An Approach to Complex Visual Data Interpretation with Vision-Language Models

Thanh-Son Nguyen[1,2*], Viet-Tham Huynh[1,2*], Van-Loc Nguyen[1,2*], and Minh-Triet Tran[1,2**]

[1] Software Engineering Laboratory, University of Science, VNU-HCM, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
{nthanhson,hvtham,nvloc}@selab.hcmus.edu.vn,
tmtriet@fit.hcmus.edu.vn

Abstract. The LAVA Workshop 2024 challenge aimed to assess the capability of Large Vision-Language Models (VLMs) to interpret and understand complex visual data accurately. This includes intricate visual formats such as data flow diagrams, class diagrams, Gantt charts, and architectural blueprints. In response to this challenge, our research focuses on adapting the MMMU (Massive Multi-discipline Multimodal Understanding) benchmarks to better align with the requirements of visual data interpretation. We propose a comprehensive approach that leverages advanced prompt engineering techniques and incorporates a voting-based ensemble method for aggregating model predictions. This method improves the model's ability to generalize across different types of visual inputs. Our approach was rigorously evaluated within the context of the challenge, resulting in a total score of 0.85, ultimately securing the top position in the competition. This result demonstrates the effectiveness of combining prompt engineering with simple yet powerful ensemble strategies for enhancing the performance of VLMs on complex multimodal tasks.

Keywords: Visual language · MMMU · AGI

1 Introduction

Advancements in AI and machine learning, especially large language models (LLMs) like OpenAI's o1 or Meta's LLaMA, demonstrate great utilization in various applications in society. These language models have shown an exceptional ability to utilize large-scale text data to deliver outstanding results in various natural language understanding tasks. As a result, the research community has paid a lot of attention to expanding the capabilities of these language models to additional data modalities, such as images, videos, or audio, forming large vision-language models (LVLMs) like GPT-V or LLaVA.

[*] Equal contributions
[**] Corresponding author

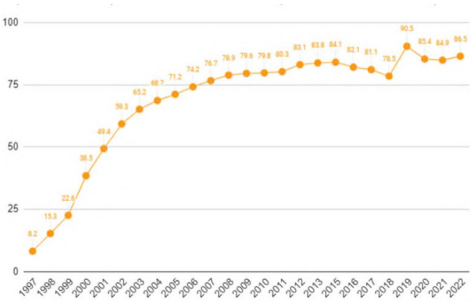


Fig. 1: Examples of the inputs and outputs of LAVA workshop challenge

Large Vision-Language Model Learning and Applications Workshop (LAVA Workshop) is held to encourage researchers to unlock the full potential of research in large vision-language models by focusing on the integration of various modalities, such as text, images, and videos. In addition, the workshop serves as a forum to explore the usage of large vision-language models across various fields, such as healthcare, education, entertainment, transportation, finance, and more.

The LAVA workshop also contains the challenge, with the main objective to enhance the ability of large vision-language models to precisely interpret and comprehend complex visual data, including data flow diagrams (DFDs), class diagrams, Gantt charts, or building design drawings. Figure 1 illustrates examples of the inputs and outputs of the LAVA workshop challenge.

In this research, we propose investigating the capabilities of different large vision-language models in answering questions with complex visual data in various fields and the effect of changing prompts on the model's performance.

The next sections of this paper will be organized as follows. Section 2 briefly describes the related research to our work. Section 3 is our proposed method, including the information about the dataset, MMMU benchmark, model selection, the architectures of chosen models, and our prompt techniques. Section 4 is the experiments and results of our method, including our prompt and results on public and private datasets. Section 5 is our conclusion and discussion about future works.

2 Related work

2.1 Multimodal Pre-Training

Significant progress has been achieved recently in multimodal pre-training, which focuses on integrating vision and language into a unified model. Early works, including VinVL [45], Oscar [21], LXMERT [34], VilBERT [28], and UNITER [5], pioneered the development of universal models for vision-language tasks. These models often relied on pre-trained visual features like Faster RCNN to reduce training complexity. More recent approaches such as ALIGN [16], CoCa [41], CLIP [33], Fuyu [44] SimVLM [37], Flamingo [3], and BLIP-2 [19] have shifted toward training visual representations from scratch using Vision Transformers (ViT) [12] and vast datasets sourced from the web. These models have performed well in visual question answering (VQA) and image captioning tasks, which require less in-depth reasoning.

