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 instruction quality during the visual instruction tuning stage significantly influences the performance of modality alignment. In this paper, we assess the instruction quality from a unique perspective termed **Writing Manner**, which encompasses the selection of vocabulary, grammar, and sentence structure to convey specific semantics. We argue there exists a substantial writing manner gap between the visual instructions and the inner Large Language Models (LLMs) of LMMs. This gap causes the well-trained inner LLMs to deviate from their original writing styles, leading to capability degradation of both LMMs and inner LLMs. To bridge the writing manner gap while preserving the original semantics, we propose directly leveraging the inner LLM to align the writing manner of soft-format visual instructions with that of the inner LLM itself, resulting in novel LLM-aligned instructions. We develop a novel perplexity-based indicator to quantitatively assess the writing manner gap, and corresponding results show that our approach successfully minimizes this gap. By utilizing LLM-aligned instructions, the baseline models LLaVA-7B and QwenVL demonstrate enhanced resistance to hallucinations and non-trivial comprehensive improvements across all 15 visual and language benchmarks.

1 Introduction

Recent visual-aligned LMMs like MiniGPT4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023c) have shown impressive capabilities in instruction-following and visual reasoning. 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 further align with user intent. During the visual instruction tuning stage, the inner LLM of LMM could 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 quality enhancement of instructions crucial for realizing robust and powerful LMMs.

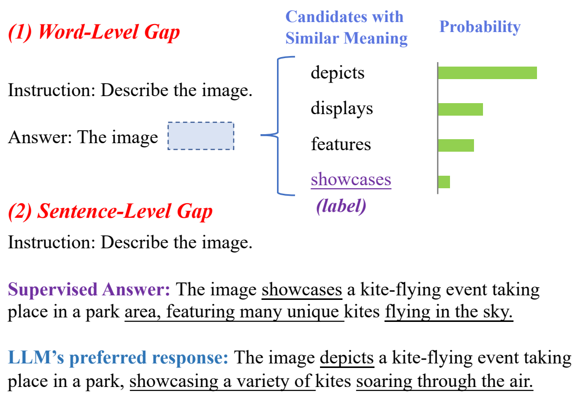


Figure 1: **The instance of word-level and sentence-level writing manner gap.**

For instruction enhancement, there are many efforts worked on building novel high-quality instruction datasets (Li et al., 2023d) or correcting factual errors in existing datasets (Wang et al., 2023; Yu et al., 2023a). In this paper, we focus on assessing the instruction quality from a unique perspective called **Writing Manner**. Writing manner refers to the specific habits of vocabulary selection, grammar usage and sentence structuring used to express particular semantics. We highlight a long-overlooked issue: there exists severe **Writing Manner Gap** between the visual instructions and the inner LLM, undermining the efficacy of LMMs.

In Figure 1, we present the instance of writing manner gap at both the word and sentence levels for illustration. Well-trained LLMs have unique writing style preferences, which are expressed in the output probabilities of candidate tokens when generating new token. The word-level gap arises when there are candidate words with similar meaning but higher probabilities than the labeled word. Further-

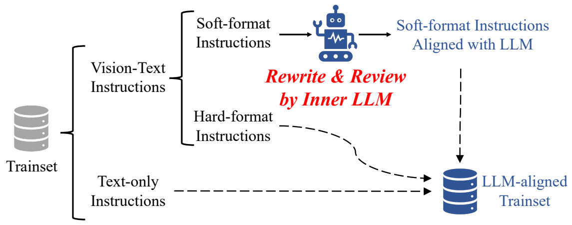


Figure 2: **The brief diagram of our LLM-aligned trainset construction.**

more, extending from the selection of single word to multiple words, the sentence-level gap is subsequently reflected in aspects of phrase, grammar and sentence structure. During the visual instruction tuning phase, the writing manner gap forces the LLM to change its original writing style, leading to performance degradation or even catastrophic forgetting. Therefore, to maintain the LLM performance and further build the robust LMM, it is essential to minimize the writing manner gap between the LLM and the training instructions.

