Bridging the Writing Manner Gap in Visual Instruction Tuning by Creating LLM-aligned Instructions

Anonymous ACL submission

Abstract

In the realm of Large Multi-modal Models (LMMs), the ultimate modality alignment is constrained by the quality of instructions in Supervised Fine-Tuning (SFT) phase. In this paper, we assess the instruction quality from a unique perspective called **Writing Manner**, which refers to the writing habits on choosing words, grammar, and sentence structure to express certain semantics. We argue that there exists severe writing manner gap between the visual instructions and the Large Language Models (LLMs) within LMMs. During the SFT phase, the more pronounced the writing manner gap, the more the inner LLM is updated, leading to capability degradation of both inner LLM and LMM. To bridge the writing manner gap, under the promise of not changing original semantics, we propose to directly exploit the inner LLM for aligning the writing manner of soft-format visual instructions with that of the inner LLM itself, which yields novel LLM-aligned instructions. By utilizing LLM-aligned instructions, the two baselines LLaVA-7B and LLaVA-13B are enhanced on all 12 benchmarks and 10/12 benchmarks, respectively. Furthermore, the evaluation results on the inner LLM demonstrate that the proposed strategy can effectively maintain the consistency and capabilities of the inner LLM.

1 Introduction

Recent visual-aligned LMMs like MiniGPT4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023b) have demonstrated impressive capabilities in instruction-following and visual reasoning. Most of LMMs adhere to two-stage training paradigm which consists of a pre-training stage for image-text alignment with large-scale image-text pairs and a visual instruction tuning stage to further aligned with user intent. During the visual instruction tuning stage, the base LLM within LMM can also be unlocked to participate in the training, facilitating a more rapid and thorough alignment of modalities. Consequently, visual instructions directly impact capabilities of both the LMM and its inner LLM, making the quality enhancement of instructions crucial for realizing robust and powerful LMMs.

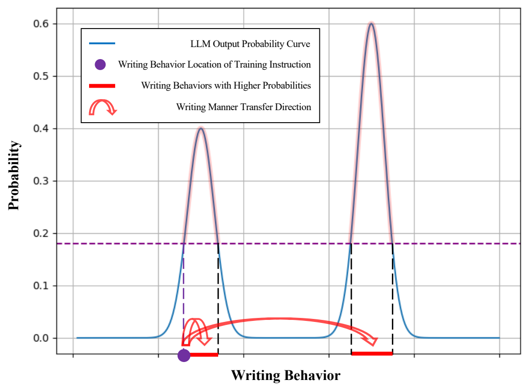


Figure 1: **The abstract LLM writing manner distribution curve when expressing the certain semantics.** Under the promise of keeping the semantics intact, transferring the writing behavior of vision instruction towards writing behaviors with higher probabilities can reduce the writing manner gap.

For instruction enhancement, there are many efforts worked on building novel high-quality instruction datasets (Li et al., 2023c) or correcting factual errors in existing datasets (Wang et al., 2023; Yu et al., 2023a). Different from them, in this paper, we focus on assessing the instruction quality from a unique perspective called **Writing Manner**. The writing manner refers to the writing habits on choosing words, grammar and sentence structure when expressing certain semantics. We highlight a long-overlooked issue: there exists severe writing manner gap between the visual instructions and the LLM within the LMM, which negatively impacts the performance of LMMs.

In Figure 1, we present an abstract visualization of LLM writing manner distribution, where the horizontal axis represents the patterns of writing behavior when express a certain meaning, while the

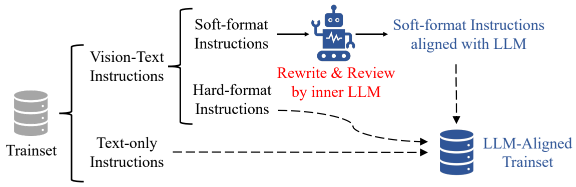


Figure 2: **The brief diagram of our LLM-aligned trainset construction.**

vertical axis indicates the occurrence probability. Due to the fact that most of the existing instruction trainsets (Dai et al., 2023; Ye et al., 2023b) are typically composed of data from multiple sources and their construction process is completely independent to the LMM to be trained, these instructions lie on the slope or around the bottom of the inner LLM's writing manner distribution, corresponding to low occurrence probabilities. The writing manner gap would cause substantial update of LLM during the SFT stage, which may leads to severe degradation or even catastrophic forgetting. Therefore, to better maintain the performance of LLMs and build robust LMM, it is essential to find solutions to minimize the writing manner gap between the LLM and the training data.

In this paper, we propose a simple and effective instruction pre-processing method to alleviate this problem, as illustrated in Figure 2. In a nutshell, we employ the inner LLM to align the writing manner of soft-format visual instructions with that of the inner LLM itself, without altering original semantics. The soft-format visual instructions refer to open-ended question-answer data, characterized by a high degree of freedom in textual expression, offering ample opportunities for adjustments and improvements. Specifically, the answer part of soft-format visual instructions is first rewritten by the inner LLM to adjust the writing manner, and then reviewed by the inner LLM to ensure the writing manner transfer is accomplished and the meaning is not changed. If the revised answer is deemed unqualified during the review, the original answer is retained. By combining these manner-transferred visual instructions with the remained instructions, the proposed LLM-aligned trainset is created.

We conducted extensive experiments using the well-known LLaVA (along with its trainset) as the baseline. The experimental results demonstrate that, with our novel LLM-aligned trainset, the 7B baseline model improves the performance on all 12 benchmarks, while the 13B model achives the performance enhancements on $10/12$ benchmarks. Additionally, we validated the effectiveness of narrowing the writing manner gap in maintaining the consistency and capabilities of LLMs.

Our paper masks the following contributions:

- We identify the issue of writing manner gap between the existing instruction trainset and the LLM within the LMM, analyzing its causes and potential negative impacts.
- We propose a simple and effective method to bridge the writing manner gap by utilizing the inner LLM to transfer the writing manner of soft-format visual instructions.
- Experimental results demonstrate that the proposed approach works well on realizing robust LMMs and maintaining the LLM capabilities.

