Non-Natural Image Understanding with Advancing Frequency-based Vision Encoders

Supplementary Material

Overview In this supplementary material, we present the following content.

1. Additional Examples

In **Figure 1** and **Figure 2**, we present additional qualitative examples illustrating the performance of our model on both captioning and question-answering tasks. These figures underscore the versatility of our model across multiple types of non-natural image understanding tasks.

2. Evaluation Details

Traditional metrics for evaluating the quality of descriptions, such as BLEU [17] and CIDEr [20], are limited by their reliance on the linguistic style of the reference descriptions. For instance, even for the GPT-4V, if not fine-tuned on a corresponding training set, may generate descriptions that are correct but receive low scores. Conversely, a model fine-tuned on a training set with a similar linguistic style can easily achieve high scores. To overcome this limitation, we are inspired by [14], and we consider using GPT-4 to evaluate the correctness and level of detail in the generated descriptions. The prompts used for geometric caption evaluation are shown in **Figure 3** and **Figure 4**(similar prompts for chart and function), which allows for a more objective assessment of the results produced by different models.

Table 1. **Statistics of Caption Training Data.**

Category	Statistic	Number
Total	- Total number	190K
	- Average length (words)	61.91
	- Average length (characters)	350.67
	- Vocabulary size	21127
Geometry	- Total number	60K
	- Average length (words)	44.23
	- Average length (characters)	230.28
	- Vocabulary size	2410
Chart	- Total number	30K
	- Average length (words)	109.40
	- Average length (characters)	726.73
	- Vocabulary size	20598
Function	- Total number	100K
	- Average length (words)	58.49
	- Average length (characters)	321.11
	- Vocabulary size	145

3. Training Data

In **Tables 1** and **Tables 2**, we present the training data used for alignment and instruction fine-tuning across three types of non-natural images: geometry, charts, and functions. The geometry data is sourced from Geo170K [4], the function data from Mavis-caption [22], and the chart data from Chartllama [6]. Notably, for the chart data, we extended the textual responses to provide more detailed analyses of the visual modality's impact on the results, rather than directly answering with a single number or word. All of our training data will be publicly available.

4. Baselines for the Caption task

To evaluate the performance of our model, we compare it against both **closed-source** and **open-source multimodal large language models (MLLMs)**. The baselines are divided into two categories:

4.1. Closed-Source MLLMs

These closed-Source models represent the state-of-the-art in multimodal understanding and captioning tasks. They include: Qwen-VL-Plus [2], Gemini-1.0-Pro [19], Qwen-VL-Max [2], and GPT-4V [16]. These models serve as the performance upper bound, given their advanced design and access to extensive training resources.

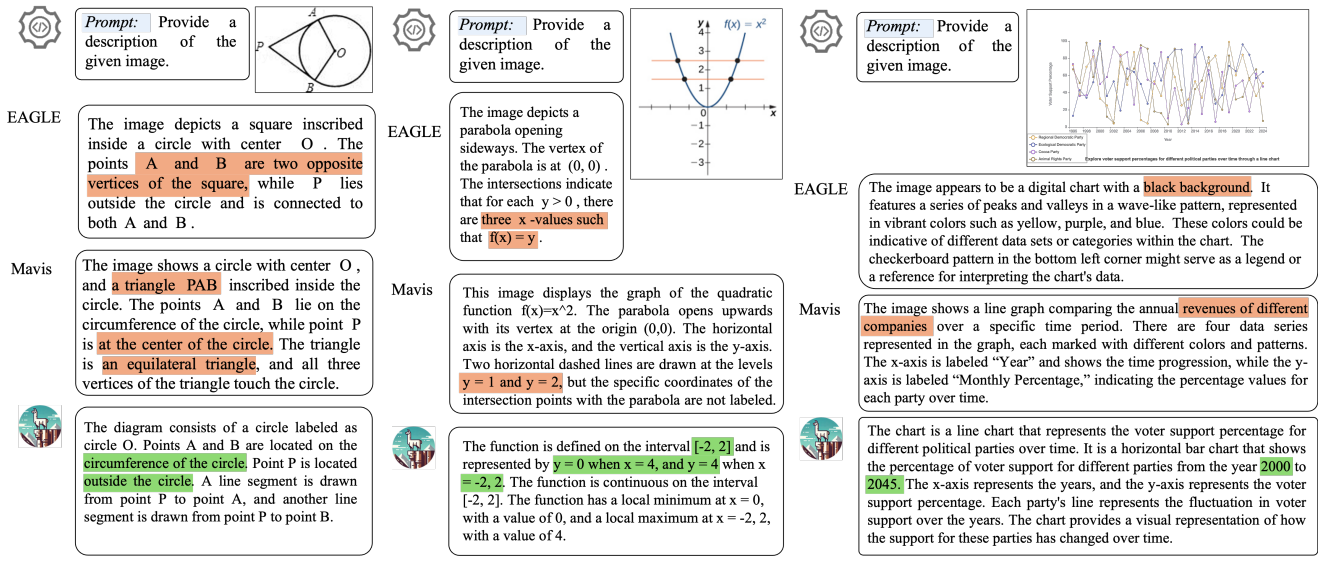

Figure 1. More presentation of caption tasks.