2.2 Multimodal Instruction Tuning

Building on the success of instruction-tuned large language models (LLMs) like Vicuna [8] and FLAN-T5 [9], new models such as MiniGPT-4 [11] and LLaVA [47] have been developed to enhance the instruction-following abilities of large multimodal models (LMMs). This has spurred advances in generating high-quality visual instruction data, with models like mPlug-OWL [39], LLaMA-Adapterv2 [13], LRV-Instruction [24], SVIT [26], and InstructBLIP [10] leading the charge. Another critical aspect of LMM research focuses on multimodal in-context learning, where models handle mixed examples of text and images. Notable models in this domain include M3IT [20], Sparkles [14], MetaVL [32], Otter [17], Flamingo [2], OpenFlamingo [4], and MMICL [46], which have contributed to improving multimodal training and instruction-following abilities.

2.3 LMM Benchmarks

The rapid advancement in multimodal pre-training and instruction tuning has outgrown traditional single-task benchmarks like MSCOCO [23], VQA [1], OK-VQA [30], and GQA [15]. These benchmarks are now insufficient for evaluating LMMs' broader abilities in perception and reasoning. Consequently, several comprehensive benchmarks have emerged, focusing on a range of LMM capabilities. These include optical character recognition (OCR) as explored in various studies, adversarial robustness, and hallucination issues, with benchmarks such as HaELM [35] and POPE [22] specifically addressing these challenges. Holistic evaluations are provided by LVLM-eHub [38], SEED [18], [27], LAMM [40], and MM-Vet [42]. However, most benchmarks still emphasize basic perception tasks, which do not require deep domain knowledge or advanced reasoning. A recent benchmark, MathVista [29], evaluates models on visually complex questions in the mathematical domain. In contrast, MMMU [43] introduces more

complex, expert-level tasks across 30 disciplines, demanding advanced perception and domain-specific knowledge for step-by-step reasoning. Concurrently, GAIA [31] presents 466 questions testing models' abilities in reasoning, multimodal comprehension, and tool usage.

3 Proposed method

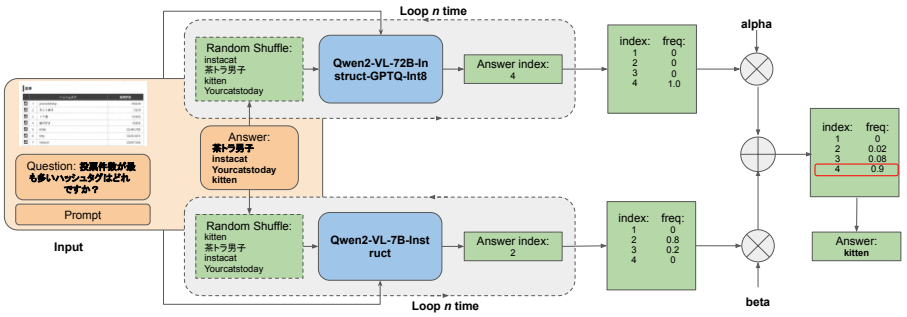


Fig. 2: Proposed method overview: our proposed method contains three steps. First, we use prompt engineering to produce input for pre-trained models. Second, the prompt is fed into the network to generate answers. Thirdly, we assemble all answers from multiple networks to decide the final result.

3.1 Overview of the Dataset

The dataset contains 2 subsets: public dataset and private dataset. The public dataset contains about 3000 data samples released by the organizing committee of the LAVA workshop. This subset is collected on the Internet. The private dataset, provided by the TASUKI team (SoftBank), consists of around 1100 samples.

Both subsets share the same structure: an image, a question, the four options to be chosen, and an indicator specifying the language of the question and options. There are 2 languages used in this dataset: English (*en-US*) and Japanese (*ja-JP*).

3.2 MMMU Benchmark

The Massive Multi-discipline Multimodal Understanding and Reasoning (MMMU) [43] benchmark, developed by Yue et al., is a novel benchmark designed to evaluate multimodal models on massive multi-discipline tasks requiring subject knowledge at a college level, and thoughtful reasoning. This benchmark is developed to measure 3 essential skills for multimodal models: perception, knowledge, and reasoning. There are 6 core disciplines covered by this dataset: Art & Design,

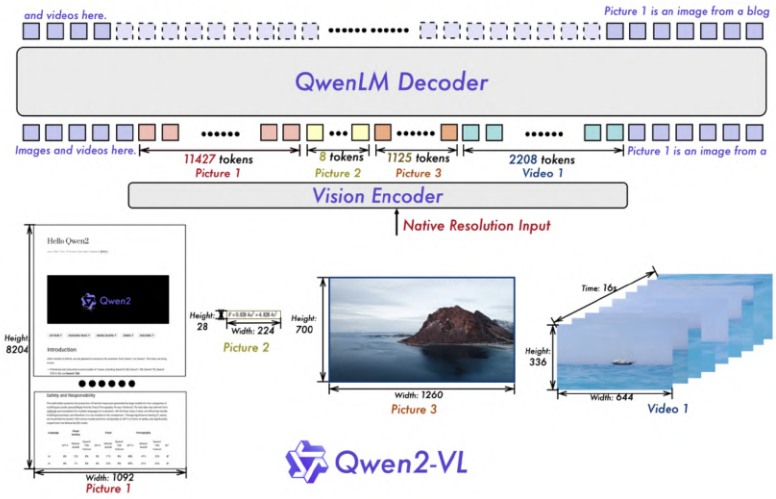


Fig. 3: Architecture of Qwen2-VL [36]

Business, Science, Health & Medicine, Humanities & Social Science, and Technology & Engineering.