In this paper, we propose a simple yet effective instruction processing approach to address this problem, as illustrated in Figure 2. We leverage the inner LLM to align the writing manner of soft-format visual instructions with that of the inner LLM itself, ensuring the original semantics of these instructions remain unchanged. Soft-format visual instructions refer to open-ended question-answer pairs 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 match its writing manner, and then reviewed by the inner LLM to ensure the alignment of writing manner while preserving the original meaning. If the revised answer is deemed unqualified during the review, the original answer is retained. By combining these writing manner-aligned visual instructions with other remaining instructions, the proposed LLM-aligned trainset is created.

We adopt advanced LLaVA-1.5 (Liu et al., 2023b) along with its trainset and QwenVL (Bai et al., 2023b) as baseline models and trainset. A novel Perplexity (PPL)-based indicator is designed for quantitative measurement, and corresponding results demonstrate the effectiveness of our approach in reducing the writing manner gap. By utilizing our LLM-aligned trainset, both LLaVA-7B and QwenVL achieve non-trivial comprehensive improvements across 15 visual and language benchmarks. Furthermore, careful cross-evaluation and ablation studies confirm that most improvements result from writing manner alignment rather than instruction revision.

Our contribution is three-folds: **1)** To our knowledge, we are the first to identify the problem of writing manner gap between training data and inner LLM of LMMs, and propose a novel perplexity-based metric for quantitative measurement. **2)** Without introducing any external data or models, we propose leveraging the inner LLM to reduce the writing manner gap by rewriting and reviewing soft-format visual instructions. **3)** Extensive experiments based on LLaVA-1.5 and QwenVL demonstrate the importance of reducing writing manner gap and the effectiveness of our approach.

2 Related Works

2.1 Large Multi-modal 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). 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 areas. As LLMs evolve rapidly, researchers are integrating knowledge from other modalities into LLMs to build LMMs for broader applications.

In model architecture, most LMMs consists of three main modules: the vision encoder, the vision-language connector, and the LLM. The vision encoder typically uses pre-trained vision backbones like Vision Transformer (Dosovitskiy et al., 2020) or ResNet (He et al., 2016). There are various designs for vision-language connector, such as the Q-former proposed by BLIP-2 (Li et al., 2023b), the linear layer or MLP used by LLaVA (Liu et al., 2023b), or the cross-attention-based re-sampler utilized in models like Flamingo (Alayrac et al., 2022) and QwenVL (Bai et al., 2023b).

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; Lu et al., 2024) opt to add a multi-task learning stage between the two stages to achieve more stable alignment.

2.2 Visual Instruction Construction

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

MiniGPT4 utilized ChatGPT (OpenAI, 2023a) 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. InstructBLIP (Dai et al., 2023) processed 26 publicly available visual datasets into a unified instruction format, enriching the quantity and diversity of instruction trainset. ShareGPT4V (Chen et al., 2023b) released 100K high-quality detailed descriptive captions generated by the powerful GPT4V (OpenAI, 2023c), effectively advancing progress in open-source LMM domain.

2.3 Visual Instruction Enhancement

Various approaches have been proposed using traditional small models, such as detectors and OCR tools, to reduce factual errors and visual hallucinations or to create specialized visual instructions (Zhang et al., 2023; Ye et al., 2023; Liu et al., 2023a). For example, HalluciDoctor (Yu et al., 2023a) designed a cross-checking paradigm to cut down visual hallucinations, while LURE (Zhou et al., 2023) evaluated underlying hallucinations based on co-occurrence, uncertainty and object position, and reconstructs less hallucinatory descriptions. Another strategy related to ours leverages LLMs or LMMs to improve existing instructions. In vision-language representation domain, LaCLIP (Fan et al., 2023) and VeCLIP (Lai et al., 2023) employed LLMs to rewrite or amalgamate image captions to enhance CLIP training. Additionally, some methods (Zhao et al., 2023; Du et al., 2023) utilized powerful external LLMs or LMMs to clean or synthesize visual instructions.

In this paper, we focus on reducing the writing manner gap by rewriting visual instructions with the internal LLM of LMM. Considering that our method ensures the original semantics remain unchanged, the proposed method complements other data augmentation and enhancement approaches.

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 particular semantics. We argue that there exists a substantial writing manner gap in the visual instruction tuning stage between the training data and the inner LLM of 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 unique writing manner. On one hand, to express particular meanings, 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, existing strategies of multi-modal instruction trainset construction have not taken the above LLM properties into account. Typically, the visual instruction datasets primarily originate from three sources: 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 evident conflict between the writing manner of the training data and the inner LLM.

