Enhancing Visual Question Answering with Pre-trained Vision-Language Models: An Ensemble Approach at the LAVA Challenge 2024

Trong-Hieu Nguyen-Mau[1,2], Nhu-Binh Nguyen Truc[1,2], Nhu-Vinh Hoang[1,2], Minh-Triet Tran[1,2], and Hai-Dang Nguyen[1,2]

[1] University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[2] Viet Nam National University, Ho Chi Minh City, Vietnam
nmthieu@selab.hcmus.edu.vn; ntnbinh21@apcs.fitus.edu.vn; hnvinh21@apcs.fitus.edu.vn; tmtriet@fit.hcmus.edu.vn; nhdang@selab.hcmus.edu.vn

Abstract. The LAVA challenge presents complex visual question answering tasks involving intricate diagrams, each accompanied by multiple-choice questions in English or Japanese. Addressing this challenge, we - the team v1olet - explore the capabilities of pre-trained Large Vision-Language Models to interpret and reason over such sophisticated visual data. We utilize models including Qwen2-VL, InternVL2, MiniCPM, and Llama-3.2-Vision-Instruct, employing a structured prompt template designed to standardize response generation and facilitate step-by-step reasoning. To enhance accuracy and robustness, we implement an ensemble method using majority voting to combine outputs from different models and configurations. Our experimental results demonstrate that the ensemble approach significantly improves performance, achieving a higher public score on the LAVA challenge dataset compared to individual models. Specifically, the ensemble of Qwen2-VL, InternVL2, and Llama-3.2 models attained the highest public score of 82, outperforming the best single model. This study highlights the effectiveness of combining multiple Large Vision-Language Models through ensemble methods and underscores the potential of prompt-based inference in enhancing model reasoning capabilities for complex VQA tasks. The provided code is here.

Keywords: Large Vision-Language Models · Visual Question Answering · Ensemble Methods

1 Introduction

Large Vision-Language Models (LVLMs) have emerged as a transformative force in artificial intelligence by seamlessly integrating visual and textual data at scale [12]. These generative models are engineered to process and understand multiple modalities simultaneously, enabling them to interpret visual content and generate coherent textual responses or outputs [17]. This multimodal capability is particularly valuable in tasks that require the fusion of visual and

textual information, making LVLMs critical tools in addressing complex challenges across various real-world applications [3, 21].

The advent of LVLMs has significantly advanced fields such as image captioning [15], visual question answering (VQA) [4], and multimodal dialogue systems [25]. In VQA, for instance, models are required not only to recognize objects within an image but also to comprehend contextual cues and infer relationships between entities to answer questions accurately [27]. This necessitates a deep understanding of both visual content and natural language, highlighting the importance of sophisticated LVLMs in achieving high performance in such tasks.

Despite these advancements, current LVLMs face significant challenges when dealing with complex visual data that go beyond straightforward photographic images. Real-world applications often involve intricate diagrams, technical schematics, and multilingual texts, which are common in fields like engineering, architecture, and data analysis. These complex visual representations require models to perform advanced reasoning and interpretation, pushing the boundaries of current multimodal understanding capabilities.

Addressing this gap, the LAVA Challenge 2024 introduces a dataset specifically designed to evaluate and enhance the capabilities of LVLMs in interpreting complex visual information [2]. The dataset consists of two parts: a public dataset with approximately 3,000 samples sourced from the internet, and a private dataset provided by the TASUKI team (SoftBank), containing around 1,100 samples. The visual data encompass a wide range of complex diagrams such as Data Flow Diagrams, Class Diagrams, Gantt Charts, and Building Design Drawings. Each image is accompanied by a multiple-choice question in either English or Japanese, with four possible answers derived from the visual content.

The LAVA Challenge presents several unique difficulties:

- **Complex Visual Structures:** The images contain detailed and abstract representations that require models to understand not just objects but also relationships, hierarchies, and flows of information.
- **Multilingual Text Understanding:** Questions are provided in both English and Japanese, necessitating models to possess or integrate cross-lingual comprehension capabilities.
- **Advanced Reasoning:** Answering the questions correctly often involves multi-step reasoning processes, including deduction, inference, and sometimes even external knowledge.

These challenges make the LAVA dataset a rigorous benchmark for testing the limits of current LVLMs and exploring new methodologies to enhance their performance.