The results of submissions to the LAVA workshop challenge are evaluated by the MMMU metric, suggesting a relationship between the challenge's input-output structure and the MMMU benchmark. Therefore, we conducted our experiments on models with the highest MMMU benchmark results.

3.3 Model Selection

In the leaderboard of the MMMU benchmark, the model o1 from OpenAI achieves the highest result, which is even better than the Human Expert (Low) performance. The following positions contain several large language models and large vision-language models, such as GPT-4o also from OpenAI, Claude 3.5 Sonnet, Gemini 1.5 Pro, Qwen2-VL, InternVL2, LLaVA, etc.

In this paper, Qwen2-VL, InternVL2, and LLaVA are the main focus, especially Qwen2-VL [36][3] and InternVL2 [6, 7][4], because these models are in top-ranking performance of the MMMU benchmark, as well as they are open-source and contain inference APIs on HuggingFace.

3.4 Model Architectures

Architecture of Qwen2-VL The Qwen2-VL architecture builds upon the Qwen-VL framework, combining a Vision Transformer [12] (ViT) model with Qwen2 language models. Below are key elements of this architecture include:

[3] https://huggingface.co/Qwen/Qwen2-VL-2B-Instruct-GPTQ-Int4
[4] https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B

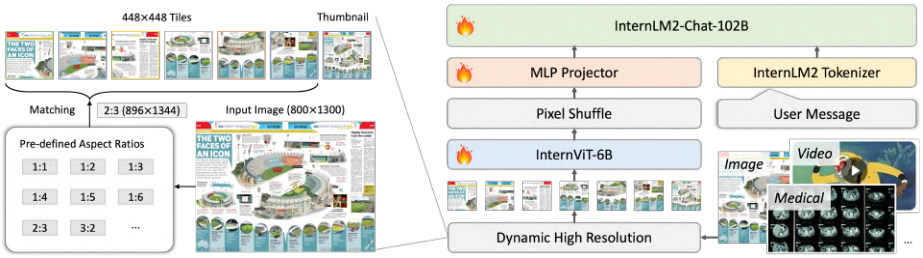


Fig. 4: Architecture of InternVL1.5 [6]

– ***Vision Transformer (ViT) with 600M Parameters:*** The model utilizes a ViT with approximately 600 million parameters to process both image and video inputs seamlessly.
– ***Naive Dynamic Resolution Support:***
  • Qwen2-VL introduces Naive Dynamic Resolution, allowing the model to handle images of arbitrary resolutions.
  • This maps images into a variable number of visual tokens, ensuring that the input is consistent with the image's inherent information.
  • This mechanism mimics human visual perception, enabling the model to process images of different clarity and sizes.
– ***Multimodal Rotary Position Embedding (M-ROPE):***
  • M-ROPE is an innovation that deconstructs the original rotary embedding into components representing temporal and spatial (height and width) dimensions.
  • This enables the model to capture and integrate positional information across 1D text, 2D images, and 3D videos, allowing simultaneous comprehension of various data types.

Architecture of InternVL2 Up until now, the architecture of InternVL2 has not been published yet. InternVL2 is an improvement of InternVL1.5, whose architecture is described in Figure 4.

– ***Strong Vision Encoder***: The majority of existing multimodal large language models use pre-trained ViTs [12]. However, these ViTs are popularly trained on image-text pairs scraped from the Internet with a low resolution, so their performance decreases when processing high-resolution images. To overcome this challenge, InternViT is introduced. The visual features learned by this model are broadly applicable, not only to specific large language models.
– ***Dynamic High-Resolution*** The authors adopt a dynamic high-resolution training approach for InternVL1.5 to adapt effectively to different input images' varying resolutions and aspect ratios.

