU-HCM, Vietnam
[2] University of Science, VNU-HCM, Vietnam
[3] John von Neumann Institute, VNU-HCM, Vietnam
[4] Vietnam National University, Ho Chi Minh City, Vietnam
{nghiatg,tuanld,thindv}@uit.edu.vn

Abstract. Visual question answering is a challenging task in computer vision and natural language processing that involves answering questions about an image using both visual and textual information. This task is more challenging when it comes to the Japanese language since there is a lack of research focus on Japanese compared to extensive studies for English and other languages. The ACCV Workshop on Large Vision – Language Model Learning and Applications (LAVA) has organized an interesting challenge that aims at benchmarking different systems on the multiple-choice visual question answering task across both Japanese and English. In this paper, we present a simple yet effective approach that competes in this LAVA Workshop Challenge. To provide a correct answer, our proposed framework needs to (1) Identify entities and understand the visual concepts and the underlying spatial relations in the image referred to in the question, (2) Align the multimedia representations of the visual content with the multiple-choice answers to determine the most accurate response. We believe that the size of the vision-language model affects the overall performance of the proposed system.

Keywords: LAVA Workshop Challenge · Vision-Language Model · Visual Question Answering

1 Introduction

Recent developments in the domains of artificial intelligence and machine learning have gained significant attention, specifically towards the advancement of large language models (LLMs) [1, 3, 18, 37, 43]. These sophisticated models have demonstrated exceptional capabilities in the processing and interpretation of extensive amounts of textual data, leading to impressive performance across a wide range of natural language processing (NLP) tasks. The success of these models has fostered a growing interest in extending their applicability beyond textual information to incorporate other modalities, including visual and auditory data, as well as the development of multi-modal inputs. This paradigm shift has led

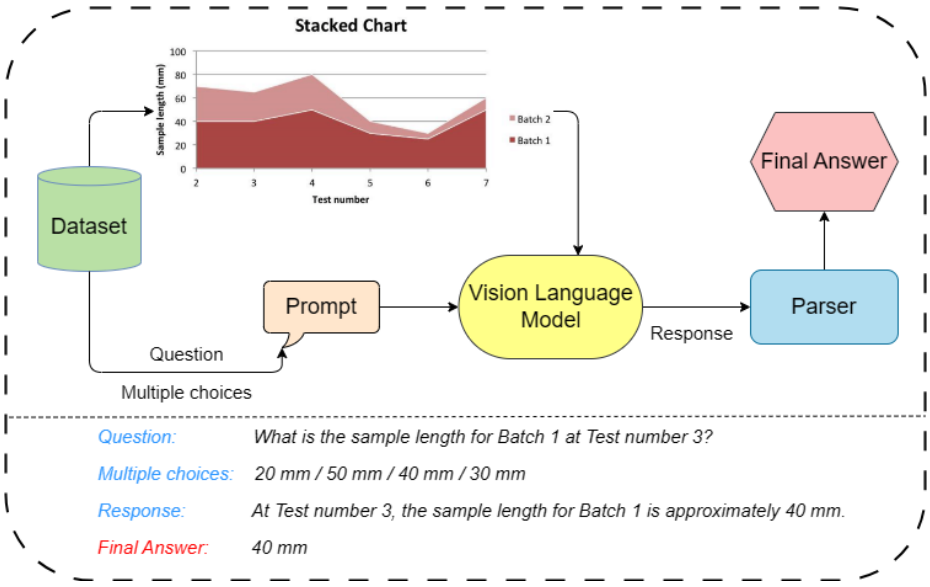

Fig. 1: Overview of our competition approach. For each sample, an image and the corresponding prompt are fed into a VLM. The model's response then goes through the parser as the post-processing step to obtain the final answer.

to the emergence of large vision-language models (VLMs), which focus on combining the strengths of both NLP and computer vision. Such integration aims to facilitate a more comprehensive and nuanced understanding of the world, bridging the gap between different forms of data.