In this paper, we introduce an ensemble approach leveraging pre-trained LVLMs to enhance visual question answering performance on the LAVA Challenge 2024. Our framework integrates multiple state-of-the-art LVLMs, including Qwen2-VL, InternVL2, MiniCPM, and Llama-3.2-Vision-Instruct, utilizing a structured prompt design to standardize responses and facilitate step-by-step reasoning. Our contributions are multifaceted and are presented as follows:

- **Structured Prompt Design for Enhanced Reasoning:** We design a structured prompt template that standardizes response generation across different models. This template guides the models through a step-by-step reasoning process, improving their ability to interpret intricate visual data and generate consistent answers.
- **Majority Voting Ensemble Method:** We implement a majority voting scheme to aggregate outputs from various models and configurations. This method leverages the strengths of individual models while mitigating their weaknesses, resulting in improved accuracy and robustness in the final predictions.
- **Significant Performance Improvement on LAVA Challenge Dataset:** Our experimental results demonstrate that the ensemble approach significantly enhances performance, achieving higher scores on the LAVA Challenge 2024 dataset compared to individual models. This highlights the effectiveness of our method in advancing the state-of-the-art in visual question answering.

This paper is structured as follows. In Section 2, we provide a concise overview of existing methods relevant to our research. Section 3 introduces our proposed approach in detail. We then discuss our experimental findings in Section 4. In Section 5, we discuss and outline open problems and future investigations. Finally, Section 6 concludes the paper.

2 Related Works

2.1 Large vision-language models

Large vision-language models (LVLMs) represent the intersection of machine vision and natural language processing, combining visual interpretation with linguistic capabilities to answer questions, perform visual reasoning, and generate textual descriptions [30]. These models leverage advancements in both machine vision—traditionally used for tasks like image classification, object detection, and counting—and the robust inference capabilities of large language models (LLMs) [8]. With access to vast amounts of data from the internet, pre-trained LVLMs are particularly adept at domain-specific adaptation [31], even when faced with previously unseen images. This flexibility is largely attributed to contrastive learning techniques [6], which enhance the model's ability to align visual and textual modalities.

In recent advancements, the alignment of modalities allows models to answer questions about images, generate captions, and even engage in multimodal dialogues, as seen in models like InternVl [23], and Qwen-VL [9]. These models connect vision encoders with LLMs using advanced architectures like vision transformer (ViT) [29], multilayer perceptrons (MLP) [23].

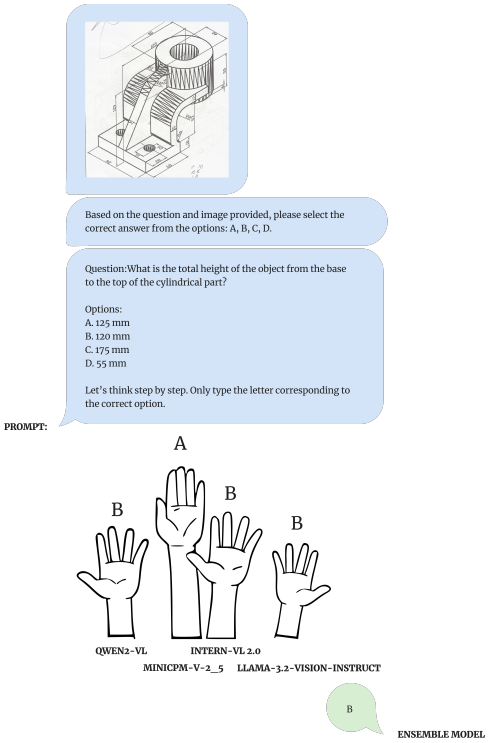

Fig. 1: Overall process of our approaches.

2.2 Visual Question Answering

The task of Visual Question Answering (VQA) involves answering natural language questions based on a given image. It represents the intersection of computer vision and natural language processing. VQA requires models to perform multimodal reasoning and integrate visual with text understanding.