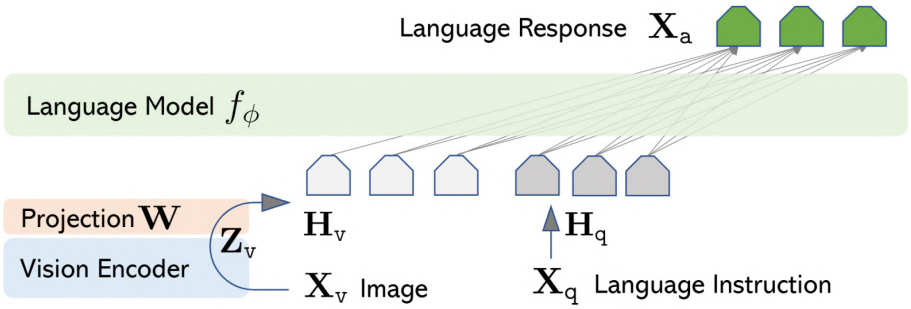


Fig. 5: Architecture of LLaVa 1.5 [25]

Architecture of LLaVa 1.5 LLaVA (Large Language and Vision Assistant) is an advanced AI model that handles both language and vision inputs. The architecture of LLaVA 1.5 is based on multimodal learning, combining both vision and language models to handle tasks like answering questions about images, image captioning, and generating detailed descriptions of visual content. Here's a high-level overview of its architecture:

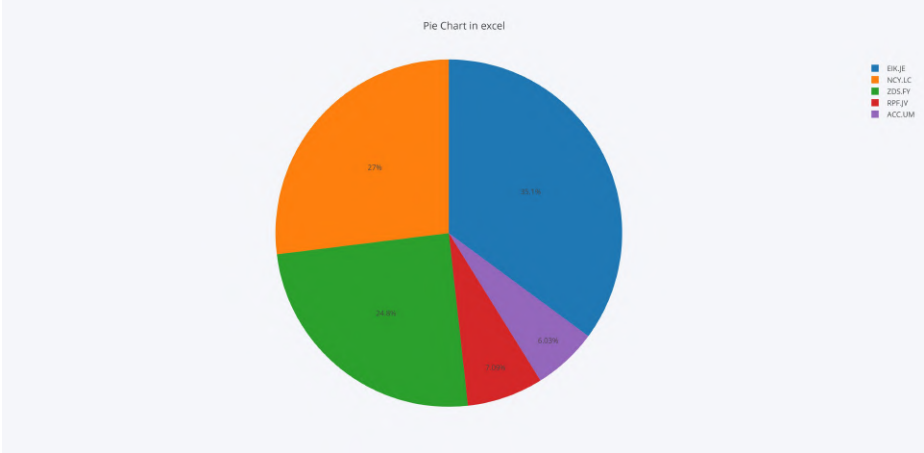
- ***Backbone Architecture***:
  - ***Language Model (LM)***: LLaVA 1.5 integrates a pre-trained large language model like GPT or LLaMA as its language backbone. The language model handles text-based tasks, generates responses, and performs reasoning based on input queries. This LM is fine-tuned to handle the combination of visual and textual information.
  - ***Vision Model***: For visual input, LLaVA employs a vision transformer (ViT) or a similar deep learning model trained on large-scale image datasets. This model processes image inputs, extracting visual features that are later combined with the language model's understanding. with linguistic tokens.
- ***Multimodal Fusion Mechanism***: The key challenge in LLaVA's architecture is combining text and image features effectively. LLaVA 1.5 uses cross-attention layers to align the features from both the language model and vision model. The vision model encodes the image into a set of embeddings (image tokens), while the language model encodes the textual inputs (word tokens). These embeddings are fused together via a multimodal transformer layer, which learns how to relate visual features with linguistic tokens.

3.5 Prompt Techniques

Alongside selecting the appropriate models, this research focuses on prompt engineering for these models. For example, some improvement is made to the prompt fed into the models so that the model can perform at its best capability in understanding and answering questions with complex visual data. Moreover, when researching the MMMU benchmark, the inference on models is performed with the prompt of the benchmark itself.

Answer shuffle. In addition, the shuffle on the order of answers is experimented to investigate the models' performance. To be more detailed, the regular order of the answers is: 1 for option A, 2 for option B, 3 for option C, and 4 for option D; after shuffling, one of the possible cases should be 3 for A, 1 for B, 4 for C, and 2 for D. The model will be asked several times, and the final option will be decided to be the answer with the most number of chosen times. See Figure 6.

Image:



Shuffle prompt 1:
Question: Which category has the smallest percentage in the pie chart?
Answer options:
A. NCY.LC B. EIK.JE C. ACC.UM D. RPF.JV
Answer with the option's letter from the given choices directly.

Shuffle prompt 2:
Question: Which category has the smallest percentage in the pie chart?
Answer options:
A. EIK.JE B. ACC.UM C. NCY.LC D. RPF.JV
Answer with the option's letter from the given choices directly.

Fig. 6: An example of the random shuffle method: In the experiments, each sample will be shuffled 5 times, and the final answer will be the most frequent candidate.

