What Makes for Good Visual Instructions? Synthesizing Complex Visual Reasoning Instructions for Visual Instruction Tuning

Yifan Du[1*], **Hangyu Guo**[1*], **Kun Zhou**[2*], **Wayne Xin Zhao**[1†], **Jinpeng Wang**[3], **Chuyuan Wang**[3], **Mingchen Cai**[3], **Ruihua Song**[1], **Ji-Rong Wen**[1]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] School of Information, Renmin University of China [3] Meituan Group
{yifandu1999, hyguo0220, batmanfly}@gmail.com, francis_kun_zhou@163.com

Abstract

Visual instruction tuning is crucial for enhancing the zero-shot generalization capability of Multi-modal Large Language Models (MLLMs). In this paper, we aim to investigate a fundamental question: "*what makes for good visual instructions*". Through a comprehensive empirical study, we find that instructions focusing on complex visual reasoning tasks are particularly effective in improving the performance of MLLMs, with results correlating to instruction complexity. Based on this insight, we develop a systematic approach to automatically create high-quality complex visual reasoning instructions. Our approach employs a *synthesize-complicate-reformulate* paradigm, leveraging multiple stages to gradually increase the complexity of the instructions while guaranteeing quality. Based on this approach, we create the **ComVint** dataset with 32K examples, and fine-tune four MLLMs on it. Experimental results consistently demonstrate the enhanced performance of all compared MLLMs, such as a 27.86% and 27.60% improvement for LLaVA on MME-Perception and MME-Cognition, respectively. Our code and data are publicly available at the link: https://github.com/RUCAIBox/ComVint.

1 Introduction

To extend the application scope of Large Language Models (LLMs) (Zhao et al., 2023; Brown et al., 2020) , a surge of work (Liu et al., 2023b; Ye et al., 2023) augments LLMs with vision encoders to endow the ability of multi-modal cognition and reasoning, leading to the emergence of Multi-modal Large Language Models (MLLMs) (Yin et al., 2023; Li et al., 2023c). To achieve good performance, most work first pre-trains the MLLM on a large collection of image-text pairs (*e.g.,* LAION (Schuhmann et al., 2021) and CC (Changpinyo et al., 2021)) to align the text and visual representations, and then fine-tunes it on visual instructions to improve the zero-shot generalization capability (Liu et al., 2023b; Zhang et al., 2023a).

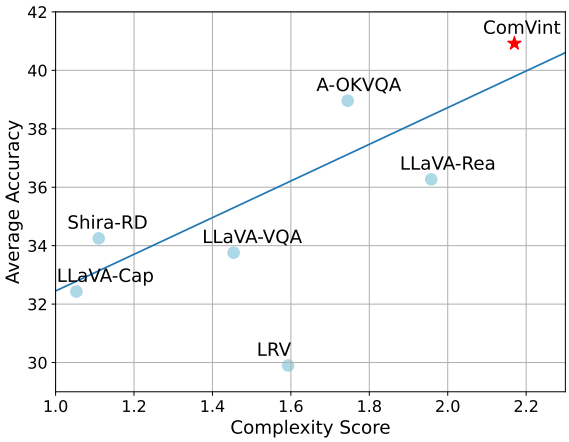

Figure 1: The relation between the complexity of the used instruction set and the average performance of four models on SEED-Bench and MME. The complexity is measured by the reasoning steps counted by ChatGPT.

A visual instruction typically consists of an image, a task description, and a text output (Liu et al., 2023b; Yin et al., 2023). Great efforts have been made to construct high-quality visual instruction datasets, including collecting existing datasets (Li et al., 2023e,a) or synthesis via LLMs (Liu et al., 2023a; Chen et al., 2023b). Despite the prosperity, there is still a lack of a systematic comparison of instruction sets in terms of effectiveness, *based on the same settings* of the backbone model and training strategies. Thus, it remains unclear which instruction sets are more effective and what factors contribute to good instruction data.

Considering this issue, we would like to investigate a more fundamental question, *i.e., "what makes for good visual instructions"*. For this purpose, we first conduct a comprehensive evaluation of existing visual instruction sets, aiming to iden-

[*] Equal Contribution.
[†] Corresponding Author

Proceedings of the 31st International Conference on Computational Linguistics, pages 8197–8214
January 19–24, 2025.

tify the key factors that contribute to effective instructions for MLLMs. Specifically, based on six representative instruction datasets and two popular MLLM, we mainly consider examining two important aspects, namely task types and instruction characteristics. According to the empirical study, we have two main findings:

- The visual reasoning task is more helpful in boosting the performance than image captioning and visual question answering tasks.
- Increasing the instruction complexity is more helpful to improve the performance, than enhancing instruction diversity and integrating fine-grained spatial information.

Additionally, as shown in Figure 1, as the complexity of visual instruction datasets increases, the average benchmark performance is also consistently improved, following an approximate linear trend (details in Section 5). All the above results motivate us to construct complex visual reasoning instructions to enhance MLLMs. However, it is hard to directly prompt GPT-4 (OpenAI, 2023) for synthesizing sufficiently complex and non-hallucination visual instructions (Li et al., 2023f). To address this, we developed a systematic multi-stage pipeline to gradually enhance the quality and complexity of the generated instructions. Concretely, our approach adopts a *synthesize-complicate-reformulate* pipeline to generate the instruction, where corresponding prompts are devised to guide GPT-4. In the complication stage, we guide GPT-4 to fully utilize both image content and outside knowledge[1] to improve the complexity, and iteratively verify the accuracy of the instruction to ensure the quality. Finally, we reformulate it into multiple formats for better adaptation to various downstream tasks.

Using the above approach, we synthesize a **Com**plex **V**isual reasoning **ins**truction dataset, namely **ComVint**, consisting of 32K examples, and fine-tune four representative MLLMs (*i.e.,* BLIP-2, LLaVA, MiniGPT-4, and InstructBLIP) on it. Evaluation results on two comprehensive benchmarks, SEED-Bench (Li et al., 2023b) and MME (Fu et al., 2023), demonstrate that our instruction dataset significantly enhances the performance of these MLLMs, outperforming existing visual instruction collections. For instance, leveraging our dataset leads to a remarkable improvement of 27.86% and 27.60% in the performance of LLaVA on MME-Perception and MME-Cognition, respectively.

[1]According to (Marino et al., 2019), outside knowledge refers to the knowledge that is not provided by the image, *e.g.,* inferring the latitude of a location in the image.

2 Background

Multi-modal Large Language Models. Multi-modal Large Language Models (MLLMs) (Li et al., 2023c) are advanced generative models capable of processing information from various modalities (*e.g.,* image, video, and audio) and generating corresponding textual responses. This work focuses on MLLMs in the visual modality, typically consisting of an image encoder, an LLM, and a connection module. The image is first encoded into patch embeddings by the image encoder and the connection module, then concatenated with text embeddings, enabling the LLM to comprehend the image and generate the response auto-regressively. MLLMs undergo vision-language pre-training to align the vision encoder and LLM, followed by visual instruction tuning to enhance instruction following and understanding ability (Liu et al., 2023b; Zhang et al., 2023b).