Early approaches used simple CNN-LSTM architectures [5] to extract visual features from images and encode questions using recurrent neural networks. Later, Stacked Attention Networks (SAN) [28] enabled models to focus on specific regions of the image based on the questions. Furthermore, multimodal fusion techniques such as Multimodal Compact Bilinear Pooling (MCB) [10] and Bilinear Attention Networks (BAN) [13] improved the interaction between visual and textual features, leading to enhanced performance. In 2019, the advent of transformer-based models like VilBERT [18] and LXMERT [22] extended the success of transformers to vision-language tasks.

2.3 Ensemble approaches for Visual Question Answering

Ensemble methods in VQA leverage the strengths of multiple models to enhance overall performance, reducing biases and increasing the robustness of predictions. One such method, the Greedy Gradient Ensemble [11], employs a strategy to sequentially fit bias models to data, thus enabling a base model to focus more effectively on unbiased data distribution. This method helps improve model generalization by minimizing biases inherent in the training data, which often skew predictions.

Additionally, the use of explicit attention models has been shown to improve VQA accuracy by allowing models to focus on the most relevant parts of an image [16]. This approach leverages separate word embedding models for textual and visual inputs, enhancing the expressive power of the attention mechanism and improving the accuracy of answering visually grounded questions.

Recently, some papers have utilized simple aggregation methods yet achieved impressive results [14]. These methods, which simply select the most frequent output from various models, have surprisingly performed well and demonstrated good generalization.

These ensemble and attention methods illustrate advanced strategies in handling the complex interplay of visual and textual information essential for robust VQA systems. They address the critical challenges of bias and focus on model responses, contributing to more accurate and reliable VQA outcomes.

3 Proposed Method

We conduct experiments with pre-trained LVLMs through prompt-based inference, utilizing the following models: Qwen2-VL [24], InternVL2 [23], MiniCPM [29], and Llama-3.2-Vision-Instruct [20]. We chose these models because they are the top-performing models on the leaderboard [1] that we could access and apply to our experiments.

We design a structured prompt template to standardize the process of generating responses, guiding them through a structured, step-by-step reasoning framework. The overall approach is shown in Figure 1.

We obtain answers to each model's input questions using the prompt template. To enhance accuracy and robustness, we employ an ensemble approach to combine the outputs from different models, known as majority voting, where the most commonly selected answer by the models is chosen as the final prediction. This ensemble method increases accuracy by leveraging the collective wisdom of several models, mitigating the risks associated with individual model errors and capitalizing on their diverse strengths for more reliable and robust results.

Specifically, we generate results from multiple inference runs with different configurations, including variations in step-by-step reasoning, or different model sizes (e.g., InternVL 26B, 40B, 76B [23] or Llama-3.2-Vision-Instruct with 11B or 90B versions [20]). The outputs from these models are then aggregated by selecting the most frequent response for each question, based on a majority

voting scheme. We gather the answers from all models for each image-question pair and select the most commonly occurring answer as the final prediction.

We will detail how we designed the prompt and the specifics of each model family in subsequent sections.

3.1 Prompt Design

In our experiments, we utilized a structured prompt template to standardize the input provided to the models and guide them toward generating accurate answers. The prompt facilitates step-by-step reasoning and ensures that the models focus on selecting the correct option from the given choices. The prompt template is as follows:

This prompt begins by instructing the model to select the correct answer based on the provided question and image, explicitly mentioning that the options are labeled A, B, C, and D. This sets clear expectations for the model's response format.

The placeholders `question` and `option1` to `option4` are filled with the actual question and options from the dataset. Presenting the options in a labeled list helps the model to reference them easily during reasoning.

The instruction "Let's think step by step." encourages the model to engage in a chain-of-thought reasoning process [26], which has been shown to improve performance on complex tasks. By prompting the model to consider the problem methodically, we aim to enhance its ability to arrive at the correct answer.

Finally, we include the directive "Only type the letter corresponding to the correct option." to constrain the model's output to the required format. This minimizes the chance of generating extraneous text and ensures that the answer can be easily parsed and evaluated.

3.2 Qwen2-VL

Qwen2-VL [24] series represents an advanced upgrade from the original Qwen Large Vision Language Model (Qwen-VL) which is proposed by Alibaba Cloud. It is a powerful tool for a variety of applications, including visual analysis, multilingual support, and autonomous agent capabilities [9].