Voting ensemble. Additionally, the ensemble of several models is developed, combining the outputs from different sources to help improve the accuracy of the answer. For example, if answers of `Model1(image, prompt)` on 5 times shuffle is `[1,2,2,3,4]` and answers of `Model2(image, prompt)` on 5 times shuffle is `[1,2,2,2,2]`, the final answer will be `2`.

Post-processing. In the majority of experiments, the models can directly provide an answer from options A, B, C, or D. However, in rare instances where

the model fails to generate a direct answer, we decide to automatically assign option A as the response, given the infrequency of such occurrences.

Additional prompt. Another approach asks the models with given questions and whether it is true or false if the answer is A. The same things are implemented for B, C, and D. Based on the result of the model, the final answer to the question will be proposed.

4 Experiments and Results

4.1 Metric.

LAVA Challenge used MMMU to evaluate submission results:

$$\text{Micro-averaged accuracy} = \frac{\sum (\text{TP})}{\sum(\ (\text{TP}) + (\text{FP}) + (\text{FN}))}$$

Which:

- ***TP(True Positives):*** Correctly predicted instances across all labels or classes.
- ***FP(False Positives):*** Instances incorrectly predicted as belonging to a class when they don't.
- ***FN(False Negatives:*** Instances that were not predicted as belonging to a class but should have been.

The total score will be calculated by:

$$\text{Total score} = 0.3(\text{Private score}) + 0.7(\text{Public score})$$

4.2 Settings.

In experiments, about the shuffle of answer order, we asked the model 5 times, and the answer chosen the most number of times would be our final solution for the question. The ideal situation is that an option is selected in more than 2 times. However, there are some cases in which 2 options are equally responded by the model, for example, the model chooses option A 2 times, option B 2 times, and option C 1 time. In such cases, we will ask the model the same question again, and the answer now contains only the option selected most time, in the example, they are A and B. The model selection in this inference would be the final answer. The same approach is applied to the true-false strategy: if there is more than one true answer, the model will be asked again with only those answers.

In the actual examination, the Qwen2-VL model is noticed to produce better accuracy than the other selected models (LLaVA and InternVL2); therefore, most of our prompt improvement focuses on different versions of this model

(Qwen-VL-7B-Instruct-GPTQ-Int4, Qwen-VL-7B-Instruct-GPTQ-Int8, Qwen2-VL-7B-Instruct, etc). Our answer order shuffle, true-false approach, and model assembly are applied to this model only.

For the LAVA Workshop challenge, the submission that receives the highest result on the public dataset is the assembly of Qwen2-VL-72B-Instruct-GPTQ-Int8 and Qwen2-VL-7B-Instruct models with 2 times vote results for Qwen2-VL-72B-Instruct-GPTQ-Int8 and 1 time for Qwen2-VL-7B-Instruct. This submission is conducted following the prompt.

Question: <Question> Answer options: A. <Option 1 (after being shuffled)> B. <Option 2 (after being shuffled)> C. <Option 3 (after being shuffled)> D. <Option 4 (after being shuffled)> Answer with the option's letter from the given choices directly.

Table 1: Prompt 1

The reason for having the last sentence (Answer with the option's letter from the given choices directly) is to force the model to answer the letter only, prevent it from answering a wall of text, and make the decision difficult.

There are different prompts used in our experiments, which are defined as follows:

Analyze the attached image and answer the following question with precision: Question: <Question> Options: 1. <Answer 1> 2. <Answer 2> 3. <Answer 3> 4. <Answer 4> Please respond by selecting the option number that best fits the image's content. Be precise and choose only one option.

Table 2: Prompt 2

Look at the image and answer the following question: Question: <Question> Options: 1. <Answer 1> 2. <Answer 2> 3. <Answer 3> 4. <Answer 4> Provide your answer by choosing the number corresponding to the correct option. For example, if the correct answer is the first option, respond with "1".

Table 3: Prompt 3

Based on the given image, answer the following question: Question: <Question> Options: 1. <Answer 1> 2. <Answer 2> 3. <Answer 3> 4. <Answer 4> Your task is to analyze the image and choose the correct answer by providing the number of options you believe are correct. Only the number is required for your response.

Table 4: Prompt 4

Please answer the question: <Question>, by looking at the attached image and the following answers: 1. <Answer 1> 2. <Answer 2> 3. <Answer 3> 4. <Answer 4> You must answer by number of the answer. For example, if you think the answer is option 1, please write 1.