Table 2. **Statistics of Instruction Training Data.**

Category	Statistic	Number
Total	- Number of unique questions	247K
	- Average question length	44.35
	- Average answer length	61.62
Geometry	- Number of unique questions	117K
	- Average question length	69.88
	- Average answer length	70.19
Chart	- Number of unique questions	100K
	- Average question length	12.51
	- Average answer length	60.51
Function	- Number of unique questions	30K
	- Average question length	50.82
	- Average answer length	31.45

4.2. Open-Source MLLMs

These models are publicly available, making them more accessible for research and development. All three data types include LLaVA [13] and its variants (LLaVA-adapter [5], LLaVA-next [8]), known for their adaptability to multi-modal inputs, with visual encoders fine-tuned on open datasets. For these models, we directly use their pre-trained weights. In addition, we introduce 3 baselines for each data type, including:

Geo-Caption: For geometric image caption we selected 3 state-of-the-art baselines respectively: G-LLaVA [4], MAVIS [22], and EAGLE [9].

Function Caption: For function image caption, to the best of our knowledge, only MAVIS has attempted, so we use the function image caption dataset provided by MAVIS and train according to G-llava and EAGLE as the corresponding baseline model.

- **G-LLaVA**: First train the projection linear layer and then both the projection linear layer and the language model are trainable.
- **MAVIS**: First Train the visual encoder and then freeze the CLIP-Math in both the alignment phase and instruction fine-tuning phase, and train the projection layer along with the LoRA-based LLM.
- **EAGLE**: First train the visual encoder and projection layer and then use Lora to fine-tune the visual encoder and train the projection layer and LLM with full parameters

Chart Caption: For the chart image caption, we compare 2 chart understanding models, ChartLlama [6] and Unichart [15]. We additionally use the same chart dataset to fine-tune MAVIS, G-LLAVA, and EAGLE. The complete experimental results are shown in the **Table 3**.

5. Baselines for the Q&A task

5.1. For GeoQA

In our experimental setup, we compare our method against three categories of baselines: heuristics-based baselines, conventional models, and multi-modal large language models (MLLMs). The heuristics-based baselines include Random Chance and Frequent, which provide simple reference points for evaluating performance. Conventional models such as Geoformer [3] and UniMath [10] represent traditional task-specific solutions designed for mathematical and geometric reasoning. The MLLMs include state-of-the-art open-source models like LLAVA-1.5 [13] (7B and 13B), Math-LLAVA [18], G-LLAVA [4] (7B and 13B), MAVIS [22], and EAGLE [9], which are based on ad-

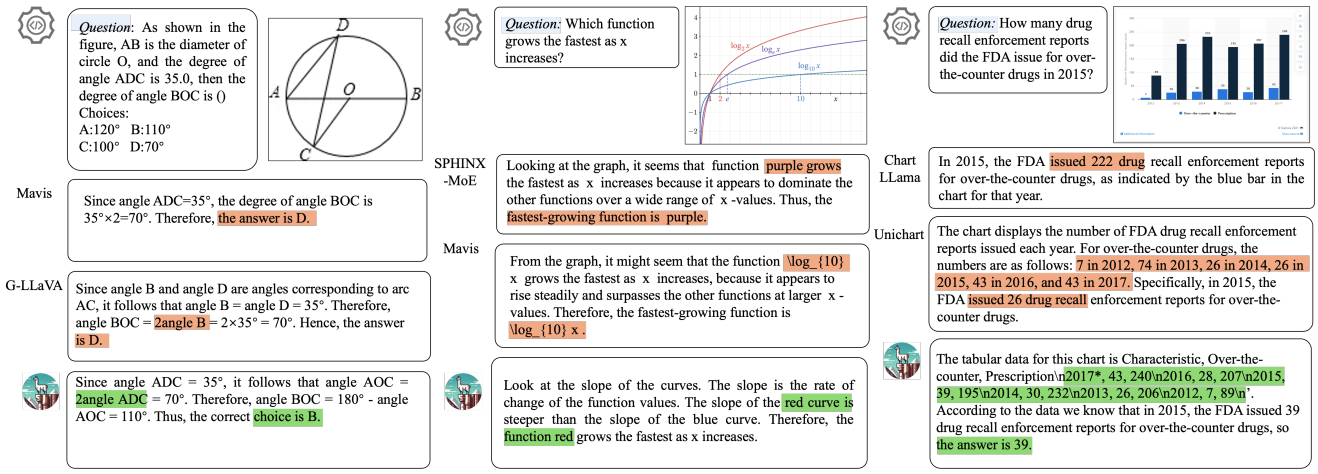

Figure 2. More presentation of question-answer tasks.