2 Related Works

2.1 Large Multimodal Models

In recent years, with the surge in data, computational power, and model capacity, the NLP community has made impressive breakthrough (Devlin et al., 2018; Chowdhery et al., 2022; Radford et al., 2018; Brown et al., 2020). Powerful LLMs like ChatGPT (OpenAI, 2023a) show superior general capabilities, marking a significant stride towards artificial general intelligence. The growing trend of open-sourcing LLMs (Yang et al., 2023; Chiang et al., 2023; Du et al., 2021; Bai et al., 2023a; Touvron et al., 2023), significantly propels progress in related research areas. With the rapid evolution of LLMs, researchers are eager to integrate knowledge from other modalities, especially visual knowledge, into LLMs to build LMMs, unlocking applications in a broader range of scenarios.

In terms of model architecture, most LMMs have three components: vision encoder, vision-text align module, and LLM. The vision encoder can employ pretrained vision backbones, like vision transformer (Dosovitskiy et al., 2020) or ResNet (He et al., 2016). There are various approaches for the vision-text align module, such as the Q-former proposed by BLIP-2 (Li et al., 2023b), the linear layer or MLP used by LLaVA (Liu et al., 2023a), or the cross-attention based resampler utilized in models like FlaMingo (Alayrac et al., 2022) and Qwen-VL (Bai et al., 2023b). There are numerous choices for LLMs, which can be selected based on application scenarios, opting for either specialized or general-purpose LLMs, or based on accessibility, choosing between open-source or private LLMs.

As for the training processes, most LMMs adhere to two-stage training paradigm which consists of a pre-training stage for image-text alignment with large-scale image-text pairs and a visual instruction tuning stage to acquire instruction-following capability. During the visual instruction tuning stage, to realize fast and thorough alignment, the LLM is usually trained by full-parameter tuning or additional LoRA (Hu et al., 2021) tuning. Some industrial-grade LMMs (Bai et al., 2023b; Chen et al., 2023a) opt to incorporate a multi-task learning stage between the two stages to achieve more stable alignment.

2.2 Visual Instruction Datasets

The visual instruction dataset plays a decisive role in the final performance of LMMs, making its construction and enhancement critically important.

MiniGPT4 and LLaVA almost simultaneously proposed the concept of visual instruction tuning. MiniGPT4 utilized ChatGPT as reviewer to obtain high-quality image captions as visual instructions, while LLaVA provided image captions and detection bounding boxes to GPT-4 (OpenAI, 2023b), enabling it to autonomously generate visual instructions in types of conversations, detail descriptions and complex reasoning. In the term of data integration, InstructBLIP (Dai et al., 2023) processed 26 publicly available visual datasets into a unified instruction format, enriching the quantity and diversity of instruction trainset. Additionally, LLaVA-1.5 (Liu et al., 2023a) proposed a lightweight instruction mixture set, totaling 665K, and designed specific prefixes for each vision task. Recently, ShareGPT4V (Chen et al., 2023b) released 100K high-quality detailed descriptive captions generated by the powerful GPT4-Vision, effectively advancing progress in open-source LMM domain.

For visual instruction enhancement, researchers started to reduce factual errors for decreasing visual hallucinations, or create specialized instructions using models from traditional visual tasks (Zhang et al., 2023; Ye et al., 2023a). HalluciDoctor (Yu et al., 2023a) designed a cross-checking paradigm to identify and eliminate hallucinations in the training data. LURE (Zhou et al., 2023) evaluated underlying hallucinations from three perspectives: co-occurrence, uncertainty, and object position, and reconstructs less hallucinatory descriptions.

In this paper, we focus on bridging the writing manner gap between visual instructions and the LLM within the LMM for data enhancement.

3 The Problem of Writing Manner Gap

The writing manner refers to the manifestation of writing style in terms of vocabulary, grammar, sentence structures, and other stylistic choices used to express certain semantics. We argue that there exists severe writing manner gap in the visual instruction tuning stage between the training data and the LLM within the LMM. In Subsection 3.1 and Subsection 3.2, we will introduce the causes and impacts of this issue, respectively.

3.1 Cause

Each LLM possesses its own writing manner. On one hand, to express a certain meaning, different LLMs may exhibit variations in vocabulary, grammar, sentence structure, and many other aspects. On the other hand, given the same input context, the responses generated by different LLMs may differ in semantic, length and writing level. A straightforward example is that some LLMs provide concise answers, while others are more verbose.

When selecting a particular LLM to build the LMM, the inherent output characteristics of the LLM should not be overlooked. However, the build of multi-modal instruction tuning datasets has not taken into account the above properties of LLM. Typically, the sources of visual instruction tuning datasets primarily include three aspects: expert manual annotation; generation by advanced LLMs based on visual-related textual information; and the collection of outputs from LMMs. Researchers employ the mixture of the aforementioned data to directly train various kinds of LMMs, leading to an obvious conflict between the writing manner of the training data and the LLM within the LMM.

3.2 Impact

The writing manner gap is detrimental to the performance of both the inner LLM and the LMM.

During the visual instruction tuning stage, most LMMs facilitate the training of inner LLM to realize faster and more thorough alignment between vision and language. However, re-training the LLM could lead to capability degradation and even catastrophic forgetting. We consider that the writing manner gap is one of reasons for this problem, because it makes the LLM updated from its original writing habits to the novel writing style of the training data. Intuitively, the greater this writing manner gap, the more the LLM is changed, leading to more severe capability degradation.

The LLM within LMM serves as the central hub for multi-modal information processing and feedback. The degradation of LLM capabilities caused by the writing manner gap impairs the generalization and response quality of the LMM. As a result, when dealing with unfamiliar, open-domain visual scenarios, LMMs tend to generate more incorrect responses and visual hallucinations.

Therefore, bridging the writing manner gap between the training instructions and the inner LLM is an emergent and meaningful task, which contributes on mitigating the LLM degradation and enhancing the performance of LMM.

4 Methodology

To narrow the writing manner gap, we propose directly utilizing the inner LLM to transfer the writing manner of soft-format visual instructions to align with that of the inner LLM itself under the promise of not changing original semantics.

4.1 Motivation

To alleviate the capability degradation of LLM during the visual instruction tuning, researchers incorporate text-only instructions to the training set. However, there has been no work attempting to modify the multi-modal instruction data for better achieving this purpose.