A particular aspect in this field of research is visual question answering (VQA) [7, 21, 24]. In VQA tasks, models are requested to analyze images and subsequently answer questions regarding the visual content. This also consists of various formats, including multiple-choice VQA, in which the system is presented with a specific question alongside several potential answer options related to the visual input. The Large Vision – Language Model Learning and Applications (LAVA) workshop challenge is a newly organized competition that promotes comparative benchmarking of different multiple-choice VQA systems across complex visual data, such as diagrams, Gantt charts, building designs, and drawings.

In this challenge, participants are presented with multiple pairs of image-question. For each pair, each team must identify the most accurate answer from a set of four distinct natural language options. The dataset for this competition covers a wide range of world knowledge, featuring visual cues that present complex information (charts, diagrams, posters, drawings, etc). Moreover, the competition incorporates both English and Japanese languages in its photos and question-answers set, adding an extra layer of complexity. The unique characteristics of the Japanese language, such as its complex writing system and

rich cultural context, pose specific challenges for this multiple-choice VQA task. Therefore, participants are required to have not only sharp analytical skills but also linguistic proficiency and knowledge across multiple domains.

In this paper, we introduce our competition approach, which takes part in the LAVA 2024 challenge. To be specific, we evaluate various pre-trained VLMs and describe our strategies to achieve our competitive results. Our method can be easily integrated in a plug-and-play manner, allowing it to work with various pre-trained VLMs without the need for retraining or fine-tuning. Figure 1 demonstrates the workflow of our proposed approach. In general, each pair of image-prompt is passed through a VLM to obtain the response. It is filtered to achieve the most accurate multiple-choice answer.

2 Related Works

2.1 Multi-modal Pre-Training

The integration of multiple modalities, such as vision and language, has become increasingly prominent in artificial intelligence research. Multi-modal pre-training aims to learn unified representations from diverse data types, enabling models to perform tasks that require understanding across different modalities. This approach leverages large-scale datasets to capture the complementary information inherent in each modality, leading to enhanced performance in downstream tasks like image captioning, visual question answering, and cross-modal retrieval.

Early works in multi-modal learning focused on task-specific architectures. For instance, VQA models [12, 30] were designed to answer questions about images by combining vision encoder and text encoder. However, these models often suffered from limited generalization due to their reliance on task-specific data and architectures. The advent of transformer architectures [44] and the success of large-scale language models spurred the development of multi-modal transformers. ViLBERT [29] and LXMERT [40] extended the BERT model [20] to handle both visual and textual inputs by learning joint representations through co-attention mechanisms. CLIP [36] and ALIGN [17] employed contrastive learning techniques to align visual and textual representations in a shared embedding space. By training on large-scale image-text pairs collected from the internet, these models achieved remarkable zero-shot performance without explicit region-based features. Building upon this foundation, UNITER [8], OSCAR [25], and SimVLM [47] scale up pre-training data to better align visual and textual modalities. Additionally, CoCa [51] and Florence [53] integrated encoders with decoders, enabling both understanding and generation capabilities within a single, unified framework. More recently, significant progress has been made with large multi-modal models bridging vision and language at scale. GPT-4 [2], Gemini 1.5 [37,41], Llama 3 [11], demonstrated advanced cross-modal reasoning, enabling tasks like image analysis and visual question answering.

2.2 Visual Question Answering Dataset

Visual Question Answering is a complex yet essential task situated at the intersection of computer vision and natural language processing. It requires AI models to accurately answer questions based on visual input, combining the ability to interpret images with natural language comprehension. The introduction of the original VQA dataset [4] marked a significant milestone by providing the first standardized benchmark for evaluating such models. Building on its foundation, VQA 2.0 [13] addressed several limitations of the initial version, notably by enhancing the balance between questions and answers and minimizing biases, which contributed to a more robust model evaluation. Recent developments in VQA have expanded the scope and complexity of the task. New datasets now demand deeper reasoning abilities from models, moving beyond surface-level object recognition to require logical inferences, spatial reasoning, and contextual understanding [16,19,39,46,55]. Furthermore, VQA research has extended the task to different variations of visual input, including videos [48, 52], scene text [6], and documents [35, 42], broadening the applicability of the task across different domains. Similarly, specialized datasets have emerged for medical VQA [15,23,26], aimed at assisting in healthcare tasks, where questions are based on medical images such as MRIs or X-rays. Additionally, other datasets [33, 34] focus on structured data in the form of plots, figures, and graphs, demanding models to interpret and generate insights from visualized data.