Visual Instruction Tuning. Instruction tuning (Wei et al., 2022; Chung et al., 2022) is important to improve the ability of LLMs in instruction following and generalization on unseen tasks (Longpre et al., 2023; Wang et al., 2023). It employs a text-formatted task description and the expected outcome to fine-tune LLMs in a supervised way. Inspired by the success of LLMs, instruction tuning has been adapted to develop MLLMs for visual tasks, termed *visual instruction tuning* (Zhu et al., 2023; Liu et al., 2023b). Typically, a visual instruction comprises an image X_I, a textual task instruction X_T, and a corresponding output text Y_T. During training, MLLMs learn to generate Y_T conditioned on X_I and X_T.

2.1 Visual Instruction Collections

To get a sense of what constitutes good visual instructions, we review and categorize existing collections in Table 1. Broadly, visual instructions are crafted to address specific tasks and incorporate various considerations (*e.g.,* diversity and complexity). Thus, we discuss prior efforts in two major aspects: task types and instruction characteristics.

Task Types. Most visual instruction datasets (Liu et al., 2023b; Li et al., 2023e) are derived from

existing multi-modal datasets and primarily focus on three types of tasks:

- *Image captioning*: it requires the model to generate a free-form description of an image.
- *Visual question answering (VQA)*: it requires the model to answer a question about the image, *e.g.,* counting the objects and recognizing the color.
- *Visual reasoning*: it requires the model to perform reasoning based on the image context, *e.g.,* conjecturing the relationship between two objects, and answering questions involving commonsense reasoning.

We examine the effects of task type using the LLaVA-Instruct (Liu et al., 2023b), which includes 23K image captions, 58K conversations, and 77K visual reasoning questions. We divide it into three subsets corresponding to the three task types, namely LLaVA-Caption, LLaVA-VQA, and LLaVA-Reasoning, respectively.

Instruction Characteristics. In addition to the task types, recent studies (Liu et al., 2023a; Chen et al., 2023b; Zhang et al., 2023c; Chen et al., 2023a) also attempt to endow visual instruction collection with special characteristics to further improve the performance of MLLMs.

- *Task diversity:* existing work (Wei et al., 2022; Liu et al., 2023a) has found that increasing the task diversity can improve the zero-shot ability for task solving. This can typically be achieved by aggregating instructions from different tasks.
- *Instruction complexity:* enhancing instruction complexity is a widely used strategy to improve the performance of LLMs (Xu et al., 2023), and we can also utilize complex multi-modal tasks (*e.g.,* multi-hop cross-modal reasoning) to improve the performance of MLLMs.
- *Fine-grained spatial information:* it is important for MLLMs to recognize fine-grained spatial details of objects in an image. For this purpose, the spatial coordinates annotations can be included in the textual instructions (Chen et al., 2023b,a).

To study the effect of these characteristics, we select LRV (Liu et al., 2023a), A-OKVQA (Schwenk et al., 2022), and Shikra-RD (Chen et al., 2023b), three representative instruction sets with diverse task types, complex outside knowledge, and fine-grained spatial information, respectively.

Instruction	Number	Task Type	Characteristics
LLaVA-Cap	23K	Cap	\
LLaVA-VQA*	256K	VQA	\
LLaVA-Rea	77K	Rea	\
LRV	150K	Cap, VQA, Rea	Diverse
Shikra-RD	4K	Rea	Fine-grained
ComVint (Ours)	32K	Rea	Complex

Table 1: Comparison of existing synthesized visual instruction collections. "Cap" shorts for Caption and "Rea" shorts for Reasoning. *We divide the multi-turn conversation in LLaVA into individual questions, resulting in 256K VQA instructions.

3 Empirical Analysis of Visual Instructions

In this section, we empirically study the effect of different task types and different in visual instruction tuning by fine-tuning two representative MLLMs (*i.e.,* BLIP-2 (Li et al., 2023d) and MiniGPT-4 (Zhu et al., 2023)) on the visual instruction collections selected in Section 2.

3.1 Experiment Setup

Backbone MLLMs. We select two models with minimal or no instruction tuning to clearly study the effect of different factors:

- *BLIP-2*: it incorporates a lightweight querying Transformer to connect a fixed vision encoder and a fixed LLM. It is only pre-trained on large-scale image-text pairs, and has not been fine-tuned with visual instructions.
- *MiniGPT-4*: it employs the similar architecture as BLIP-2 and adopts Vicuna as the LLM. It is first pre-trained on 5M image-text pairs and then fine-tuned on 3,500 image captions.

We follow the training strategies in (Zhu et al., 2023; Li et al., 2023d), fine-tuning the Q-Former in BLIP-2 and the linear layer between the Q-Former and LLM in MiniGPT-4.

Evaluation Benchmark. We select two widely-used benchmarks, *i.e.,* MME (Fu et al., 2023) and SEED-Bench (Li et al., 2023b) for evaluation:

- *MME*: it aims to measure the perception and cognition abilities of MLLMs. Each instance comprises one image and two questions. Following (Fu et al., 2023), we report the accuracy of whether both questions of an instance are answered correctly, denoted as *ACC+*.

Baseline Model		MiniGPT-4				BLIP-2			
	Benchmark	SEED-Bench Image ACC	MME-P ACC+	MME-C ACC+	Average	SEED-Bench Image ACC	MME-P ACC+	MME-C ACC+	Average
	Original	43.31	26.96	10.77	27.01	53.68	41.25	15.38	36.77
A	+LLaVA-Cap	41.60	10.12	1.54	17.75	52.31	28.38	11.54	30.74
A	+LLaVA-VQA	45.97	12.87	5.38	21.41	51.00	31.88	14.62	32.50
A	+LLaVA-Rea	43.69	**40.59**	16.15	33.48	50.94	43.33	20.77	38.35
B	+LRV	**50.92**	3.12	0.77	18.27	**54.64**	10.88	5.38	23.63
B	+Shikra-RD	41.75	8.33	1.54	17.21	52.92	36.05	16.15	35.04
B	+A-OKVQA	43.99	36.71	**21.54**	**34.08**	52.60	**46.83**	**24.62**	**41.35**

Table 2: The results on SEED-Bench and MME after fine-tuning MiniGPT-4 and BLIP-2 using different instruction collections. MME-P and MME-C short for MME-Perception and MME-Cognition, respectively. Cells shaded in orange indicate that fine-tuning enhances the performance, while blue indicates performance degradation.

- *SEED-Bench*: it contains 12 tasks to evaluate the image and video understanding capacity of MLLMs. Since most MLLMs do not consider video understanding ability, we only evaluate the models on image understanding tasks. Following (Zhang et al., 2023a), we employ accuracy in image understanding tasks as the evaluation metric and denote it as *Image ACC*.