Table 5: Prompt 5

Look at the image, based on the information in the image, answer the following question. Choose the correct answer among four options. Question and answers are in <language>. You should only choose the correct answer, without any further explanation. Question: <Question> Answers: 1. <Answer 1> 2. <Answer 2> 3. <Answer 3> 4. <Answer 4>

Table 6: Prompt 6

Of the 4 answers, based on the image and question, which is the correct answer (just write the correct answer number) to the following question: Question: <Question> Answers: 1. <Answer 1> 2. <Answer 2> 3. <Answer 3> 4. <Answer 4>

Table 7: Prompt 7

Qwen: You are an AI assistant specializing in multimodal understanding. Analyze the following question and multiple choice answers related to various topics. Your task is to select the most accurate response. Think through your reasoning carefully, but only output the letter corresponding to your final answer. Question: <Question> Answers: A. <Answer A> B. <Answer B> C. <Answer C> D. <Answer D> Provide only your final answer as a single letter: A, B, C, or D.

Table 8: Prompt 8

The answer to the question: <question>is <option (option is four options A, B, C, D). True or False? A. True B. False Answer with the option's letter from the given choices directly.

Table 9: Prompt 9

4.3 Results.

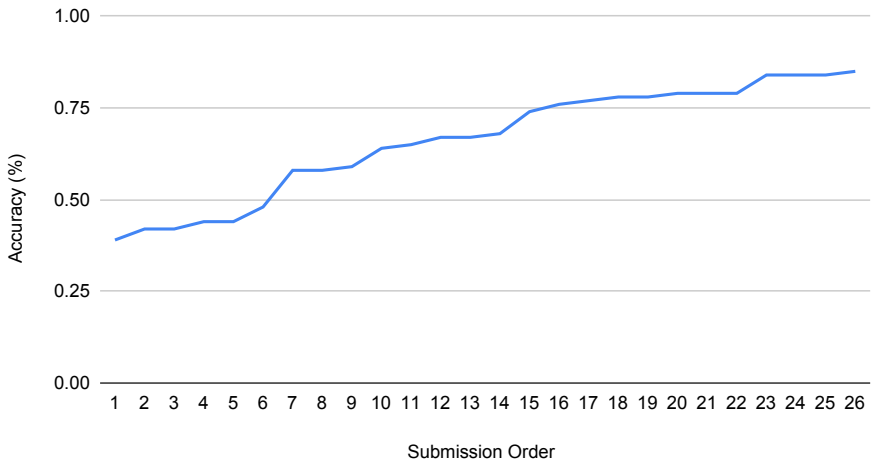


Fig. 7: Accuracy Trend by Submission Order

Table 10 illustrates our results on the public dataset of the LAVA workshop challenge, round by 2 decimal number. The highest result is 0.85, achieved by the assembly of Qwen2-VL-72B-Instruct-GPTQ-Int8 and Qwen2-VL-7B-Instruct,

with the random shuffle of answer order, and prompt 1. From table 10, we can have a brief summary of results based on the prompt:

- Prompt 1 and Prompt 8 consistently resulted in the highest scores, with top-performing models reaching up to 0.85.
- Prompts 2, 3, and 4 yielded lower results across models, with Prompt 3 showing the weakest performance.
- Prompt 7 and Prompt 9 also produced solid results, particularly for larger models like Qwen2-VL and InternV2.

Table 11 shows the results of the winning teams on the LAVA Workshop challenge. Our method scores the first prize on the challenge, with 0.85 on both public and private datasets.

Model	**Prompt**	**Answer order shuffle**	**Result**
LLAVA-v1.5-7B	Prompt 2	✗	0.39
LLAVA-v1.6-34B	Prompt 2	✗	0.42
LLAVA-v1.6-34B	Prompt 3	✗	0.42
LLAVA-1.6-13B	Prompt 3	✗	0.44
LLAVA-v1.6-34B	Prompt 4	✗	0.44
Use the vote between 3 LLAVA models	Prompt 4	✗	0.48
LLAVA-v1.6-34B	Prompt 5	✗	0.58
InternVL2-Llama3-76B	Prompt 6	✗	0.58
Qwen2-VL-2B-Instruct	Prompt 7	✗	0.59
Qwen-VL-7B-Instruct-GPTQ-Int4	Prompt 7	✗	0.64
InternVL2-8B	Prompt 6	✗	0.65
Qwen-VL-7B-Instruct-GPTQ-Int8	Prompt 7	✗	0.67
InternVL2-26B	Prompt 6	✗	0.67
InternVL2-Llama3-76B	Prompt 7	✗	0.68
Qwen2-VL-7B-Instruct-GPTQ-Int4	Prompt 1	✗	0.74
InternVL2-40B	Prompt 1	✗	0.76
Qwen2-VL-7B-Instruct-GPTQ-Int4	Prompt 2	✗	0.77
Qwen2-VL-7B-Instruct-GPTQ-Int8	Prompt 2	✗	0.78
Qwen2-VL-7B-Instruct	Prompt 8	✗	0.78
Qwen2-VL-7B-Instruct-GPTQ-Int8	Prompt 1	✗	0.79
Qwen2-VL-7B-Instruct (True/False approach)	Prompt 9	✗	0.79
Qwen2-VL-7B-Instruct	Prompt 1	✓	0.79
Qwen2-VL-72B-Instruct-GPTQ-Int8	Prompt 1	✗	0.84
Qwen2-VL-72B-Instruct-GPTQ-Int8	Prompt 8	✗	0.84
Qwen2-VL-72B-Instruct-GPTQ-Int8	Prompt 1	✓	0.84
Qwen2-VL-72B-Instruct-GPTQ-Int8 ***assembles with*** **Qwen2-VL-7B-Instruct**	**Prompt 1**	✓	**0.85**