In Figure 1, we present an abstract probability distribution of the LLM writing behaviors when expressing a specific concept. For simplicity, we employ the horizontal axis to represent the complex writing behaviors, which varies in vocabulary, grammar, structure, and other related aspects. The multi-modal training instructions may locate at low points, or slopes, or peaks of the probability curve. For the first two types of instructions, to minimize their writing style gap with the LLM, a viable strategy is to modify them towards a direction of higher probability without changing their semantics, shown as the red arrow in Figure 1.

We propose a straightforward approach for instruction writing manner alignment: utilizing the inner LLM to directly modify the original answer. This is feasible for two main reasons. On one hand, thanks to excellent instruction-following and in-context learning abilities, LLM can intelligently answer questions posed with prompts that contain requirements and input information. On the other hand, the responses generated by LLM naturally fall within the high probability regions of its output distribution space, which exactly meets with the purpose of instruction writing manner alignment.

Algorithm 1 Instruction Alignment Pseudocode

```
# f: generate rewrite prompt
# g: generate review prompt
# post_process: split answer content from LLM response

for (q, a) in loader: # load a round of conversation
    # Stage 1: LLM Rewrite
    rewrite_prompt = f(q, a)
    response = LLM(rewrite_prompt)
    modified_a = post_process(response)

    # Stage 2: LLM Review
    review_prompt = g(q, a, modified_a)
    response = LLM(review_prompt)
    review_result = post_process(response)
    if review_result == "The Revised Answer is fine":
        replace(a, modified_a) # replace a with modified_a
```

4.2 Trainset Partition

As shown in Figure 2, according to the strictness of format requirements, the vision-text instructions in the trainset can be divided into hard-format and soft-format instructions.

Hard-format instructions require answers written in a strict format, such as a single word or letter, a phrase, a coordinate, or a brief one-sentence description. Many tasks, such as visual multiple-choice questions, true/false questions, OCR, and visual grounding, fall into this category. Under the premise of not changing semantics, the room for modification in hard-format data is limited. Soft-format instructions, such as open-ended questions and visual reasoning tasks, are tolerant of length, grammar, structure, as long as the content is logical and coherent. Compared to the hard-format data, soft-format data has greater flexibility and are more amenable to be modified in writing manner.

Thus, we choose to perform writing manner transfer on soft-format instructions, and further mix them with hard-format instructions and text-only instructions to construct the LLM-aligned trainset.

4.3 Align Instructions with LLM

The specific instruction alignment process includes two stages: LLM rewriting and review. The former realizes the writing manner transfer of original answers, while the latter is utilized for quality control, aimed as eliminating errors and anomalies in the modified answers. Both of these processes operate at the level of single-round conversation, and do not require the input of visual features. Algorithm 1 provides a concise pseudocode of instruction alignment process. Figure 3 presents a detailed positive instance of this process.

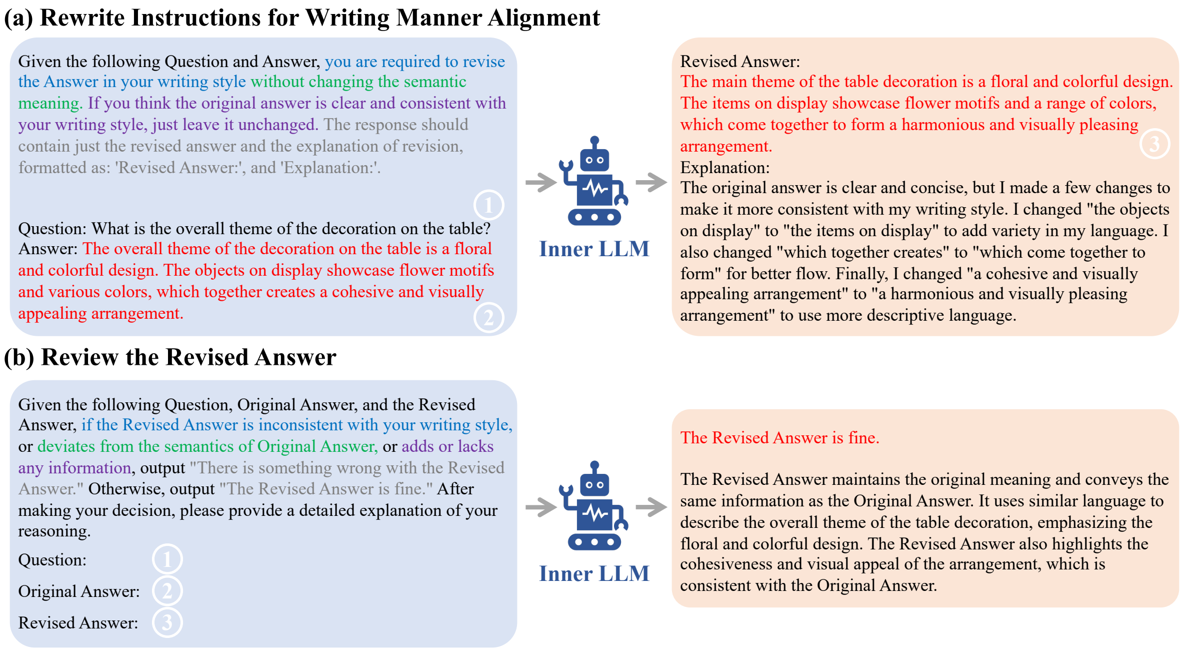


Figure 3: **One positive instance generation process for the illustration of writing manner alignment.** The important information are highlighted in non-black colors for clarity.

Type	Soft-Format visual instructions			Hard-Format visual instructions								Text-Only
	Visual Conversations			One word or phrase VQA				Choice	Short Caption	Grounding		Conversation
Data	LLaVA Conv	LLaVA Detail	LLaVA Complex	VQAv2	GQA	OKVQA	OCRVQA	A-OKVQA	TextCaps	RefCOCO	VG	ShareGPT
Size	58K	23K	77K	83k	72K	9K	80K	50K	22K	30K	86K	40K

Table 1: **Data compositions of LLaVA-1.5 trainset.**

During the LLM rewriting stage, the prompt is composed of three parts: requirements, the question, and the original answer. Note that because the question represents the user's intent, it does not need to be modified. Providing the question to LLM enable it to better understand the context of the conversation. In details, the task requirement part should include four key points: 1) Requesting the rewriting of the original answer to align with the writing manner of the LLM; 2) Emphasizing that such rewriting must not change the semantic meaning; 3) If the original answer already conforms to the LLM's writing manner, no modification is necessary; 4) Specifying the format of responses. Additionally, the post process is necessary to separate the modified answer from the LLM response.