3.2 Results and Analysis

We categorize the results into two groups based on the task types and special characteristics of the instructions: (A) includes LLaVA-Cap (image captioning), LLaVA-VQA (visual question answering), and LLaVA-Rea (visual reasoning); (B) includes LRV (diversity), A-OKVQA (complexity), and Shikra-RD (spatial annotation). We can obtain the following insights from the results in Table 2.

Finding 1: among task types in group A, the visual reasoning task yields the best performance compared to image captioning and visual question answering. Concretely, fine-tuning MLLMs on the LLaVA-Cap leads to a noticeable performance degradation across all benchmarks, while LLaVA-Rea results in significant improvements, with LLaVA-VQA results showing intermediate results in most cases. This suggests that the performance advantage may correlate positively with the difficulty of visual instructions.

Finding 2: for instruction characteristics in group B, complexity is more important than task diversity and fine-grained information. Concretely, MLLMs fine-tuned on A-OKVQA achieve the highest accuracy across all benchmarks, emphasizing the pivotal role of complex instructions in boosting model performance. In contrast, LRV and Shikra-RD show minimal or negative impacts, indicating limited benefits from increased task diversity or spatial details.

To summarize, reasoning-oriented (LLaVA-Rea) and complexity-enhanced (A-OKVQA) instruction sets are particularly useful in improving the performance of MLLMs in our experiments. However, LLaVA-Rea still has limited complexity. Though A-OKVQA contains complex task instructions, it is mainly constructed by human annotators, and the instruction complexity is limited to the annotator's abilities. Therefore, it is desirable to develop automatic approaches to produce complex visual reasoning instructions at scale.

4 Approach

Based on the findings in Section 3.2, we propose a method to create high-quality, complex visual reasoning instructions to improve the performance of MLLMs. For image X_I with captions $\{C_I\}$ and objects $\{O_I\}$, we first synthesize two common kinds of visual reasoning instructions, *i.e.,* cross-modal reasoning instruction $\langle X_T^{(C)}, Y_T^{(C)} \rangle$ and outside-knowledge reasoning instruction $\langle X_T^{(K)}, Y_T^{(K)} \rangle$. Subsequently, we use an iterative complicate-then-verify procedure to gradually improve their complexity and quality, obtaining $\langle \tilde{X}_T^{(C)}, \tilde{Y}_T^{(C)} \rangle$ and $\langle \tilde{X}_T^{(K)}, \tilde{Y}_T^{(K)} \rangle$. Finally, we merge and reformulate these instructions to increase data format diversity, resulting in the final dataset.

4.1 Visual Reasoning Instruction Synthesis

Cross-modal reasoning (Hudson and Manning, 2019) and outside-knowledge reasoning (Marino et al., 2019) are key visual reasoning tasks that focus on image content and external knowledge, respectively. We synthesize instructions for both

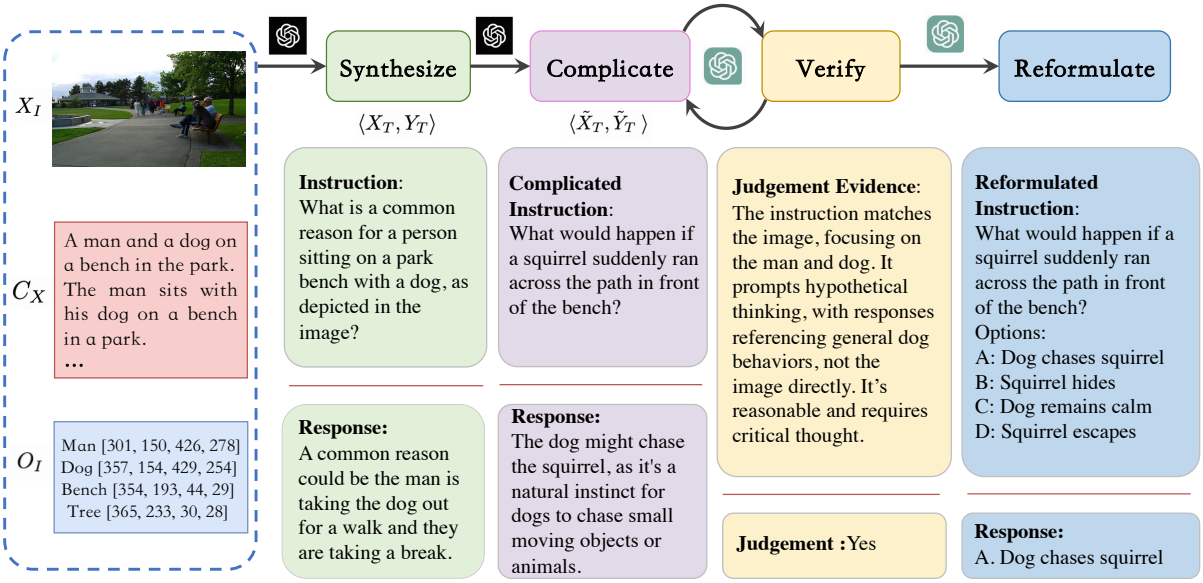

Figure 2: Our approach to synthesizing complex visual reasoning instructions involves three stages: synthesizing the primary reasoning instructions, iterative complication and verification, and reformulating instruction formats.

types to capture their essential information, resulting in the primary cross-modal and outside-knowledge reasoning instructions.

4.1.1 Cross-modal Reasoning Instructions.

We aim to synthesize cross-modal reasoning instruction that requires MLLMs to accurately map text entities to image objects and describe object relationships in natural language. To achieve this, we first select images that contain rich objects and then utilize GPT-4 to generate the instruction based on the image annotations.

Image Selection. To ensure the instruction complexity, we select images containing diverse objects and relationships from the Flickr30k Entities dataset (Plummer et al., 2015). Each image has five detailed captions linking objects with entities and coordinates. Empirically, we find that more informative images usually come with more detailed captions, so we consider the total character count in the five captions for each image as the indicator of informativeness. We filter out those with fewer than 700 characters in the caption and retain only the most informative ones for instruction synthesis.

Instruction Generation. After selecting informative images and their associated captions and objects, *i.e.*, $\{\langle X_I, C_I, O_I \rangle\}$, we employ GPT-4 to generate cross-modal reasoning instructions $\{\langle X_I, X_T^{(C)}, Y_T^{(C)} \rangle\}$ as follows:

$$X_T^{(C)}, Y_T^{(C)} = \text{GPT-4}(P^C, C_I, O_I) \quad (1)$$

where P^C is the prompt for cross-modal reasoning instructions. We carefully design the prompt to instruct GPT-4 to synthesize three instructions simultaneously, while guaranteeing instruction diversity. By incorporating specific requirements and in-context demonstrations into P^C, we can reduce the probability of synthesizing instructions that are too simple or contain irrelevant information. The prompt is shown in Figure 5 of the Appendix.