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 achieve faster and more thorough alignment between vision and language. However, fine-tuning the well-trained LLM could lead to capability degradation and even catastrophic forgetting. One of reason for this issue is the writing manner gap, which alters the LLM's original writing habits to match the novel writing style of the training data. Intuitively, the more pronounced the writing manner gap, the more the LLM is changed, leading to more severe capability degradation.

Since the LLM within LMM serves as the central hub for multi-modal information processing and feedback, it is crucial to maintain LLM capabilities (Lu et al., 2024) for building robust LMMs. 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 enhancing inner LLM and developing robust LMM.

4 Methodology

4.1 Overall Processing

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.

This approach is feasible for two main reasons. On one hand, thanks to excellent instruction-following and reasoning capabilities, LLM can intelligently answer questions posed by 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 reducing writing manner gap.

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 at 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 our instruction alignment process. Figure 3 presents a detailed positive instance for illustration.

4.2 Trainset Partition

As shown in Figure 2, depending on the strictness of format requirements, the vision-text instructions in the trainset can be categorized 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 quite limited. In contrast, 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. Therefore, soft-format instructions have greater flexibility and are more amenable to be modified in writing manner.

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

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)
    rewrite_response = LLM(rewrite_prompt)
    modified_a, status = post_process(rewrite_response)
    if status == False:
        continue

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

4.3 Align Instructions with LLM

LLM Rewriting Stage. The input text in LLM rewriting stage consists of three parts: the requirement for writing manner alignment, the question, and the answer. Note that because the question in conversation represents the user's intent, it does not need to be modified. Providing the question to LLM enables it to better understand the context of the conversation. As shown in Figure 3, the task requirement part should include four key points: 1) Request the rewriting of the original answer to align with the writing manner of the LLM; 2) Emphasize that such rewriting must not change the semantic meaning; 3) Specify that if the original answer already conforms to the LLM's writing manner, no modification is necessary; 4) Specify the format of responses. Afterwards, the post-processing operation is designed to separate the revised answer from the LLM response. The status yielded by post-processing step indicates whether the desired answer has been obtained, with the specific procedure detailed in Appendix A.3.