Table 3. The complete experimental results of Chart Caption

Model	Correctness	Detail
Closed-source MLLMs		
Qwen-VL-Plus [2]	1.75	2.15
Gemini-1.0-Pro [19]	2.00	2.20
Qwen-VL-Max [2]	**2.35**	2.30
GPT-4V [16]	1.88	**2.33**
Open-source MLLMs		
LLAVA [13]	2.05	2.45
LLAMA-Adapter [5]	2.19	2.65
LLAVA-NeXT [8]	2.26	2.72
G-LLAVA [4]	2.25	2.79
MAVIS [22]	2.31	2.78
EAGLE [9]	2.33	2.75
Ours	**2.35**	**2.84**

vanced architectures such as Vicuna and LLAMA-2. These baselines ensure a comprehensive evaluation of our method against a range of existing approaches.

5.2. For FunctionQA

In our experimental setup, we compare EDGE against three categories of baselines: heuristics-based baselines, closed-source MLLMs, and open-source MLLMs. The heuristics-based baseline includes Random Chance, which serves as a simple reference point. The closed-source MLLMs include advanced proprietary models such as CoT GPT-4 [1], PoT GPT-4 [1], Multimodal Bard [19], and GPT-4V [16], which represent the state-of-the-art in multimodal reasoning tasks. The open-source MLLMs consist of models like LLAVA [21], LLAMA-Adapter [5], LLAVA-NeXT [8], SPHINX-MoE [11], and MAVIS [22], which leverage accessible architectures such as Vicuna-1.5 and MAmmoTH-2-7B for multimodal tasks. These baselines provide a comprehensive comparison, highlighting the effectiveness of our model, which achieves competitive performance against both open-source and closed-source approaches.

Table 4. Ablation experiments of the visual encoder with different adjustment strategies in stage 3.

Vision Encoder in Stage 3	Full Fine-tune	Freeze	LoRA
Performance on GeoQA	65.7	67.3	**68.2**

5.3. For ChartQA

In our experiments on the ChartQA benchmark, we compare our method against both conventional models and multi-modal large language models (MLLMs). The conventional models include Pix2struct [7], Matcha [12], Chart-T5 [23], and Unichart [15], which are specialized methods designed for processing and reasoning with chart data. The multi-modal large language models consist of SPHINX [11], Qwen [2], and Chartllama [6].

6. Ablation of Vision Encoder in Stage 3

The visual encoder is fully fine-tuned in stages 1 and 2. In stage 3, visual encoder is frozen and LoRA is used for further adjustments. This approach is based on the fact that, as the LLM backbone is fine-tuned on the QA dataset and acquires knowledge from step-by-step rationales, the LoRA-based visual encoder progressively shifts focus toward key geometric cues critical to the resolution process, rather than the entire image. This synergy enhances a deeper understanding of geometric features. We include an ablation study of the adjustment strategy, as shown in the Table above, which confirms that LoRA tuning yields the best performance. The projection layer is fine-tuned in stages 2 and 3, while the LLM is adjusted only in stage 3.

```
// Evaluation of the correctness of the geometry
{
   "role": "system",
   "content":
      "You are an intelligent chatbot designed for evaluating the accuracy of generative outputs for geometry-based image question-answer pairs. "
      "Your task is to compare the predicted answer with the correct answer and determine its accuracy in terms of geometric shapes and properties. Here's how you can accomplish the task:"
      "------"
      "##INSTRUCTIONS: "
      "- Check if the predicted answer includes all the correct geometric shapes and properties present in the correct answer.\n"
      "- Evaluate whether the predicted answer accurately describes the geometric elements, including their relationships and characteristics.\n"
      "- Consider synonyms or paraphrases as valid matches.\n"
      "- Provide a single evaluation score that reflects the accuracy of the prediction in terms of geometric correctness."
},
{
   "role": "user",
   "content":
      "Please evaluate the following geometry-based image question-answer pair:\n\n"
      f"Question: {question}\n"
      f"Correct Answer: {answer}\n"
      f"Predicted Answer: {pred}\n\n"
      "Provide your evaluation only as an accuracy score where the accuracy score is an integer value between 0 and 5, with 5 indicating the highest level of accuracy. "
      "Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the accuracy score in INTEGER, not STRING."
      "DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. "
      "For example, your response should look like this: {'score': 4}."
}
```