4.1.2 Outside-knowledge Reasoning Instruction.

In addition to understanding the visual semantics of an image, MLLMs also require world knowledge and common sense to help understand complex relationships in complex tasks. Following the process in Section 4.1.1, we employ image selection and instruction generation for synthesizing outside-knowledge reasoning instructions.

Image Selection. To synthesize outside-knowledge reasoning instructions, we require detailed object information from images (*e.g.,* the brand of a T-shirt) to capture outside knowledge (*e.g.,* the price of the T-shirt). We select Visual Genome (Krishna et al., 2017) as the image source, which provides an average annotation of 21 objects per image, each with a corresponding caption. However, if an image contains an excessive number of object annotations, these annotations often lack detailed information about individual objects, which is essential for the generation of

outside-knowledge reasoning instructions. Hence, we set a threshold (*e.g.,* 7) and remove images exceeding this limit to ensure quality.

Instruction Generation. To generate high-quality outside-knowledge reasoning instructions, we first select suitable objects as topic entities from the chosen images, and then prompt GPT-4 to synthesize instructions about them. For topic entity selection, we mainly consider long-tail world knowledge that MLLMs may overlook. Specifically, we utilize *Inverse Document Frequency* (IDF) to measure the importance of a certain object. We select the object with the highest IDF and denote it as $O_I^{'}$. Based on the topic entity and image annotation, we utilize GPT-4 to generate the instructions $\{\langle X_I, X_T^{(K)}, Y_T^{(K)}\rangle\}$ as:

$$X_T^{(K)}, Y_T^{(K)} = \text{GPT-4}(P^K, C_I, O_I^{'}) \quad (2)$$

where P^K is the prompt for outside-knowledge instructions. Additionally, we sample knowledge categories from the category set in OK-VQA (Marino et al., 2019) and incorporate them into P^K, guiding GPT-4 to produce instructions related to these categories. This ensures balanced knowledge coverage across the instruction dataset. The detailed prompt is shown in Figure 6 of the Appendix.

4.2 Visual Reasoning Instruction Complication

Through this instruction synthesis process, we obtain two types of instruction sets. Despite the carefully designed prompts, the synthetic instructions are still relatively simple and even contain hallucinated objects. To address this, we propose an iterative *complicate-then-verify* procedure to gradually increase the complexity of the instructions and meanwhile ensure the quality and avoid contradictions or hallucinations.

Instruction Complication. Inspired by existing work (Xu et al., 2023), we instruct GPT-4 to iteratively complicate the instructions and generate the corresponding response, based on the primary instructions and image annotations, denoted as:

$$\tilde{X}_T, \tilde{Y}_T = \text{GPT-4}(P^{Comp}, X_T, Y_T, C_I, O_I) \quad (3)$$

where P^{Comp} is the prompt sent to GPT-4, shown in Figure 7 in the Appendix. Empirically, we find that very few iteration turns (*e.g.,* 1 or 2) are sufficient to obtain high-quality instructions, thereby reducing the cost of APIs invocation.

Instruction Verification. To ensure instruction quality, we use a verification process to filter out instructions that contradict the image. Specifically, we prompt ChatGPT to determine if the synthesized instruction aligns with the provided image annotations. The prompt is shown in Figure 8 in the Appendix. Based on the judgment of ChatGPT, we only retain the instructions that pass the verification and discard the failed ones.

4.3 Visual Reasoning Instruction Reformulation

After the synthesis and complication processes, we obtain many high-quality, complex instructions. However, these open-ended responses may not suit tasks requiring specific formats (*e.g.,* multiple-choice or boolean QA), potentially affecting zero-shot generalization.

To address this, we incorporate a reformulation stage, in which we sample some synthetic instructions and use ChatGPT to convert them into two distinct representative formats: boolean QA and multiple-choice QA. Boolean QA offers binary answers, *i.e.,* "yes" or "no", while multiple-choice QA provides several predefined options. After the reformulation stage, we combine the original open-ended instructions with the newly reformulated instructions to create the final **ComVint** dataset.

To evaluate the quality of the data synthesized by GPT-4, we randomly sample some instances for human review. The results in Table 5 in the Appendix show that most instructions are of high quality. We also compare cases in LLaVA-Reasoning, LRV, and our ComVint in Figure 4 in the Appendix, and find that instructions in ComVint are more complex and involve more reasoning steps.

5 Experiment

5.1 Experimental Setup

To exhibit the generality of our instructions, we fine-tune four representative MLLMs on ComVint: BLIP-2, MiniGPT-4, LLaVA, and InstructBLIP. These models were selected because they employ diverse architectures, training strategies, and training datasets, making them ideal for verifying the generalizability of our conclusions. A detailed introduction to these models is in the Appendix. We fine-tune these models on our instruction dataset and the other six representative visual instruction collections (LLaVA Caption, LLaVA VQA, LLaVA Reasoning, LRV, Shikra-RD, and A-OKVQA) used

Model	Benchmark	Original	+LLaVA Cap	+LLaVA VQA	+LLaVA Rea	+LRV	+Shikra -RD	+A-OKVQA	+ComVint
MiniGPT4	SEED-Bench	43.31	41.60	45.97	43.69	**50.92**	41.75	43.99	50.13
	MME-P	820.37	643.59	650.51	929.82	552.75	722.38	924.47	856.30
	MME-C	198.57	154.29	218.21	265.71	199.29	172.14	257.50	227.14
BLIP-2	SEED-Bench	53.68	52.31	51.00	50.94	**54.64**	52.92	52.60	53.73
	MME-P	1151.26	1115.52	907.21	1162.47	643.02	1102.43	1115.69	**1216.24**
	MME-C	241.07	224.64	238.93	**271.07**	216.43	231.43	247.14	250.71
LLaVA	SEED-Bench	49.43	48.52	46.86	46.82	**56.58**	49.01	54.01	54.74
	MME-P	949.42	1091.41	1002.42	1043.53	1154.99	915.64	1140.23	**1213.87**
	MME-C	232.86	238.21	257.14	211.07	272.14	260.00	**310.36**	297.14
InstructBLIP	SEED-Bench	56.14	55.44	52.63	55.22	56.57	**57.49**	56.12	**57.49**
	MME-P	1178.95	1211.72	972.29	1205.13	918.15	1176.63	**1262.29**	1199.91
	MME-C	301.79	**340.00**	302.86	287.14	219.64	279.29	277.50	305.36

Table 3: The performance of four representative MLLMs fine-tuned on different instruction collections. The best performance and the second-best performance are denoted in bold and underlined fonts, respectively.

in Section 3 for comparison.

Implementation Details. For BLIP-2, MiniGPT-4, and InstructBLIP, we set the learning rate to 1e-5 and train for 2 epochs, while for LLaVA, we set the learning rate to 2e-5 and train for 3 epochs. The batch size for BLIP-2, LLaVA, and InstructBLIP is 128, while the batch size for MiniGPT-4 is 64. Concerning the mixture of our synthesized instructions, the final instruction collection comprises approximately 12K cross-modal reasoning instructions and 20K outside-knowledge reasoning instructions.