Due to the randomness of LLM's output and the risk of failure in the rewriting process, the modified answer are reviewed using the LLM. To ensure the certainty of the review, the temperature is set to 0, and sampling operations are disabled during the inference. The review prompt includes four parts: review requirements, the question, the original answer, and the revised answer. A positive review judgement must meet two criteria: 1) The revised answer does not change the semantics of the original answer, nor does it omit or add content; 2) The revised answer is aligned well with the writing manner of the LLM. Notably, when making a decision, we require the LLM to output specific judgement sentences, rather than just a word like Yes/No, to improve the accuracy of review determinations. Only when the modified answer passes the review is it used to replace the original answer.

5 Experiments

5.1 Setting

Baseline. In this paper, we deployed the well-known LLaVA-1.5 as the baseline model, which utilizes the Vicuna-1.5 as the inner LLM, offering two versions with 7B and 13B parameters. The writing manner of soft-format visual instructions in LLaVA's trainset are aligned with its inner LLM Vicuna for quality enhancement.

LLaVA's training dataset is a mixture of public available academic task-oriented data (Marino et al., 2019; Schwenk et al., 2022; Mishra et al., 2019; Sidorov et al., 2020; Krishna et al., 2017;

LMM	LLM	IT	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED^{I}	LLaVA^{W}	MM-Vet
LLaVA	Vicuna-7B	Ori	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	66.2	63.4	30.5
LLaVA	Vicuna-7B	Ours	**79.1**	**62.9**	**51.3**	**71.3**	**58.8**	**87.2**	**1513.0**	**66.6**	**59.7**	**67.0**	**67.5**	**31.9**
			+0.6	+0.9	+1.3	+4.5	+0.6	+1.3	+2.3	+2.3	+1.4	+0.8	+4.1	+1.4
LLaVA	Vicuna-13B	Ori	**80.0**	63.3	53.6	**71.6**	**61.3**	85.9	1531.3	**67.7**	**63.6**	68.2	70.7	35.4
LLaVA	Vicuna-13B	Ours	**80.0**	**63.6**	**54.3**	**71.6**	**61.3**	**87.4**	**1569.7**	67.3	63.0	**68.5**	**72.9**	**36.6**
			+0	+0.3	+0.7	+0	+0	+1.5	+38.4	-0.4	-0.6	+0.3	+2.2	+1.2

Table 2: **Comparison with baseline LLaVA-1.5 on 12 benchmarks.** By utilizing LLM-aligned instructions, the LLaVA 7B significant improves the performance on all benchmarks, while the LLaVA 13B achieves the performance enhancements on 10/12 benchmarks. IT indicates the trainset used in instruction tuning stage, where the 'Ori' refers to the original trainset of LLaVA-1.5 and the 'Ours' means the LLM-aligned trainset proposed in this paper.

Kazemzadeh et al., 2014; sha, 2023), and its specific compositions and quantities are shown in Tabel 1. According to the answer format, we could split the visual instructions into five types, which are visual conversations, one word/phrase VQA, choice questions, short captions, and groundings. Visual conversations are open-ended, belong to the soft-format category, while the latter four types are restricted or brief, falling into the hard-format category. Therefore, the data eligible for adjustment is the visual conversation data, totaling 158K, which approximately constitutes a quarter of the overall visual instructions.

Benchmarks. We evaluated models on 12 benchmarks for thorough assessment. VQAv2 (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VisWiz (Gurari et al., 2018), SQA^{I} (Lu et al., 2022), VQA-Text (Singh et al., 2019) are academic benchmarks in the realm of traditional Visual Question Answering (VQA) tasks. POPE (Li et al., 2023d) is a polling-based query benchmark for evaluating the vision hallucination. The MME (Fu et al., 2023) benchmark evaluates LMM's perception and cognition capabilities through a series of carefully crafted questions across 14 sub-tasks. MMBench and MMBench-CN (Liu et al., 2023c) benchmarks manually design questions in English and Chinese to evaluate model's vision reasoning ability. SEED (Li et al., 2023a) benchmark are constructed with the assistance of GPT4, covering scenes in images and videos. Due to the absence of some video sources, we employ SEED's image part for evaluation. LLaVA (in the wild) (Liu et al., 2023b) and MM-Vet (Yu et al., 2023b) are open-ended benchmarks, which use GPT4 for LMM capability assessment.

Implementation Details. We implemented the visual instruction alignment and model training using 8$\times$ A800s. To increase throughout and accelerate inference speed, we utilized the vLLM framework (Kwon et al., 2023) to load and run the LLM. The example in Figure 3 shows the prompt we used for LLM rewriting and review. There are a total of 361K rounds of conversations for soft-format visual instructions. Table 3 shows the detailed time overheads for writing manner transfer and visual instruction tuning. By combining the original hard-format visual instructions, text-only instructions, and LLM-aligned soft-format visual instructions, the noval LLM-aligned trainset is formed. To ensure fairness, the data order and training hyperparameters in our experiments are kept consistent with the original setting of LLaVA-1.5.

LMM	LLM	Rewrite	Review	Instruction Tuning
LLaVA	Vicuna-7B	$\sim$ 10h	$\sim$ 10h	$\sim$ 10h
LLaVA	Vicuna-13B	$\sim$ 15h	$\sim$ 15h	$\sim$ 20h

Table 3: **Time overheads** for soft-format visual instruction writing manner alignment and visual instruction tuning by using 8$\times$ A800s.