Table 10: Results on public dataset of LAVA Workshop challenge

	Public Score	**Private Score**	**Total Score**
WAS (Ours)	**0.85**	**0.85**	**0.85**
MMLAB-UIT	0.83	0.84	0.84
V1olet	0.82	0.82	0.82

Table 11: Results on winning teams

5 Conclusion

Developing a Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI has become crucial in today's fast-paced and information-rich environment. This work is vital in supporting humans by enabling faster and more efficient access to knowledge across various disciplines. Understanding and reasoning through multimodal data is essential for handling complex, real-world problems that span multiple fields of expertise.

In conclusion, this research contributes to advancing the field by presenting a simple yet effective approach to solving the LAVA challenge. Using prompt engineering and an ensemble method based on voting, we demonstrate how multiple models can collaborate to improve performance on multimodal tasks. By leveraging the strengths of each model in the ensemble, the proposed method not only enhances accuracy but also showcases a scalable solution that can be applied to other similar tasks. This approach could be a foundation for future work in expert-level AGI systems, aiming to address increasingly complex multimodal problems with greater efficiency and precision.

Appendices

Table 12 shows the accuracy of the three models used in this research on the MMMU benchmark. Based on the results in this table, we have decided to proceed with Qwen2, which achieves the highest scores and has a public inference API, enabling readers to reproduce our results in most of our experiments easily.

Model	**MMMU-Pro**	**MMMU (Val)**
Qwen2-VL-72B	46.2%	64.5%
InternVL2-Pro		62.0%
InternVL2-Llama3-76B	40.0%	58.3%
LLaVA-OneVision-72B	31.0%	56.8%
InternVL2-40B	34.2%	55.2%
InternVL2-8B	29.0%	51.2%

Table 12: Accuracy of the Qwen2-VL, InternVL2, and LLaVA in MMMU benchmark

References

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Batra, D., Parikh, D.: Vqa: Visual question answering (2016), https://arxiv.org/abs/1505.00468 3
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning (2022), https://arxiv.org/abs/2204.14198 3
3. Alayrac, J.B., Recasens, A., Bolte, B., Rao, D.S., Bochkovskiy, A., Donahue, J., Korbar, B., Fidler, S., Malinowski, M.: Flamingo: A visual language model for few-shot learning. In: NeurIPS (2022) 3
4. Awadalla, A., Gao, I., Gardner, J., Hessel, J., Hanafy, Y., Zhu, W., Marathe, K., Bitton, Y., Gadre, S., Sagawa, S., Jitsev, J., Kornblith, S., Koh, P.W., Ilharco, G., Wortsman, M., Schmidt, L.: Openflamingo: An open-source framework for training large autoregressive vision-language models (2023), https://arxiv.org/abs/2308.01390 3
5. Chen, Y.C., Li, L., Yu, L., Kholy, A.E., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations. In: ECCV. Springer (2020) 3
6. 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) 5, 6
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., Dai, J.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. arXiv preprint arXiv:2312.14238 (2023) 5
8. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/ 3
9. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Mokra, S., Tay, Y., Vinyals, O., Li, X., Wang, X., Narang, S., et al.: Scaling instruction-finetuned language models. arXiv:2210.11416 (2022) 3
10. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023), https://arxiv.org/abs/2305.06500 3
11. Dai, W., Li, X., Zhang, C., Hu, X., Li, J., Yin, X.: Minigpt-4: Enhancing chatgpt with multimodal abilities. arXiv:2304.10592 (2023) 3
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021) 3, 5, 6
13. Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., Li, H., Qiao, Y.: Llama-adapter v2: Parameter-efficient visual instruction model (2023), https://arxiv.org/abs/2304.15010 3
14. Huang, Y., Meng, Z., Liu, F., Su, Y., Collier, N., Lu, Y.: Sparkles: Unlocking chats across multiple images for multimodal instruction-following models (2024), https://arxiv.org/abs/2308.16463 3