Evaluation Benchmarks. We follow Section 3 and mainly evaluate the models on SEED-Bench Image and MME. Additionally, we also incorporate three traditional visual reasoning benchmarks, *i.e.,* OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019).

5.2 Main Results

The results of the four models fine-tuned on seven instruction collections are shown in Table 3. Based on the results, we have the following findings:

Firstly, among all the existing visual instruction collections, complex visual reasoning instructions (*i.e.,* A-OKVQA and LLaVA Reasoning) generally lead to the most substantial improvements across the three evaluation dimensions compared to others. The enhancements in InstructBLIP and LLaVA are not as significant, as they had already utilized these data during instruction tuning.

Secondly, baseline instruction datasets can enhance model performance on specific benchmarks. For instance, MiniGPT-4, BLIP-2, and LLaVA fine-tuned on LRV outperform those fine-tuned on ComVint when evaluated on the SEED-Bench benchmark. This is because LRV is specially designed with valuable features, such as various task types and formats, which can be beneficial for improving MLLMs in related tasks. However, introducing certain characteristics can degrade performance on unrelated tasks. For example, BLIP-2 fine-tuned on LRV shows significant drops in MME benchmark (MME-Perception decreases from 1151.26 to 643.02, and MME-Cognition decreases from 241.07 to 216.43). In comparison, our ComVint effectively balances various capabilities and yields improvements across all benchmarks. We also display the results of LLaVA on all the sub-tasks of SEED-Bench in Table 7 in the Appendix. We can observe that ComVint significantly improves the performance on all the sub-tasks.

Instruction	OK-VQA	A-OKVQA	GQA
Original	57.82	77.12	50.98
+①: LLaVA-Rea	46.54	72.93	46.88
+②: A-OKVQA	57.48	78.86*	50.90
+③: ComVint	**58.71**	**78.08**	**51.75**

Table 4: Results of InstructBLIP on traditional VQA evaluation benchmarks. The result marked with * denotes that the model is fine-tuned on the training set of the evaluation benchmark.

Thirdly, since InstructBLIP achieves the best performance on these benchmarks, we also evaluate it on three traditional VQA benchmarks: OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). The results in Table 4 show that all previ-

	LRV	LLaVA-Rea	ComVint
Instruction	88.00	90.00	88.00
Response	65.00	86.00	84.00

Table 5: The correct rate of each instruction dataset evaluated by humans.

Instruction	SEED-Bench	MME-P	MME-C	Avg
Original	49.43	39.36	12.31	33.70
+ComVint	**54.74**	**55.53**	26.15	**45.47**
w/o Comp	50.89	44.94	12.31	36.05
w/o O-K Rea	51.40	49.29	24.62	41.77
w/o C-M Rea	53.85	46.74	**27.69**	42.76
w/o Reform	53.75	45.60	19.23	39.53

Table 6: Results of ablation study. "Comp" stands for complication, "C-M Rea" for cross-modal reasoning, "O-K Rea" for outside-knowledge reasoning, and "Reform" for reformulation.

ous instruction datasets hurt performance on these benchmarks, while ComVint is the only instruction dataset that can further boost performance.

Besides the quantitative results, we present case examples from ComVint in Figure 3 and Figure 4, comparing them to previous instruction sets. We can observe that LLaVA-Reasoning primarily focuses on scene descriptions, while LRV emphasizes object recognition. In contrast, ComVint is more complex and includes more reasoning steps.

5.3 In-Depth Analysis

Data Quality Analysis. We randomly sample 100 instances from ComVint, LRV, and LLaVA-Reasoning and have three of our authors manually assess their quality. Following prior work (Liu et al., 2023a), we separately check whether the instructions and responses are correct. The evaluation criteria for correctness are in the Appendix. We calculate the Fleiss' Kappa among the three annotators, obtaining a value of 0.91, which indicates near-perfect agreement. As shown in Table 5, most instructions are accurate. Besides, the responses in ComVint exhibit comparable quality to LLaVA-Reasoning, and superior to LRV, highlighting the high quality of our dataset.

The Effect of Complication. For each instruction set, we ask gpt-3.5-turbo-0125 to count the reasoning steps in each instruction, defining the average reasoning steps as the complexity score of the dataset. We then plot the relationship between complexity scores and the average accuracy of four models on these benchmarks. The result in Figure 1 shows that as the complexity of the visual instruction datasets increases, the average performance consistently improves, following a roughly linear trend. Notably, ComVint, with the most complex instruction set, achieves the best results. Please refer to the Appendix C for more details. Meanwhile, to test the effectiveness of the complication operation in our pipeline, we create a basic instruction set by removing all complexity constraints and skipping the complication stage (denoted as w/o Comp in Table 6). We fine-tune LLaVA on ComVint and this basic set. The results in Table 6 show significant degradation on all benchmarks, demonstrating the importance of instruction complexity.

The Effect of Two Types of Instructions. In our reasoning instruction synthesis stage, we generate cross-modal and outside-knowledge reasoning instructions. To assess their effectiveness, we remove either of them and fine-tune LLaVA on the remaining instructions (denoted as w/o O-K Rea and w/o C-M Rea in Table 6). The results show that removing either type hurts performance on SEED-Bench and MME-Perception. As for the MME-Cognition tasks, removing cross-modal reasoning instructions yields the best performance. This is because answering questions in MME-Cognition most rely on cognition ability, thus our outside knowledge reasoning instructions are more beneficial. Generally, incorporating both types of instruction leads to the best average performance.

The Effect of Instruction Format. We remove the reformulation module to obtain 32K open-ended instructions (denoted as w/o Reform) and fine-tune LLaVA on them, with the results presented in Table 6. We can observe that removing the reformulation module would lead to performance degradation compared to the instructions with diverse formats, highlighting the importance of the reformulation module. On the other hand, fine-tuning the model only on the open-ended instructions also significantly improves the performance, implying that the reformulation module is not the only factor contributing to the improvement.

5.4 Case Study

In this section, we display some instructions in our ComVint dataset in Figure 3. The first instruction requires the model to compare the painting of the woman and the painting on the wall, and then under-

[Instruction]
Is the woman replicating the work of another artist in her painting?

[Response]
Yes, the woman is painting a picture same as the one in the middle on the wall, indicating that she is replicating the work of another artist.

[Instruction]
Is it the man in the blue suit that is making the woman with the long brown hair laugh?

[Response]
No, the clown is making the woman laugh, and the man in blue suit is accompanying her.

[Instruction]
Are there three people, including a man and a woman, looking at the sculptures ?

[Response]
No, there is only one man looking at the sculptures, while another man and a woman are talking.