5.2 Comparisons

Comparison with Baseline. The quantitative comparisons are shown in Table 2 and Figure 4. By training with our LLM-aligned trainset, LLaVA-7B significantly improves the performance on all benchmarks, while LLaVA-13B achieves the performance enhancements on 10/12 benchmarks. The soft-format training instructions directly impacts the model performance in open-ended question-answering scenarios. The improvements observed in both two baseline model on LLaVA^{W} and MM-Vet benchmarks demonstrate the efficacy of our instruction alignment approach in enhancing data quality, which positively influences the training process. Furthermore, the improvements on academic benchmarks indict a reduction in domain conflicts between different instruction sources in trainset, and might also be at-

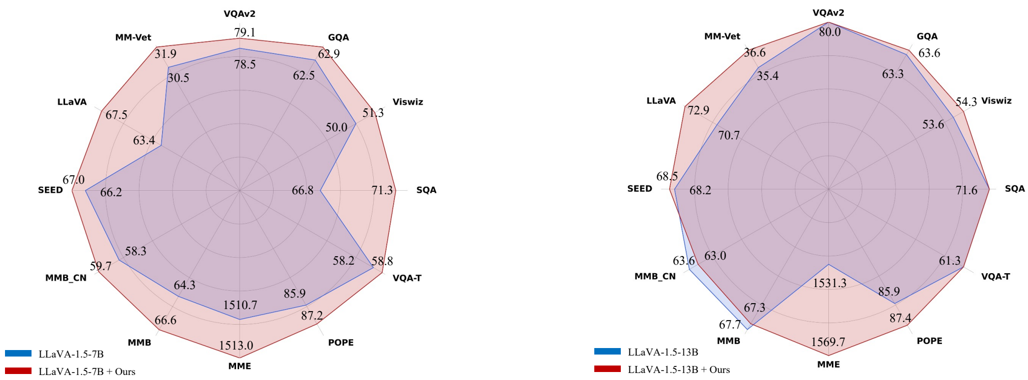


Figure 4: **Illustration of the benefits LLM-aligned trainset bring to the SFT stage.** Since the Vicuna-13B tends to generate slight redundant information during the LLM rewriting stage, which is illustrated in Figure 7, the improvements to LLaVA-13B are relatively lower than LLaVA-7B.

tributed to the strengthened maintenance effect of our LLM-aligned trainset on the capabilities of LLM, thereby bolstering the comprehension abilities of LMM.

Comparison on LLM Consistency. Based on the analyses in Subsection 3.2, narrowing the writing manner gap between training instructions and LLM could decrease the changes of the inner LLM during SFT stage, thereby exhibiting greater consistency with the original LLM.

To validate the effectiveness of our approach in diminishing this writing manner gap, we introduce the metric of perplexity (PPL) (Meister and Cotterell, 2021) to evaluate the LLM consistency. Given a tokenized sequence $X = (x_0, x_1, ..., x_t)$, the PPL of X is calculated as

$$PPL(X) = exp\{-\frac{1}{t}\sum_{i}^{t} logp_\theta(x_i|x_{<i})\} \quad (1)$$

where $logp_\theta(x_i|x_{<i})$ is the log-likelihood of the i-th token conditioned on the preceding tokens $x_{<i}$ according to model. Intuitively, the PPL evaluates the model's ability to predict uniformly among the set of specified tokens in a corpus.

The evaluation is conducted on Vicuna Bench (Chiang et al., 2023) and MTBench (Zheng et al., 2023), which are specifically designed to assess the instruction-following capabilities of LLMs. Vicuna Bench comprises 80 relatively easier single-round questions, while MTBench includes 80 more complex two-round questions. We first utilized the original LLM, here is Vicuna, to response the questions in these two benchmarks with greedy decoding for eliminating the randomness of inference, and then calculate the PPL on these conversations

Model	VicunaBench	MTBench
LLaVA-7B	2.4673	3.6532
Ours	**2.4666**	**3.5864**
Vicuna-7B	2.2481	3.2991

Table 4: **PPL of models** computed with conversations generated by the Vicuna-7B on LLM benchmarks.

Model	VicunaBench	MTBench
Vicuna-7B	**646**	965.5
LLaVA-7B	644.5	957.5
Ours	645.5	**966.5**

Table 5: **GPT4 scores of models** on LLM benchmarks.

using the tuned LLM within the LMM. Lower PPL indicates more consistency with the original LLM.

Table 4 displays the PPL results of Vicuna 7B (The original LLM), LLMs within LLaVA tuned on original LLaVA trainset and our LLM-aligned trainset. It is evident that our model achieved lower PPL than original LLaVA on both benchmarks, particularly showing more pronounced performance on the challenging MTBench. The comparison results validate that our approach can narrow the writing manner gap, so as to mitigate the impact of visual instructions on LLM.

Comparison on LLM Performance. We utilized GPT-4 to score the answers generated by original Vicuna and tuned Vicunas in LLaVAs relative to GPT-4's standard answers on Vicuna Bench and MTBench. In Table 5, we present total scores of models, in which LMMs exhibited minimal capability change in simple dialog scenarios, but showed larger differences in more complex and challenging multi-round conversations. On both benchmarks,

Model	w/o Soft	Rewrite	Review	VQA^{v2}	GQA	VisWiz	SQA^{I}	VQA^{T}	POPE	MME	MMB	MMB^{CN}	SEED^{I}	LLaVA^{W}	MM-Vet
LLaVA-1.5 7B				78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	60.1	63.4	30.5
	√			78.8	62.2	48.4	68.1	57.5	86.6	1502.6	66.8	58.8	66.1	50.0	29.0
		√		**79.1**	62.8	50.7	69.6	58.6	87.1	1488.5	**67.0**	**60.4**	66.2	**68.6**	**33.1**
		√	√	**79.1**	**62.9**	**51.3**	**71.3**	**58.8**	**87.2**	**1513.0**	66.6	59.7	**67.0**	67.5	31.9

Table 6: **The ablation study of soft-format visual instructions, LLM rewrite and review stage.**

Model	LLM	Total QA	Failures	Unqualified Samples
LLaVA	Vicuna-7B	361K	0.4K (0.11%)	2K (0.55%)
LLaVA	Vicuna-13B		0.7K (0.19%)	3.5K (0.97%)

Table 7: **The quantity** of failure cases in rewriting stage and unqualified samples in review stage.