2.3 Visual Question Answering Approaches

The field of Visual Question Answering (VQA) has undergone significant transformation, with deep learning techniques now serving as the foundational framework for most contemporary methodologies. Early approaches [4, 32, 38] often involved separate encoders for visual and textual data to extract features from images and questions, respectively. These features were then fused using various strategies to merge the multi-modal information. The resulting combined representation was processed by either a classifier or a generator, depending on whether the answer generation was approached as a classification task or a generative problem. In recent years, there has been a notable shift towards Vision-Language Pre-training (VLP) [1, 3, 5, 10, 11, 37, 41], utilizing transformer architectures [44]. These models are pre-trained on extensive datasets comprising image-text pairs to learn generalized representations that span both modalities. By effectively capturing the complex relationships between visual and textual inputs, they can be fine-tuned for downstream tasks like VQA, leading to marked improvements in performance and generalization capabilities. This evolution towards transformer-based VLP models has not only enhanced the accuracy of VQA systems but also expanded their ability to tackle more intricate questions that demand deeper reasoning and contextual understanding. The transformer architecture enables models to dynamically attend to different parts of the input data, thereby improving interpretability and fostering more nuanced interactions between visual and textual information. Furthermore, this progression has

```
###Prompt: <Image><I></Image>
Imagine you are an expert in English and Japanese with good
knowledge. Please provide the answer with ONE number from 1-4. You
MUST give reason/explanation for your choice. The output MUST be
json format. Use the following JSON format:
  {
    "answer": "number",
    "explanation": "<text>"
  }
The following is the question and 4 choices:
<Question><Q></Question>
1 <Option><o1><Option>
2 <Option><o2><Option>
3 <Option><o3><Option>
4 <Option><o4><Option>
```

Fig. 2: Prompt guidance for all of our VLM approaches.

opened up new avenues of research in VQA, including exploring zero-shot learning potentials [14, 22, 27], addressing inherent biases in datasets, and integrating commonsense knowledge to manage more sophisticated queries.

3 Our proposed approach

VLMs are designed to understand and generate text and images simultaneously. A crucial aspect of their performance lies in their ability to align both visual and textual feature representations. As mentioned earlier and shown in Figure 1, in our proposed framework, each pair of image-prompt passes through a specific VLM and returns the related response, containing a detailed explanation. The response is further processed to obtain the final answers.

To be specific, we input the image I along with the guidance prompt containing the question Q and four multiple-choice options $o1, o2, o3, o4$. The VLMs then extract the image feature and process it simultaneously with the text prompt. The expected result consists of the number answer 1-4 and the corresponding explanation for the selected answer. Figure 2 illustrates our full guidance prompt used for all of our models. Regarding the VLM backbone, we explore different methods in order to find the best VLM that can perform well in both Japanese and English. Each method is presented separately for comprehensive understanding.

3.1 MiniCPM-V

MiniCPM-V [50] is a multi-modal large language model (MLLM) designed for efficient deployment on mobile devices, using approximately 8 billion parame-

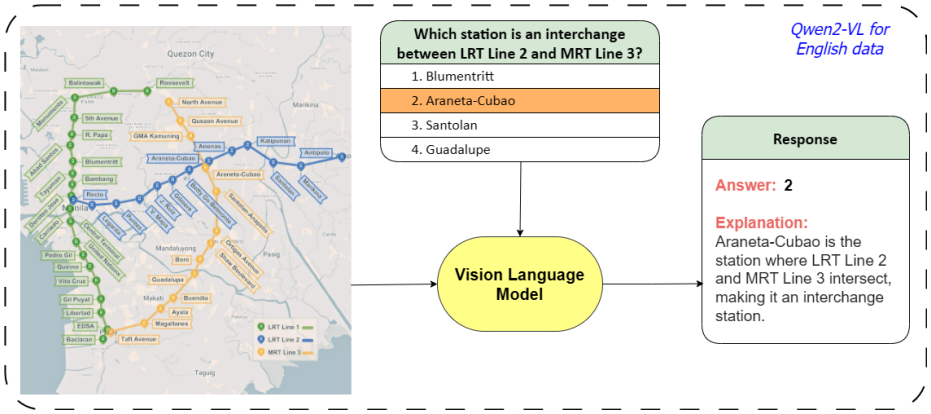

Fig. 3: An example of our best VLM approach for English data.

