Mitigating Hallucinations in Large Vision-Language Models via Entity-Centric Multimodal Preference Optimization

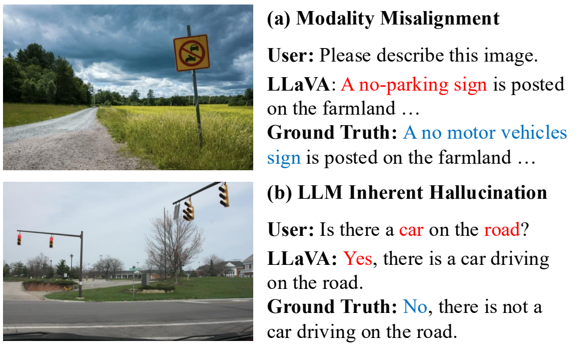
Anonymous ACL submission

Abstract

Large Visual Language Models (LVLMs) have demonstrated impressive capabilities across multiple tasks. However, their trustworthiness is often challenged by hallucinations, which can be attributed to the modality misalignment and the inherent hallucinations of their underlying Large Language Models (LLMs) backbone. Existing preference alignment methods focus on aligning model responses with human preferences while neglecting image-text modality alignment, resulting in over-reliance on LLMs and hallucinations. In this paper, we propose Entity-centric Multimodal Preference Optimization (EMPO), which achieves enhanced modality alignment than existing human preference alignment methods. Besides, to overcome the scarcity of high-quality multimodal preference data, we utilize open-source instruction datasets to automatically construct high-quality preference data across three aspects: image, instruction, and response. Experiments on two human preference datasets and five multimodal hallucination benchmarks demonstrate the effectiveness of EMPO, e.g., reducing hallucination rates by 80.4% on Object HalBench and 52.6% on MM HalBench, thereby enhancing the trustworthiness of LVLMs. The code and dataset will be made publicly available.

1 Introduction

Large Vision Language Models (LVLMs) have recently demonstrated impressive capabilities in multimodal question answering (Chen et al., 2023; Liu et al., 2023b; Bai et al., 2023; Lu et al., 2024), which typically consists of a visual encoder to extract image features, and a Large Language Model (LLM) to answer the image-related textual instructions based on the provided visual context. The LVLMs are usually learned in two steps (Li et al., 2023a; Du et al., 2022; Lin et al., 2024): (1) pre-training on large-scale image-text pairs to learn multimodal knowledge, and (2) fine-tuning on multimodal instruction-following datasets to steer their responsiveness to user instructions (Liu et al., 2023b; Wang et al., 2024c; Bai et al., 2023).



(a) Modality Misalignment

User: Please describe this image.

LLaVA: A no-parking sign is posted on the farmland …

Ground Truth: A no motor vehicles sign is posted on the farmland …

(b) LLM Inherent Hallucination

User: Is there a car on the road?

LLaVA: Yes, there is a car driving on the road.

Ground Truth: No, there is not a car driving on the road.

Figure 1: Two types of hallucinations. a) Modality misalignment: LVLM recognizes the presence of entities but confuses their semantics. b) LLM inherent hallucination: The LVLM's response is entirely dependent on textual context, disregarding the image content.

However, recent studies have identified that the LVLMs usually suffer from the hallucination problem (Li et al., 2023d; Liu et al., 2024a; Gunjal et al., 2024; Guan et al., 2024; Jiang et al., 2024b,a), akin to the LLM hallucinations (Zhang et al., 2023; Li et al., 2023b; Dhuliawala et al., 2023). Specifically, there are usually two types of LVLM hallucinations (as shown in Figure 1) (Liu et al., 2024a; Lan et al., 2024). The first type is the **modality misalignment**, which arises from the modality gap between the visual encoder and LLM, resulting in semantic mismatches between image context and textual instructions. For instance, in Figure 1(a), the LVLM (i.e., LLaVa) (Liu et al., 2023b) correctly identifies the sign on the farmland but misinterprets its meaning as *"No Parking"* instead of the correct interpretation, *"No Motor Vehicles Allowed"*. The second type is the **LLM inherent hallucination**. When the LLM inherent knowledge is either incorrect or conflicts with visual inputs, hallucinations manifest as entity co-occurrence phenomena (Lan et al., 2024). For example, as shown in Figure 1(b), *"car"* and *"road"* frequently co-occur in the LLM's pretraining corpus; the LVLM erroneously infers that whenever a *"road"* is present, a *"car"* must also be present, disregarding the image content.