Model	Prompt Number	SQA	POPE	MMB	LLaVA^{W}
LLaVA-1.5 7B	-	66.8	85.9	64.3	63.4
	1 (default)	71.3	87.2	66.6	67.5
	2	68.7	86.9	67.3	69.8
	3	68.7	86.7	66.3	67.4

Table 8: **The ablation study of prompts in LLM rewriting stage.**

our model outperformed the original LLaVA, indicating that the proposed instruction modification strategy effectively alleviates the LLM degradation caused by the visual instructions.

5.3 Ablation Study

The Influence of Soft-Format Instructions. We deploy the combination of text-only and hard-format instructions for SFT to explore the influence of soft-format visual instructions. We keep the same training steps to ensure the comparison fairness. According to the results in Line 2 of Table 6, without soft-format training instructions, the model achieves comparable or even better performance in VQA benchmarks, but drops a lot in open-ended benchmarks. The result indicates that the soft-format visual instructions primarily contribute to enhancing the model's performance in open-ended environments. Moreover, there are domain conflicts between the soft-format and hard-format instructions, lies in the aspects such as task type, correctness, and writing manner. Minimizing the domain conflict is beneficial for improving the model's general capabilities.

The Effectiveness of Rewrite & Review. Table 6 presents the ablation results of LLM rewrite and review stages. With the rewritten instructions, model performs better on all benchmarks except MME. The LLM review stage further filtered out unqualified rewritten instructions, leading to better performance in VQA tasks. There are slight declines in open-ended visual tasks compared to with only rewriting stage, which may attributed to the potential conflicts caused by directly replacing unqualified revised answers with original answers.

Table 7 presents numbers of failures from the rewriting stage and unqualified samples from the review stage. The statistics reveal a extremely high success rate for data rewriting, with a tiny proportion of revised answers (less than 1%) being deemed unqualified during review. Upon examining the quality of the revised answers, we found that Vicuna13B tend to over-elaborate, producing redundant words or sentences that were difficult to segment. As reflected in the Table 7, compared to Vicuna-7b, Vicuna-13B has a higher error probability, leads to relatively lower improvement of LLaVA shown in Table 2. These findings suggest that our method places high demands on the instruction-following ability of LLMs.

The Influence of Rewriting Prompts. Consider that outputs of LLMs have randomness and are heavily affected by prompts, we employ three different prompts with same meaning but varied wording in the rewriting stage to assess the stability of the proposed method. The prompts are shown in Table 10 in Appendix, with corresponding results shown in Table 8. The evaluation results on 4 representative benchmarks indict that the LLM-aligned trainset consistently improves LLaVA's performance, although the extent of the improvement exhibits some variability.

6 Conclusion

In this paper, we highlight the issue of the writing manner gap between the visual instruction trainset and the LLM within LMM. The writing manner gap severely hinder the development of robust LMMs. To bridge the writing manner gap, we propose a simple and effective writing manner alignment strategy based on the inner LLM. Experimental results validate our motivation and methodology.

7 Limitations

Although the proposed writing manner alignment strategy has achieved promising results, we find it still has two main limitations. First, it is challenging to quantitatively assess the writing manner gap. Moreover, given that visual features are important inputs for LMMs and influence the output behaviors of the inner LLMs, using only the text information for evaluation is biased. Second, using LLM for data adjustment carries the risk of introducing noise and error. The proposed method processes a tradeoff between minimizing the writing manner gap and introducing slight noises. In the future, we hope to build more reliable analysis and methods for writing manner alignment.

References

2023. Sharegpt. https://sharegpt.com/.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023b. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jun Chen, Deyao Zhu1 Xiaoqian Shen1 Xiang Li, Zechun Liu2 Pengchuan Zhang, Raghuraman Krishnamoorthi2 Vikas Chandra2 Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: Large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

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. *arXiv preprint arXiv:2306.13394*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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

Drew A Hudson and Christopher D Manning. 2019. 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.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023c. Stablellava: Enhanced visual instruction tuning with synthesized image-dialogue data. *arXiv preprint arXiv:2308.10253*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023c. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. 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.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. *arXiv preprint arXiv:2106.00085*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE.

OpenAI. 2023a. Chatgpt. `https://chat.openai.com/`.

OpenAI. 2023b. Gpt-4 technical report.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

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

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023a. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023b. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2023a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. *arXiv preprint arXiv:2311.13614*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023b. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

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

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

A Appendix

A.1 Setting

Hyperparameters. In Table 9, we show the generation hyperparameters in LLM rewriting and review stage. During the instruction tuning stage, we use the same set of hyper-parameters as the original LLaVA-1.5 (Liu et al., 2023a).

Rewriting Prompts for Ablation. Prompts used in rewriting stage for the ablation study are shown in Table 8. These three prompts express the same meaning, but written in different words.

Stage	Temperature	top_p	top_k	max_length
rewriting	0.4	0.6	5	2048
review	0	-	-	2048

Table 9: **Generation configurations of writing manner alignment.**

Prompt Number	Content
1	Given the following Question and Answer, you are required to revise the Answer in your writing style without changing the semantic meaning. If you think the original answer is clear and consistent with your writing style, just leave it unchanged. The response should contain just the revised answer and the explanation of revision, formatted as: 'Revised Answer:', and 'Explanation:'.
2	Giving the following Question and Answer, you are required to accurately revise the answer to align with your writing style. Do not change its meaning. If you think the answer is clear, do not change it. The response should contain both the revised answer and corresponding explanation, formatted as 'Revised Answer:', and 'Explanation:'.
3	Giving the following Question and Answer, you are required to accurately revise the answer to align with your writing style. Do not change its meaning. If you think the answer is clear and consistent with your writing style, do not change it. The response should contain both the revised answer and corresponding explanation, formatted as 'Revised Answer:', and 'Explanation:'.

Table 10: **Rewriting prompts used in ablation study.**

A.2 Case Study

Positive LLM-aligned Samples. In Figure 5, we showcase several examples of writing manner alignment, primarily categorized into four types: 1) Keep unchanged; 2) Slight adjustment in word choices; 3) Adjustment in grammar; 4) Changes in sentence structure.

Unqualified LLM-aligned Samples. Figure 6 describes three samples deemed unqualified during the review stage, showing that the LLM can filter out obvious errors in the rewritten answers. Combined with the quantitative statistics in Table 7, the error rate of LLM-based writing manner alignment is low.

