HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data

Supplementary Material

A. Overview

In this supplementary material, we present:

- More detailed analysis of HalluciDoctor (Section B).
- More experimental analysis (Section C).
- Additional examples (Section D).

B. HalluciDoctor Framework

B.1. Question Generation

This section provides specific steps for answer-based question generation. We utilize ChatGPT [13] as the powerful question generator, covering a broad spectrum of semantic chunks and various question types. Specifically, we construct the prompt template shown in Figure 1 (a), filling the context description and answer blocks into slots to generate all corresponding questions. These questions, covering various types, effectively reflect meaningful semantic information in the descriptions.

B.2. Consistency Cross-checking Analysis

Considering the importance of the threshold in consistency cross-checking for identifying hallucinatory chunks, we further explore the effect of consistency cross-checking under different threshold setups in Table 1. Here, we assess the impact of different consistency thresholds by evaluating the average performance on CHAIR and MME benchmarks.

Our observations indicate that at lower consistency thresholds, as the threshold increases, hallucinatory descriptions are progressively detected and eliminated. This provides MLLMs with higher-quality instruction data for fine-tuning, thereby reducing the likelihood of hallucinatory outputs while enhancing model performance. However, when the consistency threshold exceeds 0.5, there's a significant decline in model performance. The possible reason is that HalluciDoctor eliminates almost all answer blocks as hallucinations when the consistency threshold is too high, resulting in a substantial loss of accurate semantics. Therefore, to effectively reduce hallucinations in MLLM outputs while ensuring competitive performance, we select **0.5** as the final threshold for the consistency cross-checking stage in our HalluciDoctor framework. To enhance the precision of hallucination detection and elimination, we will explore more advanced approaches for computing consistency scores [2, 17] in the future.

B.3. Hallucination Elimination

This section provides specific steps for hallucination elimination in the visual instruction data. To eliminate detected hallucinations while preserving meaningful semantics in the original descriptions, we employ ChatGPT to refine the descriptions. Specifically, we input both the hallucinatory phrases and the original descriptions into the prompt template shown in Figure 1 (b), prompting ChatGPT to remove hallucinatory phrases without altering the sentence structure. We also show some examples in Figure 1 (b). These refined descriptions are then employed to update the visual instructions in LLaVA, efficiently creating the rectified dataset LLaVA++.

Consistency threshold	w/o HalluciDoctor	0.1	0.3	0.5	0.7	0.9
$\overline{\text{CHAIR}}$ (%) ↓	21.7	19.9	16.1	<u>13.8</u>	14.2	**13.6**
$\overline{\text{MME performance}}$ (%) ↑	1148.9	1153.2	<u>1178.6</u>	**1207.2**	1012.9	739.1

Table 1. The influence of different consistency threshold for the hallucination elimination in visual instruction data.

B.4. Visual Instruction Expansion

This section provides detailed steps for selecting target images and how to add hallucinatory objects into target scenes. Firstly, we filter target scenes based on object detection to ensure the specific hallucinatory object o is absent. We then generate candidate objects and their corresponding masks for counterfactual image synthesis using text-to-image models [14, 16] and object detection tools [11]. Subsequently, we provide the LLM with image sizes and foreground object locations of target scenes to enable it to determine suitable positions and scaling. Finally, we employ structure-preserving filtering based on the depth map L1 distance for natural image incorporation.

In this way, counterfactual instruction expansion focuses on detailed and unusual instruction modifications, necessitating MLLMs to perceive fine-grained concepts for comprehensive instruction alignment. Consequently, it will generate fewer hallucinations (*e.g.*, 13.8% → 12.0% in MiniGPT-4) and demonstrate superior proficiency in perceiving specific information (*e.g.*, shop's name). This approach can also alleviate the adverse impact of long-tail distributions in various domains [19].