Figure 3: Examples from the ComVint dataset, highlighting instructions designed to emphasize cross-modal reasoning over basic visual perception.

stand that the woman is replicating the painting on the wall. The second instruction requires the model to understand the relationships among these people. The third instruction requires the model to observe the image carefully and understand the activity of these people. We also display several cases from LLaVA-Reasoning, LRV, and ComVint in Figure 4 for comparison. Please refer to Appendix A for more details.

6 Conclusion

In this paper, we investigated key factors contributing to effective visual instructions for MLLMs. Our empirical experiments demonstrated that instructions with the visual reasoning task type and complexity characteristic were more useful for improving the capabilities of MLLMs. Based on these insights, we devised a systematic approach to automatically create high-quality complex visual reasoning instructions, employing a synthesize-complicate-reformulate paradigm to gradually improve the complexity while guaranteeing quality. Using this approach, we synthesized a visual reasoning instruction dataset, namely **ComVint**, for fine-tuning MLLMs. Experimental results have demonstrated the efficacy of our dataset in improving the capability of representative MLLMs.

7 Limitation

First, many factors influence the final performance of an MLLM: training data, model architecture, and training strategies, *etc*. In this work, we focus solely on training data, identifying what makes for good visual instructions regarding task type and instruction characteristics, while leaving other factors unchanged. In the future, we plan to explore how these factors, in coordination with training data, affect the overall performance of MLLMs. Second, the scale of ComVint is not large enough. While we achieve strong results with limited data, demonstrating the value of "less is more", we believe that scaling up the amount of high-quality visual instruction data will further enhance performance.

Acknowledgements

This work was partially supported by National Natural Science Foundation of China under Grant No. 92470205 and 62222215. This research was also supported by Meituan and the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin University of China. Xin Zhao is the corresponding author.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish,

Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3558–3568. Computer Vision Foundation / IEEE.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. 2023a. Position-enhanced visual instruction tuning for multimodal large language models. *CoRR*, abs/2308.13437.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. MIMIC-IT: multi-modal in-context instruction tuning. *CoRR*, abs/2306.05425.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125.

Chunyuan Li, Zhe Gan, Zhengyuan Yang, Jianwei Yang, Linjie Li, Lijuan Wang, and Jianfeng Gao. 2023c. Multimodal foundation models: From specialists to general-purpose assistants. *CoRR*, abs/2309.10020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023d. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Lei Li, Yuwei Yin, Shicheng Li, Liang Chen, Peiyi Wang, Shuhuai Ren, Mukai Li, Yazheng Yang, Jingjing Xu, Xu Sun, Lingpeng Kong, and Qi Liu. 2023e. M^3it: A large-scale dataset towards multi-modal multilingual instruction tuning. *CoRR*, abs/2306.04387.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023f. Evaluating object hallucination in large vision-language models. *CoRR*, abs/2305.10355.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *CoRR*, abs/2306.14565.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *CoRR*, abs/2304.08485.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The flan collection: Designing data and methods for effective instruction tuning. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 22631–22648. PMLR.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2641–2649. IEEE Computer Society.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. *CoRR*, abs/2111.02114.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. How far can camels go? exploring the state of instruction tuning on open resources. *CoRR*, abs/2306.04751.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*, abs/2304.12244.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *CoRR*, abs/2304.14178.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023a. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *CoRR*, abs/2309.15112.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023b. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *CoRR*, abs/2309.15112.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. 2023c. Gpt4roi: Instruction tuning large language model on region-of-interest. *CoRR*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A survey of large language models. abs/2303.18223.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Case Study

We also display several cases from LLaVA-Reasoning, LRV, and ComVint in Figure 4 for comparison. The instruction in LLaVA-Reasoning mainly focuses on the description of the scene, which can be answered easily once the model recognizes the helmet in the image. The instruction in LRV is relatively shorter and simpler, focusing on the object recognition ability of the model. In contrast, the instruction in ComVint is more complex and includes more reasoning steps. To answer the question in the image, the model needs to first identify the three women according to the color of their clothes. Then, the model needs to identify their actions and finally, distinguish the commonalities and differences between them.

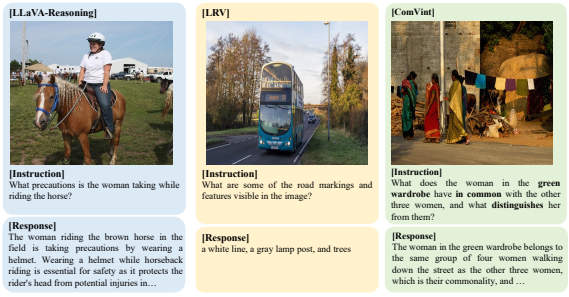

Figure 4: Sampled instructions from LLaVA-Reasoning, LRV, and ComVint, where the instruction in ComVint is more complex than LLaVA-Reasoning and LRV.

B Baseline Models

In this section, we elaborate on the details of our baseline models.

- *BLIP-2* (Li et al., 2023d) is based on FLAN-T5-XXL and utilizes a Q-Former to connect the vision encoder and the LLM. It freezes the vision encoder and the LLM during fine-tuning and only updates the parameters of the Q-Former. The training data contains 129M image-caption pairs.
- *MiniGPT-4* (Zhu et al., 2023) is based on Vicuna-7B and reuses the Q-Former in BLIP-2 to connect the vision encoder and the LLM. It freezes the Q-Former, the vision encoder, and the LLM during fine-tuning and only updates the parameters of the linear layer between the Q-Former and the LLM. The training data includes 5M image-text pairs and 3,500 high-quality image-caption pairs.
- *LLaVA* (Liu et al., 2023b) is based on Vicuna-7B and utilizes linear layers to connect the vision encoder and the LLM. It freezes the vision encoder during fine-tuning and updates the linear layers and the LLM. The training data is 158K visual instructions, including 58K in conversations, 23K in detailed descriptions, and 77K in visual reasoning.
- *InstructBLIP* (Dai et al., 2023) is based on FLAN-T5-XXL and the architecture and the trainable parameters are the same as BLIP-2. The training data includes 10 vision-language tasks as well as the instruction data used in LLaVA.

C Discussion

To analyze the impact of instruction complexity on the performance of the MLLM, we randomly sample 1000 instructions from each instruction set listed in Table 3. Utilizing the gpt-3.5-turbo-0125 model, we parse these instructions into a series of sub-questions, with each sub-question representing a reasoning step. The complexity of a question is measured by the number of reasoning steps, and we calculate the average number of reasoning steps for each instruction set to derive its complexity score. Subsequently, we average the results for the MME and SEED-Bench benchmarks as presented in Table 3. Further, we average the results across the four MLLMs to obtain the final average accuracy. The results depicted in Table 1 reveal that, overall, an increase in instruction complexity corresponds to an improvement in MLLM performance, demonstrating a trend that can be approximated by a linear function. An exception is observed with LRV, possibly due to its inclusion of more incorrect instructions, as indicated in Table 5. Moreover, LLaVA-Cap, characterized by the lowest complexity, yields the lowest accuracy, while LLaVA-Rea and A-OKVQA, with higher complexity, exhibit better performance, aligning with findings from our empirical study in Table 3. Notably, ComVint, with the most complex instruction set, achieves the best results.