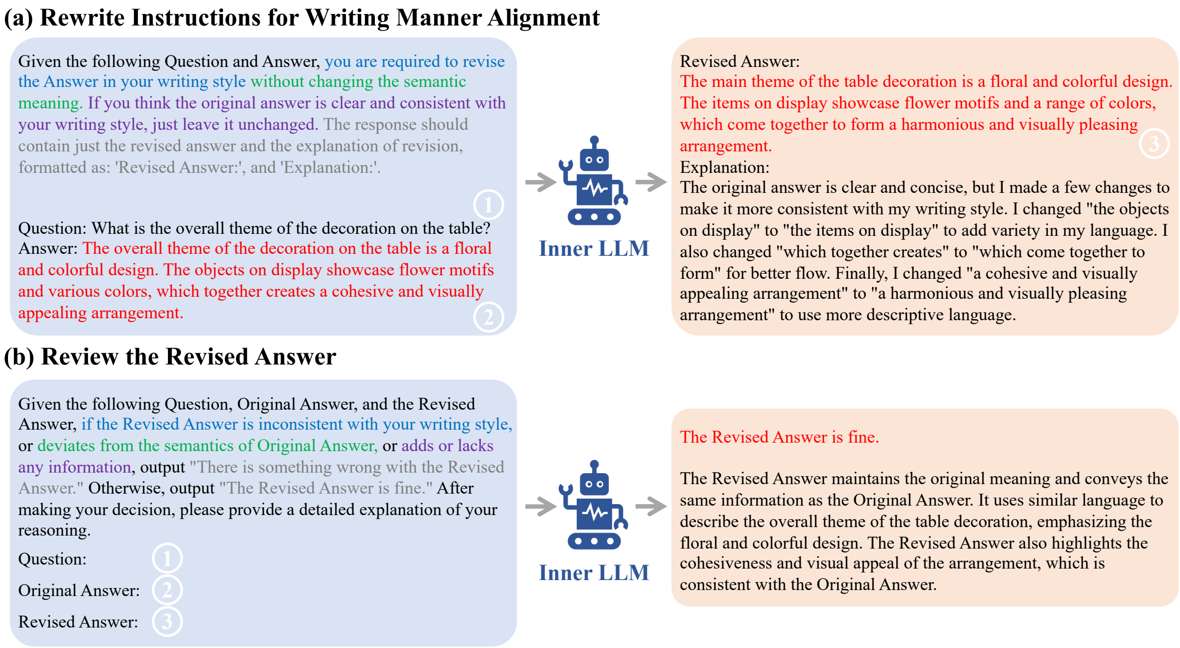


Figure 3: **One positive instance generation process for the illustration of writing manner alignment.** The important information are highlighted in non-black colors for clarity.

LLM Review Stage. Due to the randomness of LLM's output and the risk of rewriting failure, we utilize the LLM to review the modified answers for quality assurance. To ensure the certainty of the review, sampling operations are disabled during LLM inference. The review prompt includes four parts: the review requirement, 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 well-aligned 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 revised answer passes the review is it used to replace the original answer.

5 Experiments

5.1 Setting

Baseline Models and Dataset. In this paper, we utilize the well-known LLaVA-1.5 and QwenVL as the baseline model. LLaVA-1.5 employs the Vicuna-1.5 as the inner LLM, offering two versions of 7B and 13B parameters, while the QwenVL deploys the Qwen-7B as the inner LLM.

Considering LLaVA-1.5's exceptional performance and its recognition within the industry, we uniformly adopt the LLaVA-1.5's trainset as the visual instruction trainset for both LLaVA-1.5 and QwenVL pre-trained models. The writing manner of soft-format visual instructions in trainset are aligned with inner LLMs 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; Kazemzadeh et al., 2014; sha, 2023), and its specific compositions and quantities are shown in Appendix Table 8. 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.

Implementation Details. We implement the visual instruction alignment and model training using 8× A800s. To increase the throughput and accelerate inference speed, we utilize the vLLM framework (Kwon et al., 2023) to load and run LLMs. The example in Figure 3 shows the detailed prompt we used for LLM rewriting and review. There are a total of 361K rounds of conversations for soft-format visual instructions. Table 9 in Appendix shows the detailed time overheads for writing manner transfer and visual instruction tuning. By combining original text-only instructions, hard-format visual instructions and LLM-aligned soft-format visual instructions, the novel LLM-aligned trainset is formed. To ensure fairness, the order of training instructions is consistent with LLaVA-1.5, and the training hyper-parameters are same with official settings of LLaVA-1.5 and QwenVL.

5.2 Quantitative Measurement of Writing Manner Gap

Perplexity-Based Indicator. To quantitatively measure the writing manner gap between the visual instruction set and the inner LLM, we propose a PPL-based indicator. To begin with, given a tokenized sequence $X = (x_0, x_1, ..., x_t)$, the PPL of X is computed 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.

Assuming there is a pre-trained LMM M and a visual instruction set S which is divided into training set S_t and evaluation set S_e, the proposed metric is obtained in two steps. We first freeze the inner LLM of M and train M on S_t till convergence to get M', and then calculate the PPL score of the M' only on the answer part of S_e.

Why can this indicator represent the writing manner gap? When the inner LLM is frozen, its inherent writing manner remains unchanged during training. In this way, the LMM controls the subsequent output of the inner LLM sorely by adjusting visual prompts. When the LMM converges on the training set S_t, it indicates that the model has aligned as closely as possible with the content and style of the training set. At this point, the PPL measures how well the inner LLM accepts the style of the dataset. Therefore, for a specific LMM, the smaller the PPL score brought by dataset, the closer that dataset's writing manner is to that of the inner LLM.

Results and Analysis. We utilize only the original and LLM-aligned soft-format visual instructions to conduct the aforementioned evaluation to the LLaVA-7B and QwenVL, where the last $3,000$ data entries serve as the S_e, and the remaining instructions constitutes the S_t.

Model	Soft-format Instructions	
	Original	LLM-aligned
LLaVA-7B	3.413	**3.298**
QwenVL	4.208	**3.932**

Table 1: **PPL indicator of writing manner gap**.