One of the key innovations in Qwen2-VL is its Naive Dynamic Resolution mechanism [24], allowing the model to dynamically adjust the resolution of input images and closely mirroring the way humans perceive visual information. Additionally, Qwen2-VL incorporates Multimodal Rotary Position Embedding (M-RoPE), enabling effective fusion of positional data across text, images, and videos, hence improving overall comprehension of complex, multimodal inputs.

The Qwen2-VL series also scales across different parameter sizes—2 billion, 7 billion, and 72 billion parameters—making it adaptable to a range of tasks, from efficient on-device performance to tackling more complex visual reasoning tasks. The largest model, Qwen2-VL-72B, achieves results comparable to state-of-the-art systems like GPT-4 and Claude 3.5 [24].

3.3 InternVL2

InternVL represents a significant leap in vision-language foundation models by scaling up the vision foundation model to 6 billion parameters [7]. The architecture of InternVL2, consistent with the ViT-MLP-LLM configuration of previous version InternVL 1.5, has been further enhanced in version 2.0 with a variety of instruction-tuned models, ranging from 1 billion to 108 billion parameters [23].

Through these advanced techniques, InternVL2 establishes itself as a competitive alternative to other large-scale vision models, delivering robust performance in chart comprehension and infographics QA. Its ability to seamlessly integrate with LLMs positions it as an essential tool for developing comprehensive multimodal AI applications.

3.4 MiniCPM

MiniCPM-Llama3-V 2.5 [29], the latest in the MiniCPM-V series, builds on SigLip-400M and Llama3-8B-Instruct, offering significant performance enhancements over its predecessor, MiniCPM-V 2.0 [29]. This model excels in chart comprehension and infographics QA, achieving an impressive 65.1 average score on OpenCompass across 11 benchmarks, and outperforms leading models like GPT-4V-1106 and Gemini Pro with its 8 billion parameters. Additionally, its advanced OCR capabilities allow for adept processing of images up to 1.8 million pixels in any aspect ratio, excelling in data extraction from charts and infographics with a score over 700 on OCRBench [29]. Enhanced instruction-following and reasoning abilities improve multimodal interactions and utility in applications requiring nuanced visual content understanding.

3.5 Llama-3.2-Vision-Instruct

Llama-3.2-Vision-Instruct [20] includes multimodal large language models for both image and text processing in 11B and 90B sizes, as well as lightweight, text-only models in 1B and 3B sizes that support a context length of 128K tokens and are state-of-the-art in their class for on-device use cases. These models are pre-trained and fine-tuned for visual recognition, image reasoning, captioning, and answering general questions about images.

The model's evaluation shows that the Llama-3.2-Vision model performs similarly to leading models like Claude 3 Haiku and GPT4o-mini in tasks of image recognition and visual understanding. This 3B model performs better than the Gemma 2 2.6B and Phi 3.5-mini models in tasks involving following instructions, summarization, and prompt rewriting, while the 1B model is comparable to Gemma [20].

3.6 Post-processing Approach

In our experiments, we observed that individual models sometimes produced inconsistent or incorrect answers due to the complexity of the visual questions

and the diversity of the data. To address this issue and enhance the overall accuracy of our system, we implemented a post-processing step that aggregates the outputs from multiple models using a majority voting scheme.

Our post-processing approach operates as follows:

1. For each image-question pair in the dataset, we obtain the predicted answers from multiple models (e.g., Qwen2-VL-7B-Instruct, InternVL2 models, Llama-3.2-Vision-Instruct models).
2. We collect these predictions into a list of candidate answers.
3. We apply majority voting to determine the final answer, selecting the option that appears most frequently among the models' predictions.
4. In cases where there is a tie (i.e., multiple options receive the same highest number of votes), we apply a predefined tie-breaking strategy by selecting the answer from the model with the highest individual performance, which is the Qwen2-VL model.

This ensemble method utilizes the strengths of individual models while mitigating their weaknesses. By aggregating the predictions, it diminishes the effect of any single model's errors and enhances the robustness of the overall result. This method is straightforward, efficient, and scalable, enabling us to process the entire dataset with minimal computational overhead.