ters. The latest version of the MiniCPM-V series is utilized *(version 2.6)*, which is built on the SigLip-400M [54] and Qwen2-7B frameworks [49], achieving significant performance improvements over its predecessor, MiniCPM-Llama3-V 2.5 By being lightweight, this model is optimized explicitly for end-side deployment, meaning it can run efficiently on devices like mobile phones, tablets, and personal computers.

3.2 Qwen2-VL

Qwen2-VL [45] is the latest version of the VLM in Qwen model families [5, 49] developed by Alibaba Cloud. It is the current state-of-the-art on several open-source visual understanding benchmarks, such as MathVista [31], DocVQA [35]and RealWorldQA[5]. Qwen2-VL has its text encoder as QwenV2 [49], including dense models and a mixture-of-experts model. These models are designed to perform well in multilingual settings, supporting over 29 languages, with the number of parameters ranging from 0.5 billion to 72 billion. Figure 3 and 4 demonstrate effective examples of both languages in the LAVA 2024 contest.

3.3 InternVL2

InternVL2 is one of the most powerful open-source MLLMs designed to handle complex multi-modal tasks involving text, images, and videos. It is the latest version in the InternVL series [9, 10], with various options for the model size, ranging from 1 billion to 108 billion parameters, making it highly versatile and scalable across various applications. It is particularly notable for its multi-modal input support, multitask output capabilities, and progressive alignment training strategy, which align its vision models natively with LLMs.

[5] `https://x.ai/blog/grok-1.5v`

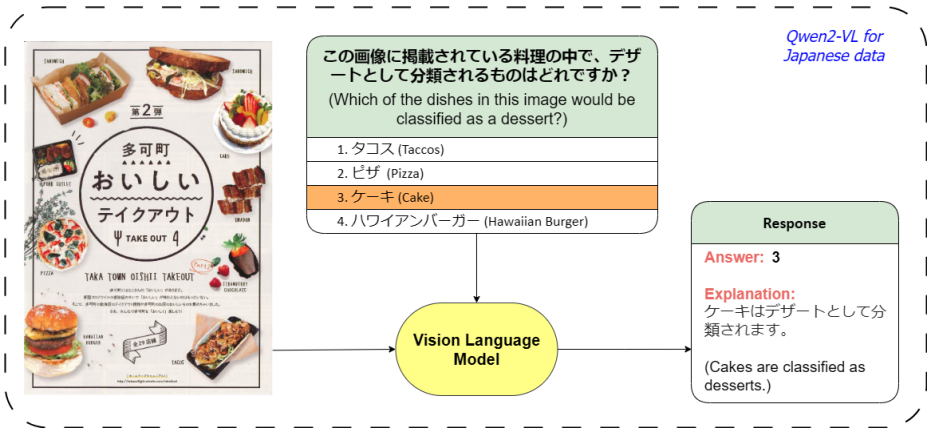

Fig. 4: An example of our best VLM approach for Japanese data. English translations are shown in parentheses.

InternVL2 excels in a wide range of benchmarks, including tasks like visual question answering (VQA), OCR, and grounding. The models achieve performance levels comparable to commercial closed-source models like GPT-4V and Claude 3.5 Sonnet. For instance, InternVL2-8B and InternVL2-Pro have demonstrated superior performance in benchmarks such as MathVista [31], DocVQA [35], and ChartQA [33], showcasing its strengths in visual-linguistic tasks. Since the InternVL2-Pro is the paid commercial version, we decided not to deploy it, as we focus on open-source versions only. We evaluate the InternVL2 series with 4 different model sizes (8B, 26B, 40B and 76B). Since there is a huge gap between the number of parameters, each version will have different vision encoders and language encoders (see Table 1).