D More Experimental Results

We display the accuracy of LLaVA on all the sub-tasks on the SEED-Bench Image in Table 7. The results show that ComVint can improve the performance of LLaVA on all the sub-tasks on SEED-Bench Image, especially on instance attributes, instance interaction, visual reasoning, and text recognition. This demonstrates the effectiveness of our instruction dataset.

Inst	instance counting	instance attributes	scene understanding	instance identity	instance interaction	visual reasoning	instance location	spatial relations	text recognition	Average
Original	38.95	51.90	57.09	53.69	47.42	47.43	41.21	38.36	30.59	49.43
+ComVint	**42.54**	**59.65**	**62.13**	**56.96**	**57.73**	**53.47**	**44.38**	**40.79**	**43.53**	**54.74**

Table 7: The performance of LLaVA on all the sub-tasks of SEED-Bench Image.

E Prompt Design

We display the prompt used to synthesize visual instructions. The prompt for instruction synthesizing is in Figure 5 and Figure 6, the prompt for complication is in Figure 7, the prompt for verification is in Figure 8, and the prompt for reformulation is in Figure 9 and Figure 10.

F Data Quality Evaluation Criteria

In Section 5.3, we have three authors to assess the quality of the visual instructions manually. Specifically, the evaluation criteria for correctness are as follows:

Instructions: (1) necessitate visual information from the image to response (2) are clear and interpretable, and (3) align with the image context without ambiguity.

Responses: (1) contain no hallucinations or contradictions to the image, (2) accurately follow the instruction, and (3) are factually and logically coherent.

System prompt:

Here are 5 captions for an image, some entities in the caption are followed by "[x1, y1, x2, y2]" to indicate the bounding box coordinates of the entity in the image. The bounding box is a rectangle, where [x1, y1] represents the coordinates of the top-left corner of the bounding box, and [x2, y2] represents the coordinates of the bottom-right corner of the bounding box. Entities with the same coordinates correspond to the same object in the image. Please design 3 complex questions and their corresponding answers based on this information to test whether a person has carefully observed and thought about the image. These 3 questions and answers must meet the following conditions:

1. Do NOT mention anything that is not mentioned in the 5 captions.
2. The questions must be definitely answered according to the 5 provided captions.
3. The questions should involve details in the image and use as many entities as possible.
4. Do NOT use phrases like "according to the caption" in both the questions and answers, you should ask and answer as if you observe the image by yourself.
5. The questions and answers should be as diverse as possible.

Some useful tips: you can swap some entities to confuse the test-taker, which will make the questions more challenging. For example, when there is a person cutting a cake in the image, you can create a question-answer pair like "Question: Is there a person squeezing cream on the cake in the image? Answer: No, the person is cutting the cake instead of squeezing cream." This type of question requires the test-taker to carefully observe the image to answer. Do not ask simple questions like "What color is the cake in the image?" because it is too simple.

Demonstration:

user:

Captions:\n1. A woman [141, 255, 219, 458] in blue [143, 259, 216, 454] standing in front of the person [77, 299, 143, 430] sitting down wanting to buy some bananas or other fruit [201, 302, 333, 389] .\n
2. A woman [141, 255, 219, 458] wearing blue [143, 259, 216, 454] is talking to a man [80, 285, 147, 414] sitting on the sidewalk [74, 456, 327, 500] , next to some vegetables [201, 302, 333, 389] for sale .\n
3. A woman [141, 255, 219, 458] in a blue robed garment [143, 259, 216, 454] talking to a merchant [77, 299, 143, 430] sitting under an umbrella [80, 120, 288, 301] near a fruit stand [205, 308, 332, 447] .\n
4. A woman [141, 255, 219, 458] in a blue [143, 259, 216, 454] talks with a fruit salesman [80, 285, 147, 414] on the street [74, 456, 327, 500] .\n
5. A woman [141, 255, 219, 458] talks to a man [80, 285, 147, 414] at the market .

assistant:

Question1: Who is selling the fruits and who is the buyer? Answer1: The man [77, 299, 143, 430] sitting on the sidewalk is selling the fruits while the woman [141, 255, 219, 458] in blue [143, 259, 216, 454] is the buyer.\n
Question2: Is the woman in blue selling fruits to the man? Answer2: No, the man is selling fruits to the woman in blue.\n
Question3: What is the woman in blue buying? Answer3: The woman in blue is buying fruits or vegetables.

Figure 5: The prompt we give to GPT-4 for constructing cross-modal reasoning instructions.

System prompt:
I want you to act as a Prompt Generator. Give an image with the following information: region captions and their corresponding bounding box, positions that are the region left-top corner coordinates (X, Y), region covered sizes (Width, Height). Highly overlapping bounding boxes may refer to the same object.
According to the region description provided above, design some instructions and corresponding responses. Your instructions need to come up with complex knowledge of one or more hops according to the image, but please ensure that the answers to this knowledge are accurate. The knowledge needs to be hard enough that only well-educated people and experts in the relevant field can reply.
The instructions must meet the following conditions:1. Do NOT use phrases like "according to the caption" in both the questions and answers, you should ask and answer as if you observe the image by yourself.\n2. The questions and answers should be as diverse as possible.\n3. Please don't ask some simple questions about the color and position of the objects in the picture, your questions must involve some knowledge.\n4. Your instructions should not be answered directly based on the image and your instructions. Instead, it requires the test-taker to carefully observe the image and have a deep knowledge of the content within the image in order to answer correctly.\n4. If a question cannot be answered, please do not ask."
Topic: Vehicles and Transportation; Brands, Companies, and Products; Objects, Materials and Clothing; Sports and Recreation; Cooking and Food; Geography, History, Language and Culture; People and Everyday Life, Plants and Animals; Science and Technology; and Weather and Climate.
Come up with 3 diverse instructions for the knowledge topics above with different language styles and accurate answers. The instructions should contain interrogative sentences and declarative sentences. The answers should be less than 30 words.
Output format:
Instruction: instruction example
Answer: answer example
Knowledge topic: The specific knowledge topic

Demonstration:
user:
man has short hair X: 664 Y: 241 Width: 93 Height: 126\nblack computer is off X: 320 Y: 243 Width: 188 Height: 122\na soda bottle has some sodas X: 111 Y: 350 Width: 75 Height: 107\ncoke can is on the table X: 305 Y: 345 Width: 40 Height: 57\nman is wearing black tshirt X: 650 Y: 329 Width: 136 Height: 139\nmonitor is white and on X: 749 Y: 255 Width: 50 Height: 162\nkeyboard is white in colour X: 357 Y: 353 Width: 116 Height: 50\nbrown shelf is wooden X: 448 Y: 153 Width: 195 Height: 126\nman is busy working on the computer X: 650 Y: 235 Width: 128 Height: 128\nwindow is clean and clear X: 180 Y: 153 Width: 122 Height: 144\……
assistant:
Instruction: Where was the company that produces the red beverage bottle on the production table established? In what year was it established?
Answer: Atlanta, Georgia in 1892. The drink on the table is produced by the Coca-Cola Company, which was founded in Atlanta, Georgia in 1892
Knowledge topic: Brands, Companies and Products

Instruction: What is the purpose of this room?
Answer: Office room. You can see that there is only an office desk with two office computers on it, and it appears that someone is working.
Knowledge topic: People and Everyday Life
……

Figure 6: The prompt we give to GPT-4 for constructing outside-knowledge reasoning instructions.