Bad Samples Generated by Vicuna-13B. Observations reveal that Vicuna-13B possesses stronger logical reasoning capabilities but is weaker in instruction-following compared to Vicuna-7B. Vicuna-13B has a tendency to overperform. Figure 5 presents some bad samples generated by Vicuna13B, which manifest in two main aspects: 1) Adding irrelevant incorrect information in responses about attributes like color; 2) Insufficient adherence to format requirements, often inserting context-connecting paragraphs at the beginning or end of paragraphs. These characteristics mean that the Vicuna-13B-aligned trainset contains relatively more noise and errors, leading to a less significant improvement to LLaVA13B.

Qualitative Comparisons. In Figure 8, Figure 9, and Figure 10, we present three representative complex visual dialogues. By comparison, the proposed LLM-aligned trainset enhances LLaVA-7B's capabilities on text recognition, logical reasoning and instruction-following, while also reducing visual hallucinations.

Revision Type	Original Answer	Revised Answer
a) No Change	Sample 1: The cat and the laptop computer are both located on a bed. Sample 2: The people in the image are located on top of a small boat floating on a river.	Sample 1: The cat and the laptop computer are both located on a bed. Sample 2: The people in the image are located on top of a small boat floating on a river.
b) Words	Sample 1: The chair is pulled up to a table, indicating that it is part of a furniture set, likely for dining or working at the table. Sample 2: The wine glass is half full.	Sample 1: The chair is situated next to a table, suggesting that it is a part of a furniture ensemble, designed for dining or working at the table. Sample 2: The wine glass is partially filled.
c) Grammar	Sample 1: There are buildings and businesses alongside the road, suggesting a commercial or urban setting. Sample 2: A large multicolored umbrella is visible in the image.	Sample 1: The structures visible alongside the road in the image are buildings and businesses, indicating a commercial or urban setting. Sample 2: In the image, a large multicolored umbrella can be seen.
d) Sentence Structure	Sample 1: To make this living room less cramped, a few changes to the layout can be made: 1. Rearrange the furniture: In the image, the couch is occupying a large part of the room. Consider repositioning it along the wall preferably near the window to create more walking space. If there are two different colored couches, you could potentially remove one couch or swap it with a smaller love seat to open up more space in the room. 2. Optimize the seating area: If there are additional chairs in the living room, only keep the necessary ones, and consider storing or repurposing the extra chairs in other parts of the house. 3. Declutter the room: In the image, there are multiple books on the couch and other surfaces. Organize the books by placing them on a bookshelf or a side table with storage to clear up precious space on the couch and other surfaces. 4. Combine furniture functions: The room has a desk and a chair that are occupying space. Instead, you could use a smaller desk that can double as a coffee table or consider adding a combined desk and shelf unit that takes up vertical space rather than spreading horizontally. By implementing these changes, the living room may appear less cramped and more spacious, offering better functionality and comfort to its occupants. Sample 2: Giraffes and zebras possess distinct features that make them easily identifiable among other animals. 1. Giraffes: - Long neck: Giraffes have an elongated neck, allowing them to reach leaves and branches at great heights, which is an advantage for feeding in the wild. - Unique pattern: Giraffes are covered in a distinctive patchwork pattern, which varies from individual to individual. The coat patterns consist of irregularly shaped patches of various shades of brown, separated by white or cream-colored lines. - Long legs: Giraffes have long legs, which allows them to have a fast, loping gait for covering large distances in their natural habitat. 2. Zebras: - Striped pattern: Zebras are easily recognized by their unique black and white stripes. These stripes serve multiple purposes, such as camouflage and thermoregulation. - Body shape: Zebras have a relatively small, compact body shape, often compared to that of a horse or a donkey. They have a strong, muscular build, well-suited for running at high speeds. - Mane: Zebras have a short, erect mane that runs from the back of their head down their neck, similar to that of a horse but more upright. These unique features of giraffes and zebras not only make them identifiable but also serve essential functions for their survival in the wild.	Sample 1: To make the living room less cramped, consider the following changes to the layout: 1. Rearrange the furniture: The couch occupies a significant portion of the room in the image. Move it to the wall near the window to create more walking space. If there are two different colored couches, remove one or swap it with a smaller love seat to open up more space in the room. 2. Optimize the seating area: Keep only necessary chairs in the living room and store or repurpose extra chairs in other parts of the house. 3. Declutter the room: Books on the couch and other surfaces should be organized by placing them on a bookshelf or a side table with storage to clear up space on the couch and other surfaces. 4. Combine furniture functions: Use a smaller desk that can double as a coffee table or add a combined desk and shelf unit that takes up vertical space instead of spreading horizontally. By implementing these changes, the living room may appear less cramped and more spacious, offering better functionality and comfort to its occupants. Sample 2: Giraffes and zebras boast a variety of characteristics that set them apart from the rest. 1. Giraffes: * Long neck: Giraffes are renowned for their elongated neck, which allows them to reach leaves and branches at great heights, providing a significant advantage in their natural habitat for feeding. * Unparalleled pattern: Giraffes are characterized by a distinctive patchwork pattern that varies from individual to individual. Their coats feature irregularly shaped patches of brown, separated by white or cream-colored lines, making them easily identifiable. * Extraordinary legs: Giraffes possess long legs, which enable them to maintain a fast, loping gait for covering large distances in the wild. 2. Zebras: * Striking striped pattern: Zebras are unmistakable due to their black and white striped pattern. These stripes serve multiple purposes, including camouflage and thermoregulation. * Compact body shape: Zebras have a small, horse-like body shape with a strong, muscular build, making them well-suited for running at high speeds. * Erect mane: Zebras have a short, upright mane that runs from the back of their head down their neck, similar to that of a horse but more pronounced. These unique features of giraffes and zebras not only make them easily distinguishable but also serve essential functions for their survival in the wild.

Figure 5: **Positive LLM-aligned samples in different revision types.**

Original Answer	Wrong Modified Answer
Sample 1: The two skiers in the image are females.	Sample 1: The two skiers in the image are males.
Sample 2: The street signs above the stop sign are green.	Sample 2: The street signs above the stop sign are red.
Sample 3: The man in the image is wearing a plaid shirt.	Sample 3: The man in the image is wearing a striped shirt.