According to the results in Table 1, both LLaVA-7B and QwenVL achieve lower PPL scores on the LLM-aligned instructions compared with the original instructions, indicating that our approach effectively reduces the writing manner gap. Additionally, there is an interesting contrast that QwenVL exhibits higher PPL scores compared to LLaVA-7B. This is because the trainset of LLaVA-7B's inner LLM Vicuna and the current soft-format visual instruction set both originate from ChatGPT, whereas QwenVL's inner LLM Qwen-7B performs a significant writing manner difference from soft-format visual instructions.

5.3 Visual Performance Comparisons

Comparison with Baseline. The quantitative comparison results on 12 benchmarks are shown in Table 2. Please refer to Appendix A.1 for details of these benchmarks. By training with our LLM-aligned trainset, LLaVA-7B and QwenVL significantly improve the performance on all benchmarks, while LLaVA-13B achieves performance enhancements in 10 out of 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 indicate a reduction in domain conflicts between different instruction sources in trainset, and might also be attributed to the strengthened maintenance effect of our LLM-aligned trainset on the capabilities of LLM, thereby bolstering the comprehension abilities of LMM.

Moreover, we also investigate the impact of LLM-aligned trainset to the LMM with frozen inner LLM. According to the last two lines in Table 2, LLaVA-7B achieves comprehensive improvements once again, which indicates that LLaVA-7B per-

LMM	inner 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**
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**
QwenVL	Qwen-7B	Ori	81.2	63.0	50.8	71.5	62.6	87.1	1576.8	71.8	64.6	68.4	70.5	41.7
QwenVL	Qwen-7B	Ours	**81.4**	**63.1**	**51.0**	**71.6**	**62.9**	**87.2**	**1589.3**	**72.0**	**65.0**	**68.9**	**72.3**	**44.3**
LLaVA	Vicuna-7B*	Ori	69.5	47.2	40.3	57.3	39.7	84.1	1104.1	45.5	32.6	50.2	58.8	28.9
LLaVA	Vicuna-7B*	Ours	**69.7**	**47.6**	**43.3**	**58.1**	**39.8**	**85.2**	**1161.1**	**46.8**	**34.7**	**50.6**	**59.4**	**29.8**

Table 2: **Performance comparisons of baseline models on 12 visual benchmarks.** "IT" indicates the trainset used in instruction tuning stage, where "Ori" refers to the original trainset of LLaVA-1.5 and "Ours" means the LLM-aligned trainset proposed in this paper. The $*$ represents the inner LLM is frozen during the fine-tuning.

Model	IT	Pope	HallusionBench Figure Acc	HallusionBench Question Acc
LLaVA-7B	Ori	85.9	14.16	44.82
LLaVA-7B	Ours	**87.2**	**16.19**	**46.32**
QwenVL	Ori	87.1	16.47	42.69
QwenVL	Ours	**87.2**	**19.08**	**43.14**

Table 3: **Visual and textual hallucination evaluation.**

Model	IT	MTBench	Alpaca-eval
LLaVA-7B	Ori	5.98	5.19
LLaVA-7B	Ours	**6.04**	**5.28**
QwenVL	Ori	4.89	2.99
QwenVL	Ours	**5.01**	**3.16**

Table 4: **Performance comparisons of baseline models on LLM evaluation benchmarks.**

forms better convergence extent to LLM-aligned trainset than original trainset.

Hallucination Evaluation. Hallucinations seriously impair the usability of LMMs. To investigate the impact of the proposed LLM-aligned instruction set on model hallucinations, we conduct hallucination assessments using POPE and HallusionBench (Guan et al., 2023), with the corresponding results presented in Table 3. The comparisons indicate that our method effectively enhances the LMM's accuracy in both visual and textual scenarios. Recalling the analysis in Subsection 3.2, our method successfully reduces the writing manner gap, thereby mitigating the disturbances to the inner LLM during the visual instruction tuning stage and improving the LMM performance. In addition to quantitative evaluation, we also present case study, please refer to Subsection D.3 in Appendix.