C. More Experimental Analysis

C.1. Experiment Details

Implementation Details. As for MiniGPT4, we initialize from its checkpoint of the first pretraining stage and only fine-tune the linear projection layer of the model for 10000 steps. As for mPLUG-Owl, we train the text encoder with

Dataset	Accuracy	F1
w/ LLaVA [10]	75.1	77.8
w/ LRV [9]	64.4	72.6
w/ LLaVA+	79.1	80.0
w/ LLaVA++	**80.1**	**80.4**

(a) Random Setting

Dataset	Accuracy	F1
w/ LLaVA [10]	65.6	71.7
w/ LRV [9]	63.4	72.3
w/ LLaVA+	74.0	75.3
w/ LLaVA++	**76.3**	**75.9**

(b) Popular Setting

Dataset	Accuracy	F1
w/ LLaVA [10]	63.2	70.5
w/ LRV [9]	60.7	70.6
w/ LLaVA+	68.5	72.0
w/ LLaVA++	**71.9**	**74.2**

(c) Adversarial Setting

Table 2. Zero-shot object hallucination results for MiniGPT-4 [21] fine-tuned with various visual instructions on POPE [9] evaluation. We follow the official setup, which involves using three different strategies (*i.e.*, random, popular, and adversarial setting) to sample objects not present in the images and then computing the corresponding accuracy and F1 scores.

the LoRA [3] strategy and prepare for 4000 steps. Due to limited computing resources, we set the micro-batch size to 4 and only fine-tuned the 7B model with NVIDIA RTX 3090. To make a fair comparison in our experiments, we only change the visual instruction data under different setups and keep other parameters the same as original models.

Evaluation Setups. MSCOCO [8] is a comprehensive dataset with 80 object categories used for diverse vision tasks. Visual Genome [5] is another vision dataset with more detailed visual information like bounding boxes and region captions. We select the overlapped images from MSCOCO and VG to construct validation images, aiming to encompass annotations of various objects, relationships, and attributes. Additionally, we employ powerful visual foundation models [6, 11, 18] to identify objects, relations, and attributes of images in the validation set, thereby enriching the ground truth labels. In the validation stage, we will extract object, relation, and attribute phrases from the description of MLLMs that are fine-tuned on different visual instruction datasets, and calculate the corresponding hallucinatory metrics by matching them to ground truth labels.

GPT-4 Evaluation. We show the GPT-4 evaluation's prompt templates for detailedness and accuracy in Figure 2.

C.2. POPE Results

We compare the MLLM fine-tuned on our more robust dataset LLaVA+ and LLaVA++ against the baseline dataset on POPE evaluation [7] in Table 2. Although POPE is tailored for close-ended questions of object hallucinations, rendering it unsuitable for our comprehensive evaluation of various hallucinations in visual instruction data, our approach also shows a similar tendency to our main results that LLaVA+ and LLAVA++ from HalluciDoctor achieve consistent gains in all accuracy and F1 score. The results indicate that HalluciDoctor is effective in correcting object-level hallucinations. In addition, the MLLM fine-tuned on LLAVA++ obtains the highest accuracy and F1 score, demonstrating that the more robust visual instruction dataset can enhance MLLMs' ability to discern negative instructions, especially in the more challenging adversarial setting.

Dataset	Captioning	VQA	
	NoCaps (val) ↑	GQA ↑	AOK-VQA ↑
Faithful Prompt	101.5	40.5	56.1
LURE [20]	93.9	41.4	58.3
VIGC [15]	96.6	41.0	58.9
MiniGPT4$_{+LRV}$ [9]	103.9	40.7	57.6
MiniGPT4$_{+LLaVA++}$	104.1	**43.7**	60.1
mPLUG-Owl$_{+LLaVA++}$	**104.4**	43.3	**61.0**

Table 3. Overview performance comparison on conventional zero-shot vision-language tasks (*i.e.*, captioning, VQA).

C.3. Zero-shot Vision-Language Task Results

As a versatile MLLM, the model's performance cannot be compromised by instruction fine-tuning. On the contrary, by eliminating hallucinatory information in the training data, the MLLM demonstrates stronger generalization capabilities for conventional visual tasks. We perform the quantitative evaluation on the zero-shot vision-language tasks based on captioning (NoCaps [1]) and visual question answering (GQA [4], AOK-VQA [12]). Table 3 provides an overview of the performance of HalluciDoctor on various zero-shot vision-language tasks. Compared to other works on hallucination elimination, our method achieves better generalization performance on traditional vision tasks.