15. Hudson, D.A., Manning, C.D.: Gqa: A new dataset for real-world visual reasoning and compositional question answering (2019), https://arxiv.org/abs/1902.09506 3
16. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T.: Align: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. PMLR (2021) 3
17. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning (2023), https://arxiv.org/abs/2305.03726 3
18. Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y.: Seed-bench: Benchmarking multimodal llms with generative comprehension (2023), https://arxiv.org/abs/2307.16125 3
19. Li, J., Hu, D., Zhao, H., Zhang, L., Li, X., Gao, J.: Blip-2: Bootstrapped language-image pre-training with frozen image encoders and large language models. arXiv:2301.12597 (2023) 3
20. Li, L., Yin, Y., Li, S., Chen, L., Wang, P., Ren, S., Li, M., Yang, Y., Xu, J., Sun, X., Kong, L., Liu, Q.: M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. arXiv preprint arXiv:2306.04387 (2023) 3
21. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Choi, Y., Gao, J.: Oscar: Object-semantics aligned pretraining for vision-language tasks. In: ECCV. Springer (2020) 3
22. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models (2023), https://arxiv.org/abs/2305.10355 3
23. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context (2015), https://arxiv.org/abs/1405.0312 3
24. Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L.: Aligning large multi-modal model with robust instruction tuning. arXiv preprint arXiv:2306.14565 (2023) 3
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023) 7
26. Liu, Y., Gehrig, M., Messikommer, N., Cannici, M., Scaramuzza, D.: Revisiting token pruning for object detection and instance segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) (2024) 3
27. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., Chen, K., Lin, D.: Mmbench: Is your multi-modal model an all-around player? (2024), https://arxiv.org/abs/2307.06281 3
28. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019) 3
29. 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 (2024), https://arxiv.org/abs/2310.02255 3
30. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge (2019), https://arxiv.org/abs/1906.00067 3
31. Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., Scialom, T.: Gaia: a benchmark for general ai assistants (2023), https://arxiv.org/abs/2311.12983 4
32. Monajatipoor, M., Li, L.H., Rouhsedaghat, M., Yang, L.F., Chang, K.W.: Metavl: Transferring in-context learning ability from language models to vision-language models (2023), https://arxiv.org/abs/2306.01311 3

33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. arXiv:2103.00020 (2021) 3
34. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP-IJCNLP. Association for Computational Linguistics (2019) 3
35. Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al.: Evaluation and analysis of hallucination in large vision-language models. arXiv preprint arXiv:2308.15126 (2023) 3
36. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., Fan, Y., Dang, K., Du, M., Ren, X., Men, R., Liu, D., Zhou, C., Zhou, J., Lin, J.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 5
37. Wang, Y., Li, Y., Ma, T., Ebrahimi, S., Zhong, V., Wu, F., Roy, S.: Simvlm: Simple visual language model pretraining with weak supervision. arXiv:2108.10904 (2021) 3
38. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models (2023), https://arxiv.org/abs/2306.09265 3
39. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qian, Q., Zhang, J., Huang, F., Zhou, J.: mplug-owl: Modularization empowers large language models with multimodality (2024), https://arxiv.org/abs/2304.14178 3
40. Yin, Z., Wang, J., Cao, J., Shi, Z., Liu, D., Li, M., Huang, X., Wang, Z., Sheng, L., Bai, L., et al.: Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. Advances in Neural Information Processing Systems **36** (2024) 3
41. Yu, J., Lu, J., Yin, X., Cichy, B., Goh, G., Ramesh, A., Zoph, B., Federici, M., Yu, J., Misra, I., Brock, A., Barham, P., Elsen, E., Sutskever, I., Fong, R.: Coca: Contrastive captioners are image-text foundation models. arXiv:2205.01917 (2022) 3
42. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities (2023), https://arxiv.org/abs/2308.02490 3
43. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of CVPR (2024) 3, 4
44. Zhang, M., Zhang, P., Li, X., Yang, J., Hu, X., Zhang, L., Gao, J.: Fuyu: Fully unified vision-language models for multimodal tasks. arXiv:2305.05999 (2023) 3
45. Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, Y., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: CVPR. IEEE (2021) 3
46. Zhao, H., Cai, Z., Si, S., Ma, X., An, K., Chen, L., Liu, Z., Wang, S., Han, W., Chang, B.: Mmicl: Empowering vision-language model with multi-modal in-context learning (2024), https://arxiv.org/abs/2309.07915 3
47. Zhu, Y., Yang, T., Lu, J., Fu, C., Yu, J., Xu, Z., Ni, B., Li, X., Zoph, B.: Llava: Large language and vision assistant. arXiv:2304.08485 (2023) 3