System prompt:
I want you to act as a Prompt Rewriter. Your goal is to draw inspiration from the given visual question and given image annotations to create a brand new outside-knowledge visual question to make those famous AI systems (e.g., GPT4-V and GPT4) and well-educated humans a bit harder to handle. \nThis new prompt should belong to the same domain as the given visual instruction but be even rarer. \nThe given question contain and your new generated question should contain one of the following four types of knowledge: 1. Commonsense Knowledge: this type of knowledge is about human social behavior (e.g., that many donuts being made in a cart implies they are for sale rather than for personal consumption) 2. Visual Knowledge: this type of knowledge is about visual knowledge in the image (e.g., muted color pallets are associated with the 1950s) 3. Knowledge Bases: (e.g., hot dogs were invented in Austria) 4. Physical knowledge: this type of knowledge is about the world that humans learn from their everyday experiences (e.g., shaded areas have a lower temperature than other areas). I will provide you with a preliminary visual question, corresponding image annotation, and their knowledge type. The image annotation is some captions for an image, some entities in the caption are followed by "[x1, y1, x2, y2]" to indicate the bounding box coordinates of the entity in the image. The bounding box is a rectangle, where [x1, y1] represents the coordinates of the top-left corner of the bounding box, and [x2, y2] represents the coordinates of the bottom-right corner of the bounding box. Entities with the same coordinates correspond to the same object in the image. \nThe process of generating more complex visual instructions should take into consideration all of the following points: 1. Focus on introducing multi-hop the four types outside-knowledge about the key entities in the image or employ complex reasoning to enhance the instruction's complexity. 2. The rewritten instruction must be reasonable and must be understood and responded by humans. 3. The instruction must be definitely answered according to the provided image annotations. 4. You should try your best not to make the new generated question and its answer become verbose, and can only add 10 to 20 words into new instruction as most. 5. Your generated question must require the test taker to carefully observe the given image before they can answer it. The correct answer cannot be answered just based on the text-only question.
Output with the following format:
Complicated Instruction: <your new generated instruction here>
Answer: <the response of the complicated instruction here>
Knowledge Type: <The specific knowledge type of question>
Demonstration:
user:
Captions:\n1. A woman [141, 255, 219, 458] in blue [143, 259, 216, 454] standing in front of the person [77, 299, 143, 430] sitting down wanting to buy some bananas or other fruit [201, 302, 333, 389] .\n ...
Preliminary Instruction: Based on the image, what is the girl's reaction to the praying mantis crawling on her arm?
Answer: The girl looks in awe at the praying mantis.
assistant:
Complicated Instruction: In the image, what would be a likely scenario if the woman had been wearing a formal dress
instead?
Answer: The woman would probably not engage in play with the puppies, considering the formal attire.
Knowledge Type: Commonsense Knowledge

Figure 7: The prompt we give to GPT-4 for complicating instructions.

System prompt:
I want you to act as a Prompt Judge. I will provide you with image annotations, visual instruction, and its response, which is generated based on the given image annotations.
The image annotations are some captions for an image, some entities in the caption are followed by "[x1, y1, x2, y2]" to indicate the bounding box coordinates of the entity in the image. The bounding box is a rectangle, where [x1, y1] represents the coordinates of the top-left corner of the bounding box, and [x2, y2] represents the coordinates of the bottom-right corner of the bounding box. Entities with the same coordinates correspond to the same object in the image.
You need to judge whether the quality of this visual instruction is high enough. The principles for high-quality visual instructions are as follows:
1. Both the instruction and the response must align with the image content consistently.
2. The visual instruction requires the test-taker to carefully observe the image and understand the instruction's content in order to provide a correct response; both components are indispensable.
3. Instruction that can be answered solely based on the textual component without relying on image information is not considered good visual instruction.
4. The instruction must be reasonable and must be understood and responded by humans.
Note: Good visual instruction should meet all these conditions.
When deciding whether an instruction is good, please first provide a full explanation of your judgment and then simply answer "yes" or "no" to indicate whether you think the instruction meets the principles of good visual instruction.
Output with the following format:
Judgement evidence: <your judgement explanation here>
Judgement result: <your judgement result here>

Figure 8: The prompt we give ChatGPT for verifying instructions.

System prompt:
I will provide you a question. According to this question and answer, generate two questions, in which the positive instruction's answer should be yes, and the negative instruction's should be no. Please note that the new positive instruction should have the same inner meaning as the provided question answer. The negative example needs to be very similar to the positive directive, but with the wrong factual substitutions for important entities.

Output format:
Positive Instruction: <the modified question and the answer must be yes>
Negative Instruction: <the modified question and the answer must be no>

Demonstration:
user:
The provide question: Where was the company that produces the red beverage bottle on the production table established? In what year was it established?
The question's answer: The drink on the table is produced by the Coca-Cola Company, which was founded in Atlanta, Georgia in 1892

assistant:
Positive Instruction: Was the company that produced red drink bottles on the bench founded in Atlanta, Georgia in 1892?
Negative Instruction: Was the company that produced red drink bottles on the bench founded in Los Angeles, California in 1898?

Figure 9: The prompt we give ChatGPT for generating Bool QA instructions.

System prompt:
I will provide you with a question.
Please modify this question into a multiple-choice format.
The correct answer to the multiple-choice question has the same meaning as the answer to this question, and the choice of correct answer must be A!!!
Please make sure that the other three options B, C, and D are all wrong, and the wrong answers cannot be too simple.
Some useful tips: For the wrong option, you can replace some entities in the correct answer to accomplish this.

Output format:
Instruction: same with the provide question
A: <the correct answer>
B: <the wrong answer>
C: <the wrong answer>
D: <the wrong answer>
Answer: A

Demonstration:
user:
The provide question: Where was the company that produces the red beverage bottle on the production table established? In what year was it established?
The question's answer: Atlanta, Georgia in 1892. The drink on the table is produced by the Coca-Cola Company, which was founded in Atlanta, Georgia in 1892.

assistant:
Question: Where was the company that produces the red beverage bottle on the production table established? In what year was it established?
A: Atlanta, Georgia in 1892
B: Paris, France in 1898
C: San Francisco, California in 1912
D: Los Angeles, California in 1889
Answer: A

Figure 10: The prompt we give ChatGPT for generating multi-choice QA instructions.