To mitigate hallucinations in LVLMs, many recent studies (Zhou et al., 2024; Sun et al., 2023; Yu et al., 2024c) adopt preference alignment algorithms such as Direct Preference Optimization (DPO) (Rafailov et al., 2024), to align the model's multimodal response capabilities with human preferences. However, existing multimodal preference optimization methods extend DPO by simply adding images to preference conditions, without paying sufficient attention to entity-centric factual fragments, which are highly related to hallucinations. On the other hand, incorporating new modalities into preference conditions increases the range of possible preference combinations, necessitating a comprehensive exploration of preference dimensions to ensure LVLM responses effectively overcome the inherent hallucinations of LLMs based on images and user instructions.

To address these questions, we propose Entity-Centric Multimodal Preference Optimization (EMPO), which efficiently aligns images, user instructions, and model responses through preference optimization on comprehensive aspects. Specifically, we first construct a multimodal preference dataset based on open-source image-text instruction datasets (Zhou et al., 2024; Yu et al., 2024c) by automatically editing the entities, attributes, and relationships within the image-instruction-response triplets. Then, we apply the DPO loss across three aspects—image instruction conditioning and model responses, helping LVLMs align image entity features with the corresponding textual semantics in user instructions and model responses. To validate the effectiveness of EMPO, we evaluate it on five benchmarks under the LLaVA-1.5 (Li et al., 2024) framework with two training datasets. Experimental results demonstrate that EMPO achieves lower hallucination rates than GPT-4V (Chen et al., 2024a) across three hallucination benchmarks. Furthermore, compared to the well-known DPO algorithm, EMPO reduces hallucination rates by 85.9% on Object HalBench (Rohrbach et al., 2018) and by 49.8% on MM HalBench (Sun et al., 2023).

We summarize our contributions as: (1) We propose EMPO to effectively mitigate hallucinations by addressing the insufficient alignment of image and text modalities in existing multimodal DPO algorithms. Through preference optimization across three aspects, EMPO helps LVLMs align entity features and semantic concepts better. (2) We propose automatically constructing multimodal preference data using open-source instruction datasets to address the lack of high-quality multimodal preference data caused by the increasing complexity of preference combinations. Our data construction method can be applied to any existing instruction dataset without additional manual annotation. (3) Experimental results on two preference training datasets across five widely-used benchmarks show that EMPO enhances multimodal semantic alignment and effectively reduces hallucinations.

2 Related Work

Large Vision Language Models. Recent research on LVLMs (Zhu et al., 2023; Liu et al., 2023b; Bai et al., 2023; Lu et al., 2024; Zhang et al., 2024; Li et al., 2023a) constructs LVLMs by aligning LLMs with visual models, demonstrating superior performance across various visual–language tasks compared to earlier studies (Jia et al., 2021; Radford et al., 2021; Ju et al., 2022; Alayrac et al., 2022). These LVLMs typically adopt a two-stage training strategy: (1) **Pretraining** on large-scale image–text pairs to learn fundamental multimodal knowledge (Li et al., 2023a; Du et al., 2022; Lin et al., 2024; Bai et al., 2023), and (2) **Instruction fine-tuning** using instruction datasets to improve instruction-following abilities (Chen et al., 2024b; Wang et al., 2024c; Bai et al., 2023; Wang et al., 2024a; Li et al., 2023a, 2024). For instance, LLaVA (Li et al., 2024) introduces synthetic instructions to fine-tune an instruction-following LVLM, while MiniGPT-v2 (Chen et al., 2023) employs unique task identifiers during fine-tuning to reduce instruction ambiguity.