Figure 3. Evaluation of the correctness of the geometry.

```
//Evaluation of geometric details
{
   "role": "system",
   "content":
      "You are an intelligent chatbot designed for evaluating the detail orientation of generative outputs for geometry-based image question-answer pairs. "
      "Your task is to compare the predicted answer with the correct answer and determine its level of detail, considering both completeness and specificity. Here's how you can accomplish the task:"
      "------"
      "##INSTRUCTIONS: "
      "- Check if the predicted answer covers all major points from the image. The response should not leave out any key aspects related to the geometric elements.\n"
      "- Evaluate whether the predicted answer includes specific details rather than just generic points. It should provide comprehensive information that is tied to specific elements of the image.\n"
      "- Consider synonyms or paraphrases as valid matches.\n"
      "- Provide a single evaluation score that reflects the level of detail orientation of the prediction, considering both completeness and specificity."
},
{
   "role": "user",
   "content":
      "Please evaluate the following geometry-based image question-answer pair:\n\n"
      f"Question: {question}\n"
      f"Correct Answer: {answer}\n"
      f"Predicted Answer: {pred}\n\n"
      "Provide your evaluation only as a detail orientation score where the detail orientation score is an integer value between 0 and 5, with 5 indicating the highest level of detail orientation. "
      "Please generate the response in the form of a Python dictionary string with keys 'score', where its value is the detail orientation score in INTEGER, not STRING."
      "DO NOT PROVIDE ANY OTHER OUTPUT TEXT OR EXPLANATION. Only provide the Python dictionary string. "
      "For example, your response should look like this: {'score': 4}."
}
```

Figure 4. Evaluation of geometric details.

7. Limitation and Future Work

Despite the significant progress achieved by FM-ViT and the edge model in understanding non-natural images, there are still some limitations that warrant further investigation. First, while the frequency modulation technique enhances the ability to capture high-frequency information, it remains insufficient in extracting fine-grained details, such as subtle protrusions on a line or isolated points. Second, this study primarily focuses on improving the visual encoding capabilities for non-natural images, without delving into the impact of larger and more diverse datasets, leaving this aspect for future exploration.

In future work, we plan to further enhance the model's ability to extract fine-grained information from non-natural images. Additionally, we aim to generate synthetic datasets with greater diversity and realism through advanced data augmentation methods to further improve the model's understanding and generalization capabilities for broader applications, such as education, accessibility tools, and data-driven policymaking.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 1, 3

[3] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. *arXiv preprint arXiv:2212.02746*, 2022. 2

[4] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*, 2023. 1, 2, 3

[5] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2, 3

[6] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. 1, 2, 3

[7] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 3

[8] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 2, 3

[9] Zhihao Li, Yao Du, Yang Liu, Yan Zhang, Yufang Liu, Mengdi Zhang, and Xunliang Cai. Eagle: Elevating geometric reasoning through llm-empowered visual instruction tuning. *arXiv preprint arXiv:2408.11397*, 2024. 2, 3

[10] Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. Unimath: A foundational and multimodal mathematical reasoner. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133, 2023. 2

[11] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 3

[12] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. Matcha: Enhancing visual language pretraining with math reasoning and chart derendering. *arXiv preprint arXiv:2212.09662*, 2022. 3

[13] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 2, 3

[14] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 1

[15] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*, 2023. 2, 3

[16] OpenAI. Openai: Gpt-4v(ision) system card. https://openai.com/research/gpt-4v-system-card, 2023. 1, 3

[17] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 1

[18] Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *arXiv preprint arXiv:2406.17294*, 2024. 2

[19] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1, 3

[20] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evalua-

tion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1

[21] Ruyi Xu, Yuan Yao, Zonghao Guo, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, Maosong Sun, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. *arXiv preprint arXiv:2403.11703*, 2024. 3

[22] Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. Mavis: Mathematical visual instruction tuning. *arXiv preprint arXiv:2407.08739*, 2024. 1, 2, 3

[23] Mingyang Zhou, Yi R Fung, Long Chen, Christopher Thomas, Heng Ji, and Shih-Fu Chang. Enhanced chart understanding in vision and language task via cross-modal pre-training on plot table pairs. *arXiv preprint arXiv:2305.18641*, 2023. 3