5.4 Textual Performance Comparisons

Comparison with Baseline. We evaluate the performance of LMMs in textual scenarios by using MTBench (Zheng et al., 2023) and Alpaca-eval (Li et al., 2023c). There two benchmarks utilize GPT-4 to score or rank model answers compared with reference answers. Table 4 displays the scores of LMMs trained with different instruction sets on MTBench (where the mean score of two assessments is taken here to mitigate the randomness of GPT-4 scoring), as well as win rates on Alpaca-eval. On both benchmarks, LLM-aligned trainset bring improvements to all baseline models compared with original instructions, demonstrating that our approach effectively alleviates the LLM degradation caused by soft-format visual instructions.

5.5 Ablation Study

The Influence of Soft-Format Instructions. We deploy the combination of text-only and hard-format instructions for tuning 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 5, 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 scenarios. 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 5 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

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

LMM	IT	PPL↓	GQA	VQA^{T}	MMB	LLaVA^{W}
LLaVA-7B	Original	3.413	62.0	58.2	64.3	63.4
LLaVA-7B	Self-aligned	3.298	62.9	58.8	66.6	67.5
LLaVA-7B	Cross-aligned	3.421	62.4	57.9	64.4	63.8
QwenVL	Original	4.208	63.0	62.6	71.6	70.5
QwenVL	Self-aligned	3.932	63.1	62.9	72.0	72.3
QwenVL	Cross-aligned	4.231	61.8	61.9	71.3	71.0

Table 6: **Cross-evaluation results for sanity check.** The "Cross-aligned" means the trainset is aligned by the other LMM, either LLaVA-7B or QwenVL.

LMM	IT	PPL↓	GQA	VQA^{T}	MMB	LLaVA^{W}
LLaVA-7B	Original	3.413	62.0	58.2	64.3	63.4
LLaVA-7B	"Plain English" Style	3.465	62.2	58.3	64.6	62.5
LLaVA-7B	Revision & No Align	3.395	62.4	57.9	66.4	66.2
LLaVA-7B	Self-aligned (Ours)	**3.298**	**62.9**	**58.8**	**66.6**	**67.5**

Table 7: **Comparison with two other revision strategies:** 1) Specific writing style of "plain English"; 2) Just revision with no writing manner alignment requirement.

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.

Sanity Check by Cross-Evaluation. We design a cross-evaluation experiment to determine whether the improvements are primarily due to bridging the writing manner gap rather than enhancements from LLM revision. Specifically, we train LLaVA-7B using the Qwen-7B-aligned trainset and Qwen-VL using the Vicuna-7B-aligned trainset, with the results shown in Table 6. In this cross-evaluation setup, there is noticeable writing manner gap between trainsets and models, as indicated by the PPL scores in Table 6. Given that both models are improved by their respective aligned trainsets, if the cross-evaluation shows better performance, we can infer that the LLM revision is the key factor. If not, it indicates that reducing the writing manner gap is crucial. As seen in Table 6, compared to using the original trainset, the LLaVA-7B in the cross-evaluation setting shows slight fluctuations, while Qwen-VL with cross-aligned trainset exhibits significant declines on GQA and VQA^{T} benchmarks. This result strongly demonstrates the importance and effectiveness of reducing writing manner gap.

Comparison with Other Revision Strategies. To further validate the effectiveness of bridging the writing manner gap, we compare the default setting with two different rewriting prompts.

The first strategy specifies a particular writing style of "plain English" by replacing the "your writing style" in default prompt with "plain English as you explain it to your children". In Table 7, we see that aligning the trainset's style to "plain English" results in larger PPL score than using the original trainset, from 3.413 to 3.465, which indicates that this style significantly differs from the default writing style of inner LLM. As for downstream evaluations, this revision method leads to poor performance on LLaVA^{W} benchmark.

The second strategy lets LLM just revise the answer without any writing manner alignment requirements for ablation. In this setting, we remove "in your writing style" and "and consistent with your writing style" in default prompt. As shown in Table 7, the inner LLM naturally generates responses similar to its default writing manner when there is no writing style constraint, indicated by PPL score drops from 3.413 to 3.395. However, this PPL decrease is not as significant as using the proposed rewriting prompt, confirms the necessity of adding writing manner alignment constraints for better reducing writing manner gap. The downstream evaluations show that this strategy enhances the model performance on most downstream tasks, except for VQA^{T} benchmark. By comparison, the overall improvement brought by these two competing strategies is far more lower than that of the proposed method, which strongly validates the importance of writing manner alignment.