4 Experiments

4.1 Datasets and Evaluation Metrics

The datasets employed in our study are integral components of the LAVA challenge [2], designed to evaluate the effectiveness of LVLMs in understanding and interpreting complex visual data accompanied by textual queries. The public dataset comprises around 3,000 visual samples, each paired with a multiple-choice question in either English or Japanese. These samples predominantly include intricate diagrams such as Data Flow Diagrams, Gantt Charts, and Building Designs, which demand a high level of visual-textual comprehension. On the other hand, the private dataset contains approximately 1,100 samples provided by the TASUKI team at SoftBank, featuring a similar composition and complexity but used exclusively for private leaderboard evaluation.

The evaluation of this competition will be based on the MMMU metric. The final score will be calculated as follows: Final score $= 0.3 \times$ Public dataset score$+$ $0.7 \times$ Private dataset score. We can only access the Public dataset score.

4.2 Implementation Details

We utilize PyTorch and the `transformers` library from Hugging Face for our experiments. Our study explores the range of image dimensions within our dataset, noting a maximum height of 7041 pixels and a maximum width of 9600 pixels, with minimums at 60 pixels in height and 130 pixels in width. We retain

Table 1: Comparison results on Public dataset.

Index	Method	Public Score
(1)	Qwen2-VL-7B-Instruct	**80**
(2)	MiniCPM-Llama3-V-2_5	64
(3)	InternVL2-26B	75
(4)	InternVL2-40B	73
(5)	InternVL2-76B	79
(6)	Llama-3.2-11B-Vision-Instruct	70
(7)	Llama-3.2-90B-Vision-Instruct	77
(8)	(3) + (4) + (5)	78
(9)	(1) + (3) + (4) + (5)	80
(10)	(1) + (3) + (4) + (5) + (6) + (7)	**82**

the original image sizes, which are processed according to each model's specific requirements, as detailed in our code. The inference pipeline is constructed using the prompt template described in 3.1. To minimize randomness in our experiments, we try to set `do_sample=False`, and if not, we try to adjust the `temperature=0.000001` and `top_k=1`. These experiments were conducted on a single NVIDIA H100 80GB graphics card, with a batch size set to 1.

4.3 Results

Table 1 presents the performance of various LVLMs and their ensembles on the public dataset, evaluated using the Public Score metric. Among the individual models, Qwen2-VL-7B-Instruct (Index 1) achieves the highest score of 80, outperforming larger models like InternVL2-76B (Index 5) and Llama-3.2-90B-Vision-Instruct (Index 7), which score 79 and 77 respectively. This indicates that model performance is not solely dependent on the number of parameters; factors such as model architecture and training data play significant roles. Notably, MiniCPM-Llama3-V-2_5 (Index 2) scores the lowest at 64, suggesting limitations in handling the complexity of the dataset.

Ensemble methods show improved performance over individual models. The ensemble of all InternVL2 models (Index 8) increases the score to 78, while adding Qwen2-VL-7B-Instruct to the ensemble (Index 9) maintains the score at 80. The highest score of 82 is achieved by the comprehensive ensemble (Index 10) combining Qwen2-VL, InternVL2 models, and Llama-3.2 models, excluding the lowest-performing MiniCPM-Llama3-V-2_5. This demonstrates that combining models with diverse strengths through ensemble methods like majority voting can enhance predictive accuracy, mitigating individual model errors and leveraging complementary capabilities.

5 Discussion

Although our model performed well, there is still room for improvement. The challenge dataset does not provide the correct answers, so we cannot confirm if the model's predictions are accurate. We noticed that the model's performance varies across different versions and configurations. This observation led us to use an ensemble approach, which helps to stabilize the model's performance by combining predictions from multiple models. This method aims to reduce errors and increase the likelihood of selecting the most accurate answer, especially in difficult or ambiguous cases.

In future work, we plan to develop more advanced ensemble methods beyond simple majority voting. We will explore options like decision trees or assigning specific models to particular tasks or languages, which could improve the model's accuracy. Another promising method is the Mixture of Experts (MoE) [19], which divides a model into specialized parts that each handle a subset of the data. We believe that incorporating MoE could greatly enhance our model's effectiveness. Additionally, we intend to fine-tune our models specifically for the tasks in the LAVA challenge if we could find the dataset with the actual label. This should lead to better performance, particularly for processing both English and Japanese, as current models may not handle different languages equally well.