D. Additional Examples

D.1. Evaluation of Visual Instruction Data

Dataset visualization. Figure 3 shows some more visualized examples in the rectified dataset LLaVA+ and more robust dataset LLaVA++.

Dataset evaluation. Similar to Sec. 5.5, we perform a manual evaluation of the generated data for more accurate results. We sample 200 instructions from LLaVA+ and LLaVA, assessing their accuracy and quality. LLaVA+ not only shows higher accuracy scores than LLaVA (451 v.s. 371) but also maintains comparable quality (405 v.s. 412).

D.2. MLLMs' Inference Analysis

We compare the outputs of MiniGPT-4 [21] fine-tuned on LLaVA [10], LRV-Instruction [9], LLaVA+, and LLAVA ++ on various types of images and show the visualized results in Figure 4. The results verified that LLaVA+ effectively helped MLLMs eliminate hallucinatory descriptions, and LLAVA ++ further added reliable detailed descriptions.

Answer-based Question Generation

You are a language assistant that helps to generate appropriate questions according to the given answer chunks and the context description.

Examples:
Description:
Please give me meaningful and answerable questions corresponding to the following answers based on the given context to help me understand the context. Please ensure that each question doesn't involve 'How many' and is concise to exactly match the corresponding answer.
Answer:
["sky is cloudy", "man fish on lawn", "man next to river", "man in background", "trees on side of river"]
Question:

1. What is the current weather condition?
2. Where is the man fishing?
3. What is the man's proximity to the river?
4. Who can be seen in the background?
5. What can be observed on the other side of the river?

…

Description:
{***description***}
Please give me meaningful and answerable questions corresponding to the following answers based on the given context to help me understand the context. Please ensure that each question doesn't involve 'How many' and is concise to exactly match the corresponding answer.
Answer:
{***answer***}

(a) The details of the prompt design for *Question Generation* in HalluciDoctor.

Hallucination Elimination

You are a language assistant that helps to refine a passage with wrong phrases removed. Given a passage and wrong phrases, you are required to remove all of them in the passage and output the refined passage in a fluent and natural style, following these rules:
1. Try to remove wrong phrases and do not use other phrases to replace

Examples:
Passage:
In addition to the sandwiches of various sizes, a bowl, a cup, and a spoon can be seen on the table, suggesting that the guests are sharing food and drinks.
Wrong phrases:
['spoon', 'drinks', 'sandwiches is various sizes']
Refined passage:
In addition to the sandwiches, a bowl and a cup can be seen on the table, suggesting that the guests are sharing food.

Passage:
The image depicts a scene of two giraffes standing on a dirt road near a fence. There are three cars parked in the background, with one on the left side and two more on the right side.
Wrong phrases:
['cars', 'cars are three']
Refined passage:
The image depicts a scene of two giraffes standing on a dirt road near a fence.

Passage:
{***passage***}
Wrong phrases:
{***hallucination phrase***}
Refined passage:

(b) The details of the prompt design for *Hallucination Elimination* in HalluciDoctor.

Figure 1. The details of the prompt design in HalluciDoctor. There are injectable slots in the prompts, such as ***description***, ***answer***, ***passage***, and ***hallucination phrase***. These slots are uniformly replaced with the corresponding text before being fed into the LLM.

GPT-4 Evaluation-Detailedness

Suppose you are an image detail annotator who judges the degree of sentence diversity based on the number of objects, relations, and attributes.
Please just provide the diversity score(1-5) for the below descriptions without any explanation, where longer caption with more content give a higher diversity score. The output format: [x,...]
Descriptions:
caption 1: {description_1}
caption 2: {description_2}
caption 3: {description_3}
caption 4: {description_4}
caption 5: {description_5}
Output:

GPT-4 Evaluation-Accuracy

Suppose you are a hallucination annotator who judges the degree of hallucination based on the number of errors in the description of objects, relations, and attributes, and you have the following real image information.
Reference captions: {coco_captions}
Bounding box: {bounding_box}
Please just provide the hallucination score(1-5) for the below descriptions without any explanation, where the fewer descriptive errors in the caption, the higher the hallucination score given. The output format: [x,...]
Descriptions:
caption 1: {description_1}
caption 2: {description_2}
caption 3: {description_3}
caption 4: {description_4}
caption 5: {description_5}
Output:

Figure 2. The details of the prompt design for GPT-4 evaluations.