3.4 GPT-4o mini

GPT-4o mini[6] is a smaller and more cost-effective version of the larger GPT-4o model developed by OpenAI. It is designed to provide a balance between performance and cost-efficiency, making it suitable for a wide range of applications, especially for businesses and developers looking for powerful AI solutions at a lower price point. The model is created through a distillation process, where it learns to mimic the behavior and performance of the larger GPT-4o model, resulting in a smaller, more affordable version that retains much of the original's capabilities.

3.5 Gemini 1.5 Flash

Gemini 1.5 Flash [37] is a lightweight, fast, and cost-efficient AI model developed by Google DeepMind as part of the Gemini model family. It is optimized for

[6] `https://platform.openai.com/docs/models/gpt-4o-mini`

high-speed and high-volume tasks, making it ideal for applications that require low latency and scalability. Gemini 1.5 Flash is designed to handle multi-modal reasoning, including text, images, and videos, and features a breakthrough long context window of up to one million tokens. This capability allows it to process large amounts of information, such as long documents, extensive codebases, and multimedia inputs.

3.6 Post-processing

As shown in Figure 2, the response is expected to have the JSON format. However, in some cases, the VLM fails to produce the proper output in the first attempt. Since the model's output still contains the selected answer with its explanation, we do not want to regenerate the response, as the answer can differ from the first attempt. Instead, we prompt the VLM to fix the output in order to maintain our pre-designed JSON format. Finally, the final multiple-choice answers and their corresponding descriptions are extracted.

3.7 Ensemble

We also deploy an ensembling method to aggregate answers from multiple model results to produce a unified prediction for each data point. The primary objective of this ensembling approach was to combine outputs from various models and determine the most common prediction (or answer) for each sample across the different model outputs. By leveraging multiple model outputs and ensembling their predictions, we expect to reduce the variance of individual model predictions, thus enhancing the robustness and accuracy of the final result. The ensemble approach is considered as a major voting mechanism that ensures that the most common prediction among the models is selected, providing a reliable consensus output.

4 Experimental Results

All of our VLM approaches are implemented in a zero-shot setting, without any training or fine-tuning. We conducted our evaluations on two A100 80G GPUs and achieved comparable results. Table 1 presents an overview of our team approaches for each VLM separately with the related public scores. The closed-source GPT 4o-mini and Gemini-1.5 Flash share the same performance on the public dataset with a score of 0.71. Regarding the open-source models, the performance on the public leaderboard increases as we use the larger size of VLM. Qwen2-VL, with 76 billion parameters, performs the best, achieving a public score of 0.83. However, this only accounts for 30% of the competition's final score, with the remaining 70% is the score of the private dataset.

Table 2 presents different ensemble configs of our attempt. In the first two settings (without Qwen2-VL), the ensemble approach improves the overall performance. However, when it comes to Qwen2-VL, the ensemble performance

Table 1: Performance comparison of various VLMs based on the LAVA challenge public dataset. Names of vision encoders and text encoders, along with #params and pixel image size, are also displayed. Model with the highest score is highlighted in red.

Model	Open Source	Vision Part	Language Part	Public Score
GPT-4o-mini [1]	✗	-	-	0.71
Gemini-1.5 Flash [37]	✗	-	-	0.71
MiniCPM-V-2.6-8B [50]	✓	SigLip-400M	Qwen2-7B	0.67
InternVL2-8B [10]	✓	InternViT-300M-448px	internlm2_5-7b-chat	0.70
InternVL2-26B [10]	✓	InternViT-6B-448px	internlm2-chat-20b	0.73
InternVL2-40B [10]	✓	InternViT-6B-448px	Nous-Hermes-2-Yi-34B	0.77
InternVL2-Llama3-76B [10]	✓	InternViT-6B-448px	Hermes-2-Theta-Llama-3-70B	0.76
Qwen2-VL-72B [45]	✓	**ViT - 675M - 224px**	**Qwen2-72B**	**0.83**

Table 2: Name of VLM chosen for the ensemble approach with the public scores.