6 Conclusion

In summary, our study demonstrates the effectiveness of pre-trained Large Vision-Language Models (LVLMs) in tackling complex visual question answering tasks using the LAVA challenge dataset. By employing a structured prompt template that encourages step-by-step reasoning, we enhanced the models' ability to interpret intricate visual data. Among individual models, Qwen2-VL-7B-Instruct achieved the highest public score, highlighting its capability despite a smaller parameter size. Moreover, ensemble methods using majority voting further improved performance, with the comprehensive ensemble attaining a public score of 82. These results underscore the benefits of combining different LVLMs to leverage their diverse strengths and improve predictive accuracy. Our work highlights the potential of ensemble approaches and prompt-based inference in advancing multimodal AI applications, particularly for tasks requiring sophisticated visual and textual understanding.

References

1. Openvlm leaderboard, https://huggingface.co/spaces/opencompass/open_vlm_leaderboard 5
2. ACCV workshop on large vision – language model learning and applications (2024), https://lava-workshop.github.io/ 2, 8
3. Akshay Gopalkrishnan, Ross Greer, M.T.: Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving (2024), https://arxiv.org/abs/2403.19838 2

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) 2
5. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. CoRR **abs/1505.00468** (2015), http://arxiv.org/abs/1505.00468 4
6. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330 (2024) 3
7. Chen, Z., Wang, W., Tian, e.a.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821 (2024) 7
8. Cheng, S., Guo, Z., Wu, J., Fang, K., Li, P., Liu, H., Liu, Y.: Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14291–14302 (2024) 3
9. Emanuilov, S.: Qwen2-vl — a new milestone in vision-language ai, https://unfoldai.com/qwen2-vl/ 3, 6
10. Fukui, A., Park, D.H., Yang, D., Rohrbach, A., Darrell, T., Rohrbach, M.: Multimodal compact bilinear pooling for visual question answering and visual grounding. CoRR **abs/1606.01847** (2016), http://arxiv.org/abs/1606.01847 4
11. Han, X., Wang, S., Su, C., Huang, Q., Tian, Q.: Greedy gradient ensemble for robust visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1584–1593 (2021) 5
12. Jiang, Y., Yan, X., Ji, G.P., Fu, K., Sun, M., Xiong, H., Fan, D.P., Khan, F.S.: Effectiveness assessment of recent large vision-language models. Visual Intelligence **2**(1), 17 (2024) 1
13. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. CoRR **abs/1805.07932** (2018), http://arxiv.org/abs/1805.07932 4
14. Le, B.H., Nguyen-Mau, T.H., Nguyen-Vu, D.K., Ho-Ngoc, V.P., Nguyen, H.D., Tran, M.T.: Leveraging large vision-language models for visual question answering in vizwiz grand challenge 5
15. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International conference on machine learning. pp. 12888–12900. PMLR (2022) 2
16. Lioutas, V., Passalis, N., Tefas, A.: Explicit ensemble attention learning for improving visual question answering. Pattern Recognition Letters **111**, 51–57 (2018) 5
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 1
18. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. CoRR **abs/1908.02265** (2019), http://arxiv.org/abs/1908.02265 4
19. Masoudnia, S., Ebrahimpour, R.: Mixture of experts: a literature survey. Artificial Intelligence Review **42**, 275–293 (2014) 10
20. Meta: Llama 3.2: Revolutionizing edge ai and vision with open, customizable models, https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ 5, 7
21. Saxena, S., Sharma, M., Kroemer, O.: Mrest: Multi-resolution sensing for real-time control with vision-language models (2024), https://arxiv.org/abs/2401.14502 2

22. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. CoRR **abs/1908.07490** (2019), http://arxiv.org/abs/1908.07490 4
23. Team, O.: Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy (2024), https://internvl.github.io/blog/2024-07-02-InternVL-2.0/ 3, 5, 7
24. 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) 5, 6
25. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024) 2
26. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022) 6
27. Yang, A., Miech, A., Sivic, J., Laptev, I., Schmid, C.: Just ask: Learning to answer questions from millions of narrated videos. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1686–1697 (2021) 2
28. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. CoRR **abs/1511.02274** (2015), http://arxiv.org/abs/1511.02274 4
29. Yao, Y., Yu, T., Zhang, e.a.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint 2408.01800 (2024) 3, 5, 7
30. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024) 3
31. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., Gao, J.: Unified vision-language pre-training for image captioning and vqa. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 13041–13049 (2020) 3