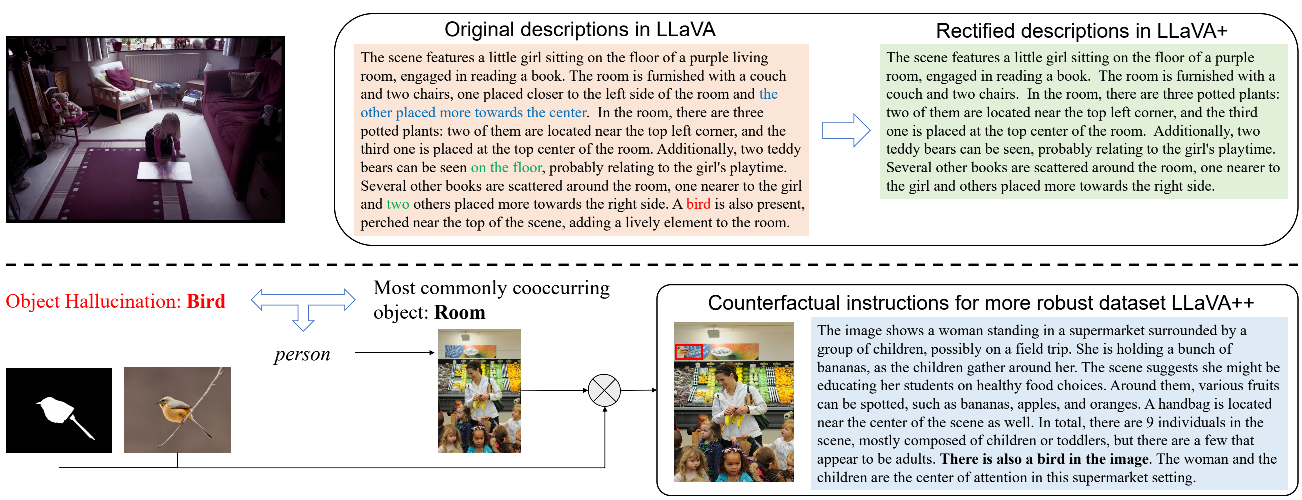

Figure 3. The instruction data visualization of LLaVA+ after hallucination elimination and LLaVA++ after visual instruction expansion.