Models	Public Score
InternVL2-(26B + 40B+ 76B)	0.77
InternVL2-(26B + 40B+ 76B) + Gemini-1.5-flash + Gpt-4o-mini	0.79
InternVL2-(26B + 40B+ 76B) + Qwen2-VL-72B	0.8
InternVL2-(26B + 40B+ 76B) + Gemini-1.5-flash + Gpt-4o-mini + Qwen2-VL-72B	0.83

decreases compared to the score of the Qwen2-VL itself. This proves that Qwen2-VL outperforms other VLMs since the majority vote in the ensemble approach may neglect the correct multiple-choice answer from Qwen2-VL.

5 Discussion

5.1 About the proposed approach

Our method in the LAVA 2024 workshop challenge offers specific benefits: (1) **Robustness**: By exploring different large VLMs as well as implementing an ensemble mechanism to obtain the highest vote answers, we provide accurate answers with detailed explanations. However, explicitly verifying every explanation is time-consuming and requires much human force. Instead, we look for several samples in which the correct answers are effortlessly inferred by humans and use them as our validation set. (2) **Adaptability**: As presented earlier, our approach is capable of utilizing multiple pre-trained VLMs [9, 28, 37, 45] without fine-tuning for retraining. This emphasizes that our method is a plug-and-play module and is capable of utilizing various multi-modal models.

Apart from the advantages, our approach still displays some limitations. Firstly, we do not control the language of the generated explanation. As a result, there are such cases where the language of the description differs from the language of the question. Secondly, our approach can be sensitive to the hallucination problem that exists in VLMs. Therefore, the explanation created by the model sometimes includes unnecessary details that are not related to the visual input. Resolving these limitations would necessitate improvement in the structure of the LLMs, which is not the scope of our competition approach.

Fig. 5: Both Japanese and English appear in the content of the images as well as the multiple-choice answers, elevating the level of complexity. English translations are shown in parentheses.

5.2 About the Dataset

The LAVA 2024 workshop challenge is the first competition tackling the understanding of general knowledge for the task of multiple-choice VQA, especially where the Japanese language with a complex writing system is taken into account. The dataset covers a broad spectrum of domains (e.g., healthcare, education, or entertainment) by providing complex visual data (e.g., diagrams, charts, or drawings). One challenge in this competition is the mixed appearance in the visual content as well as the question-answers in a single sample. This happens in both public and private datasets, adding an extra layer of complexity. Figure 5 demonstrates a sample of mixed languages in the dataset, where the question is present in Japanese while the multiple-choice answers and the image content contain both English and Japanese sentences.

Additionally, we recognize an ambiguous case related to the dataset that can affect the performance in general. As shown in Figure 6, there are 100 samples labeled `‘‘ja"` instead of `‘‘ja-JP"` or `‘‘en-US"`. In fact, these 100 samples are all written in the Japanese language. The problem of label inconsistency may hurt some automatic paradigms, where samples in both languages are executed in different pipelines.

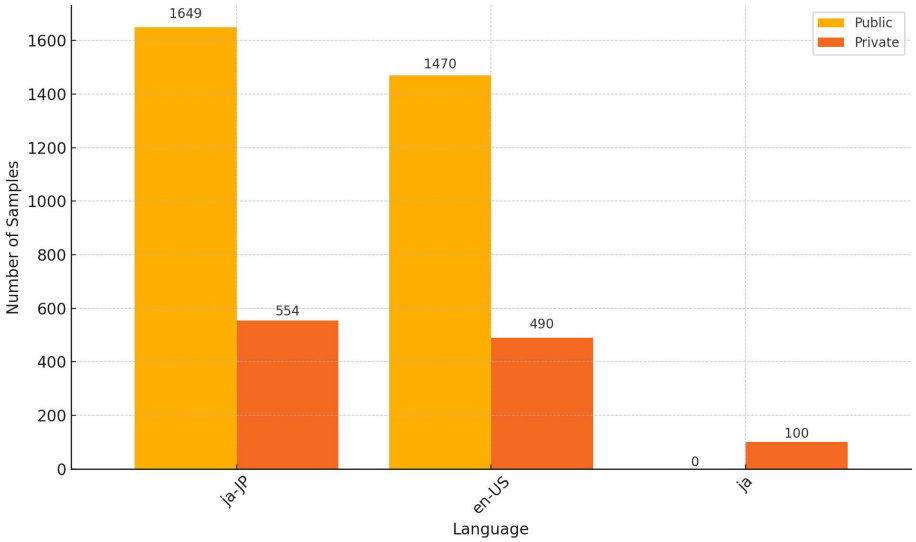

Fig. 6: Overview of the number of samples of each language in LAVA challenge.