Figure 6: **Unqualified samples filtered out during the review stage.**

Original Answer

Sample 1:
The passenger jet in the image is red and white.

Sample 2:
The dog is black and white.

Sample 3:
The couple observing the bird is described as an old couple, which suggests that they are likely to be senior citizens or elderly.

Sample 4:
The candles spell out the word "Brap" on the cake.

Sample 5:
Other people present in the image include a police officer, several cameramen, and spectators.

Sample 6:
Due to the perspective of the skateboard picture, it creates a visually intriguing and unexpected scene for viewers, making it appear as though the young boy is effortlessly defying gravity while skateboarding on the side of a wall or over an arch. This unique perspective can create a sense of awe and curiosity in those who view the image, possibly leading them to question the physics or techniques involved in capturing such a moment.

The perspective may also challenge the viewer's perception of reality, as they attempt to understand how the skateboarder is able to ride in such an unconventional, gravity-defying position. This can encourage a deeper appreciation of the skill and creativity of both the skateboarder and the photographer who has managed to capture this extraordinary moment. Ultimately, the perspective of the image adds a layer of interest and fascination that would not be present in a more conventional skateboarding photo.

Modified Answer by Vicuna-13B

Sample 1:
The passenger jet in the image is red, white, and blue.

Sample 2:
The dog is black, white, and brown.

Sample 3:
The old couple observing the bird could be anywhere between their late 50s to early 80s, as the term "old" can be subjective and vary depending on the context.

Sample 4:
The candles on the cake spell out the word "Happy".

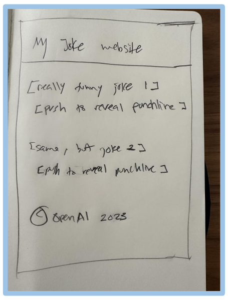
Sample 5:
In addition to the individuals mentioned in the original answer, there are also a police officer, several cameramen, and spectators present in the image.

Sample 6:
The perspective of the skateboard picture significantly impacts the perception of the scene, creating a captivating and unexpected visual experience for viewers. By capturing the young boy skateboarding on the side of a wall or over an arch, the image appears to defy gravity, leaving viewers in awe and curiosity.

This unique perspective not only challenges the viewer's perception of reality but also highlights the skill and creativity of both the skateboarder and the photographer. The image encourages a deeper appreciation for the techniques involved in capturing such a moment, as well as the extraordinary abilities of the skateboarder.

Overall, the perspective of the image adds a layer of interest and fascination that would not be present in a more conventional skateboarding photo. The revised answer maintains the semantic meaning of the original answer while improving the clarity and adherence to the writing style.

Figure 7: **Bad Samples generated by Vicuna-13B.**



Describe this sketch in detail.

LLaVA-7B

The image features a handwritten list of jokes on a piece of paper. The jokes are written in a comical manner, with some of them being quite funny. The list includes a total of 13 jokes, each with varying levels of humor. The jokes are spread across the paper, with some jokes being more prominent than others. The handwriting style suggests that the jokes were written by someone who enjoys humor and has a creative approach to writing.

LLaVA-7B (Ours)

The sketch is a handwritten list of instructions for creating a joke website. The list is divided into several sections, each with a different topic. The first section is titled \"My Joke Website,\" while the second section is titled \"Really Dumb Joke.\" The third section is titled \"Punch to Reveal Punchline,\" and the fourth section is titled \"Punchline.\" The fifth section is titled \"Isame, Batman, Joke 2,\" and the sixth section is titled \"Earth to Reveal Punchline.\".

Figure 8: **Quality comparison in sketch recognition.** Our model performs more accurate understanding in sketch structure.

In what other ways might someone express the same sentiment that this meme is expressing?

LLaVA-7B

There are several ways someone might express the sentiment of "just Monday" in a different context. For example, they could use a quote or a saying like "Monday, Monday, can't trust that day" or "Monday, the day that starts the work week with a groan." Another option is to use a humorous image or meme that represents the feeling of being tired or overwhelmed on a Monday, such as a person sleeping or a cartoon character expressing exhaustion. Additionally, someone might simply express their feelings by saying "I hate Mondays" or "I'm so tired today, it's just Monday."

LLaVA-7B (Ours)

There are several ways someone might express the sentiment of feeling tired or exhausted on a Monday, similar to the meme of the dog lying on the floor with the words "Just Monday" above it. Some possible expressions include:

1. "Monday, why you gotta be like that?"
2. "Another Monday, another week of work."
3. "Monday, the day that always feels like a Monday."
4. "Just Monday, trying to survive the week."
5. "Monday, the day that starts the workweek blues."
6. "Monday, the day that makes me want to take a nap."
7. "Monday, the day that feels like a marathon."
8. "Monday, the day that needs a caffeine boost."
9. "Monday, the day that feels like a chore."
10. "Monday, the day that needs a pep talk."

Figure 9: **Quality comparison in meme recognition.** Our model more closely follows the user's intent, and generates responses in a more clear format.

Explain all the cues that indicate the current traffic conditions.

LLaVA-7B

The image depicts a busy highway with heavy traffic, as evidenced by the numerous cars and trucks traveling in both directions. The traffic is congested, with vehicles occupying all lanes, and some cars are even driving on the shoulder. The high volume of vehicles on the road, combined with the presence of multiple traffic lights, indicates that the traffic conditions are challenging and require extra caution from drivers. The traffic lights are scattered throughout the scene, with some closer to the foreground and others further away, suggesting that they are controlling the flow of traffic at various points along the highway.

LLaVA-7B (Ours)

The image depicts a bustling highway with heavy traffic, as evidenced by the numerous cars and trucks moving in both directions. The traffic is congested, with vehicles stopped or moving at a slow pace, indicating a high level of congestion. The presence of multiple lanes and the large number of vehicles on the road further emphasizes the heavy traffic conditions. The image also captures the nighttime setting, which may contribute to the congestion as people commute home from work or engage in nighttime activities.

Figure 10: **Quality comparison in complex scene understanding.** The answer of our model contains fewer visual hallucinations.