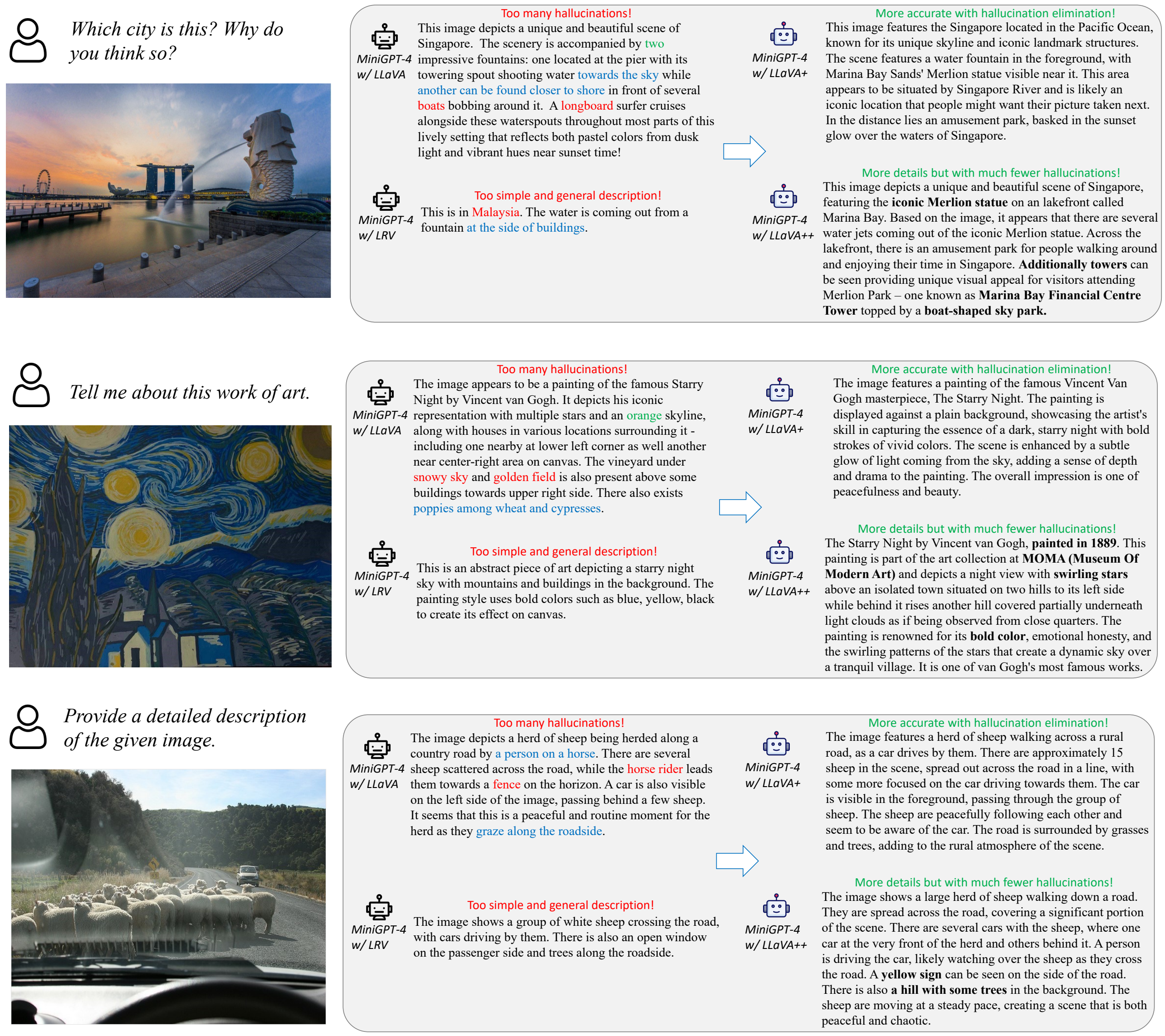

Figure 4. More visualization of MLLM comparison on various visual perception cases (multi-round conversation, single-round conversation, detailed description, etc.).

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 2

[2] Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. *arXiv preprint arXiv:2305.08281*, 2023. 1

[3] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[4] Drew A Hudson and Christopher D Manning. 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*, pages 6700–6709, 2019. 2

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 2

[6] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2

[7] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2

[9] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Aligning large multi-modal model with robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023. 2

[10] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 1, 2

[12] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 2

[13] OpenAI. ChatGPT. https://openai.com/blog/chatgpt/, 2023. 1

[14] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1

[15] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. *arXiv preprint arXiv:2308.12714*, 2023. 2

[16] Qifan Yu, Juncheng Li, Wentao Ye, Siliang Tang, and Yueting Zhuang. Interactive data synthesis for systematic vision adaptation via llms-aigcs collaboration. *arXiv preprint arXiv:2305.12799*, 2023. 1

[17] Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. Alignscore: Evaluating factual consistency with a unified alignment function. *arXiv preprint arXiv:2305.16739*, 2023. 1

[18] Wenqiao Zhang, Haochen Shi, Jiannan Guo, Shengyu Zhang, Qingpeng Cai, Juncheng Li, Sihui Luo, and Yueting Zhuang. Magic: Multimodal relational graph adversarial inference for diverse and unpaired text-based image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3335–3343, 2022. 2

[19] Wenqiao Zhang, Changshuo Liu, Lingze Zeng, Bengchin Ooi, Siliang Tang, and Yueting Zhuang. Learning in imperfect environment: Multi-label classification with long-tailed distribution and partial labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1423–1432, 2023. 1

[20] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 2

[21] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2