6 Conclusion

In this paper, we highlight the issue of the writing manner gap between the visual instruction trainset and the inner LLM of LMM. The writing manner gap severely hinder the development of robust LMMs. Without introducing any external data or models, we leverage the inner LLM to bridge writing manner gap. Experimental results validate the effectiveness of our motivation and methodology.

7 Limitations

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 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*.

Yifan Du, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Jinpeng Wang, Chuyuan Wang, Mingchen Cai, Ruihua Song, and Ji-Rong Wen. 2023. What makes for good visual instructions? synthesizing complex visual reasoning instructions for visual instruction tuning. *ArXiv*, abs/2311.01487.

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*.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*.

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.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models.

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*.

Zhengfeng Lai, Haotian Zhang, Wentao Wu, Haoping Bai, Aleksei Timofeev, Xianzhi Du, Zhe Gan, Jiulong Shan, Chen-Nee Chuah, Yinfei Yang, et al. 2023. From scarcity to efficiency: Improving clip training via visual-enriched captions. *arXiv preprint arXiv:2310.07699*.

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 pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023c. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.

Yanda Li, Chi Zhang, Gang Yu, Zhibin Wang, Bin Fu, Guosheng Lin, Chunhua Shen, Ling Chen, and Yunchao Wei. 2023d. 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. 2023e. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. 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. 2023d. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024. Deepseek-vl: Towards real-world vision-language understanding.

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.

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.

OpenAI. 2023c. Gpt-4v(ision) system card.

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. 2023. mplug-docowl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

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*.

Zhiyuan Zhao, Linke Ouyang, Bin Wang, Siyuan Huang, Pan Zhang, Xiao wen Dong, Jiaqi Wang, and Conghui He. 2023. Mllm-dataengine: An iterative refinement approach for mllm. *ArXiv*, abs/2308.13566.

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*.

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 8: **Data compositions of LLaVA-1.5 trainset.**

LMM	LLM	Rewrite	Review	Instruction Tuning
LLaVA	Vicuna-7B	$\sim$ 10h	$\sim$ 10h	$\sim$ 10h
LLaVA	Vicuna-13B	$\sim$ 15h	$\sim$ 15h	$\sim$ 20h
QwenVL	Qwen-7B	$\sim$ 5h	$\sim$ 5h	$\sim$ 22h

Table 9: **Time overheads** for soft-format visual instruction writing manner alignment and visual instruction tuning by using 8× A800s.

LLM	Stage	Temperature	top_p	top_k	max_length
Vicuna	rewriting	0.4	0.6	5	2048
Qwen	rewriting	0.2	0.6	5	2048

Table 10: **Generation configurations** of writing manner alignment.

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%)
QwenVL	Qwen-7B		0.3K (0.08%)	0.8K (0.22%)

Table 11: **The quantity** of failure cases in rewriting stage and unqualified samples in review stage.

A Setting

A.1 Evaluation Benchmarks

By utilizing LLM to transfer writing manner of visual instructions, our approach involves a trade-off between minimizing the writing manner gap and introducing noise. To validate that our method prioritizes the former and that the impact of noise is limited, we evaluated models on 12 benchmarks for thorough assessment.

VQA^{v2} (Goyal et al., 2017), GQA (Hudson and Manning, 2019), VisWiz (Gurari et al., 2018), SQA^{I} (Lu et al., 2022), VQA^{T} (Singh et al., 2019) are academic benchmarks in the realm of traditional Visual Question Answering (VQA) tasks. POPE (Li et al., 2023e) 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., 2023d) manually design questions in English and Chinese to evaluate model's vision reasoning ability. SEED (Li et al., 2023a) benchmark is 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., 2023c) and MM-Vet (Yu et al., 2023b) are open-ended benchmarks, which use GPT4 for LMM capability assessment.

A.2 Hyperparameters

In Table 10, 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., 2023b) and QwenVL (Bai et al., 2023b).