6 Conclusion

In this paper, we present our multi-modal vision-language approach for the task of multiple-choice visual question answering. Our approach effectively utilizes large pre-trained VLMs and achieves competitive results in the LAVA 2024 workshop challenge. Through extensive experiments, we demonstrate that our method can handle the cases of Japanese, English, and even the mixture of both languages that appeared in the data samples. The limitations we identified suggest potential directions for future development to enhance the ability of multiple-choice visual question answering.

Acknowledgements: This research is funded by the University of Information Technology - Vietnam National University Ho Chi Minh City under grant number D1-2024-55.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
3. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023)

4. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
5. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
6. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019)
7. Chen, K., Wu, X.: Vtqa: Visual text question answering via entity alignment and cross-media reasoning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27218–27227 (2024)
8. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
9. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024)
10. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
11. Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
12. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. arXiv preprint arXiv:1606.01847 (2016)
13. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: 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. pp. 6904–6913 (2017)
14. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10867–10877 (2023)
15. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
16. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
18. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
19. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elemen-

tary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
20. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. vol. 1, p. 2. Minneapolis, Minnesota (2019)
21. Khan, Z., Fu, Y.: Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10854–10863 (2024)
22. Lan, Y., Li, X., Liu, X., Li, Y., Qin, W., Qian, W.: Improving zero-shot visual question answering via large language models with reasoning question prompts. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 4389–4400 (2023)
23. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
24. Li, L., Peng, J., Chen, H., Gao, C., Yang, X.: How to configure good in-context sequence for visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 26710–26720 (2024)
25. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., et al.: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16. pp. 121–137. Springer (2020)
26. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1650–1654. IEEE (2021)
27. Liu, C., Wang, C., Peng, Y., Li, Z.: Zvqaf: Zero-shot visual question answering with feedback from large language models. Neurocomputing **580**, 127505 (2024)
28. Liu, Y., Liang, Z., Wang, Y., He, M., Li, J., Zhao, B.: Seeing clearly, answering incorrectly: A multimodal robustness benchmark for evaluating mllms on leading questions. arXiv preprint arXiv:2406.10638 (2024)
29. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems **32** (2019)
30. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems **29** (2016)
31. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
32. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: Proceedings of the IEEE international conference on computer vision. pp. 1–9 (2015)
33. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv preprint arXiv:2203.10244 (2022)
34. Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1697–1706 (2022)

35. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2200–2209 (2021)
36. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
37. Reid, M., Savinov, N., Teplyashin, D., Lepikhin, D., Lillicrap, T., Alayrac, J.b., Soricut, R., Lazaridou, A., Firat, O., Schrittwieser, J., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024)
38. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. Advances in neural information processing systems **28** (2015)
39. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
40. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
41. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
42. Tito, R., Karatzas, D., Valveny, E.: Document collection visual question answering. In: Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16. pp. 778–792. Springer (2021)
43. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
44. Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)
45. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
46. Wang, P., Wu, Q., Shen, C., Dick, A., Van Den Hengel, A.: Fvqa: Fact-based visual question answering. IEEE transactions on pattern analysis and machine intelligence **40**(10), 2413–2427 (2017)
47. Wang, Z., Yu, J., Yu, A.W., Dai, Z., Tsvetkov, Y., Cao, Y.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904 (2021)
48. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
49. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)
50. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., Zhu, H., Cai, T., Li, H., Zhao, W., He, Z., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
51. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)

52. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)
53. Yuan, L., Chen, D., Chen, Y.L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., et al.: Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432 (2021)
54. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11975–11986 (2023)
55. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C.: Raven: A dataset for relational and analogical visual reasoning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5317–5327 (2019)