A.3 Post-process Step

Procedure. The objective of post-processing is to separate the desired answers from the responses of LLM and to filter out apparent errors. The post-process step in LLM rewriting step contains two aspects. Firstly, based on the prompt depicted in Figure 3, the response of LLM is expected to contain two segments, starting with "Revised Answer:" and "Explanations: ". The portion between these two keywords is the desired modified answer. If these keywords are absent, the attempt is considered a rewrite failure. Secondly, we detect the presence of certain sensitive words that indicate obvious errors in the modified answer. If these sensitive words are found, this rewrite is deemed a failure. The sensitive words include "revised answer", "original answer", "revision", "semantic meaning", and "Question". In cases of rewrite failure, the original answers are reserved.

Statistics. Table 11 presents numbers of failures in 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%) be ing 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 11, compared to Vicuna-7B and Qwen-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.

Prompt Number	Content
No.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:'.
No.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:'.
No.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 12: Rewriting prompts in stability validation.

Model	Prompt Num	SQA	POPE	MMB	LLaVA^{W}
LLaVA-1.5 7B	-	66.8	85.9	64.3	63.4
	No.1	71.3	87.2	66.6	67.5
	No.2	68.7	86.9	67.3	69.8
	No.3	68.7	86.7	66.3	67.4

Table 13: The stability validation results of using rewriting prompts in same meaning but different expressions.

B More Ablation Study

B.1 The Stability Validation

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 12 in Appendix, with corresponding results shown in Table 13. The evaluation results on four representative benchmarks indicate that the LLM-aligned trainset consistently improves LLaVA's performance, although the extent of the improvement exhibits some variability.

C Discussion on Implementation Details

Q1: Why the generated explanations in LLM rewriting stage are not used afterward?

A1: We have attempted to instruct LLMs to directly output the rewritten answers without any additional information, but their instruction-following abilities are not strong enough. LLMs always append some extra explanations after outputting the revised answer, which hinders the subsequent extraction of the desired answers from the LLMs' responses. Therefore, we have decided to require LLMs to output in current format.

Q2: Why are text-only instructions not subject to going through the proposed method?

A2: It is not feasible for two main reasons. Firstly, the adopted LLMs are not powerful enough to achieve this goal. The text-only instruction set is somewhat chaotic, lengthy, diverse in task types, and encompasses various languages. For the LLM, simply maintaining the original content is challenging, let alone achieving writing manner alignment. Secondly, refining the text-only instructions would affect the analysis of the effect of improving the multi-modal instructions, which is the focus of this paper. To be honest, the quality of text-only instructions is poor. It is no doubt that improving the quality of these text-only instructions can improve the model's performance. Given that the difficulty of rewriting text-only instructions, even if the LLM successfully rephrases the text and brings improvement, it is still hard to determine whether this improvement stems from aligning writing manner or eliminating errors.

Q3: Why does QwenVL perform better than LLaVA-7B in Table 2 even though the perplexity of QwenVL is higher than that of LLaVA-7B in Table 1?

A3: The ultimate performance of LMM depends on multiple factors such as parameters, scale and pre-training data, etc, rather than just fitting to instructions. Qwen-VL's pre-training is far more comprehensive that LLaVA's (1.4B v.s. 558K image-text-pairs in pre-training stage). Therefore, comparing the PPL scores across different models and then mapping it to their performance on downstream tasks is unreasonable. Moreover, PPL scores are computed with frozen inner LLM, it cannot reflect the fitting ability of LMM with unlocked LLM.

D Case Study

D.1 Positive LLM-aligned Samples.

In Figure 4, 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.

D.2 Unqualified LLM-aligned Samples.

Figure 5 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 11, the error rate of LLM-based writing manner alignment is low.

D.3 Qualitative Comparisons.

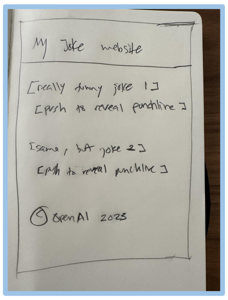
In Figure 6, Figure 7, and Figure 8, we present three representative complex visual dialogues. By comparison, the proposed LLM-aligned trainset enhances LLaVA-7B's capabilities on text recognition, logical reasoning and instruction-following, while also reducing visual hallucinations.

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

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

Figure 5: **Unqualified samples filtered out during the review stage.**



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 6: **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 7: **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 8: **Quality comparison in complex scene understanding.** The answer of our model contains fewer visual hallucinations.