Hallucination in LVLMs. Despite their impressive performance, LVLMs suffer from *hallucinations*, where model responses conflict with the images, instructions, or context (Du et al., 2022; Sun et al., 2023; Xiao et al., 2025). To mitigate the hallucination, some methods have been proposed to filter out long-tail or entity co-occurrence data (Liu et al., 2023c; Yu et al., 2024a; Hu et al., 2023; Liu et al., 2023a), though this involves high annotation costs. Others recognized modal misalignment as a key factor (Li et al., 2023e; Tong et al., 2024; Cao et al., 2024; Jiang et al., 2024d,b), yet overlooked inherent LLM errors. Post-processing techniques—optimizing decoding strategies (Zhang et al., 2025; Huang et al., 2024; Yang et al., 2024;

Gao et al., 2024; Leng et al., 2024) or applying post-hoc corrections (Lee et al., 2023; Zhou et al., 2023; Yin et al., 2023)—reduce hallucinations but add inference cost. Human preference alignment has also emerged as an effective approach to mitigate hallucinations (Lan et al., 2024). LLaVA-RLHF (Sun et al., 2023) pioneered this exploration in LVLMs, while RLHF-V (Yu et al., 2024b), RLAIF-V (Yu et al., 2024c), and POVID (Zhou et al., 2024) further refined the approach with improved visual localization, text segment scoring, and preference example generation. However, these methods primarily focus on response-level preferences while neglecting the multimodal task's requirements to align human preferences with images and instructions in multiple aspects. MDPO (Wang et al., 2024b) proposed image-conditional preference alignment but overlooked aligning instructions with human preferences. In contrast, our EMPO incorporates preferences across comprehensive aspects—image, instruction condition, and model response—while leveraging entity-centric preferences for efficient alignment between visual content and semantic concepts.

3 Method

In this section, we detail the data construction and training process in the proposed EMPO. Section 3.1 introduces the foundation of preference optimization. Section 3.2 details the procedure to construct multimodal preference data. Section 3.3 introduces how EMPO addresses the hallucination by an entity-centric alignment training framework.

3.1 Preliminaries

Direct Preference Optimization (DPO) (Rafailov et al., 2024) is a primary method for human preference alignment that implicitly models the reward function in Reinforcement Learning from Human Feedback (Yu et al., 2024b). Given an instruction q, chosen response y_w, and rejected response y_l, DPO formulates the reward function as

$$r(q, y) = \beta \log \frac{\pi_\theta(y \mid q)}{\pi_{\text{ref}}(y \mid q)} + Z(q), \qquad (1)$$

where $Z(q)$ is a partition function, π_θ represents the policy model, π_{ref} is the reference model, and β is a hyperparameter controlling deviation from the reference model. DPO directly optimizes:

$$\begin{aligned}\mathcal{L}_{\text{DPO}} &= -\log \sigma(r(q, y_w) - r(q, y_l)) \\ &= -\log \sigma(\beta \log \frac{\pi_\theta\,(y_w \mid q)}{\pi_{\text{ref}}\,(y_w \mid q)} - \beta \log \frac{\pi_\theta\,(y_l \mid q)}{\pi_{\text{ref}}\,(y_l \mid q)}).\end{aligned} \qquad (2)$$

3.2 Preference Dataset Construction

As aforementioned, most existing multimodal preference datasets (Zhou et al., 2024; Yu et al., 2024c) merely align human preferences with the overall response, lacking focus on entity-centered key facts in the images and instructions. Therefore, we perform comprehensive multimodal alignment via constructing multimodal preference datasets across the full aspects: image context, instruction condition, and model response. As shown in Figure 2, we keep the original images, instructions, and response data as chosen samples, and then edit the entities, attributes, and relationships in three aspects respectively to construct rejected preference samples.

Image Preference Data To construct the rejected image, we initially employ GPT4o-mini (Achiam et al., 2023) to identify entities in both the instruction and response, ensuring close alignment between the edited image and text. Subsequently, we use an object detection model to locate these entities. Next, we apply Stable-diffusion-2 (SD2) (Rombach et al., 2022) to either remove 30% entities or substitute them with visually plausible alternatives, thereby generating an edited image as a rejected image sample v_l. Finally, we use CLIP (Jiang et al., 2023) to calculate the similarity between the edited image regions and the entity labels to ensure the image has been correctly edited. Based on the rejected image samples, these selected entities will be weighted as described in Section 3.3. We introduce two strategies to construct image preference rejected samples q_l: (1) *Entity Deleting*: Use SD2 to delete the chosen entities, helping the LVLM reduce the occurrence of non-existent entities generated by the LVLM. (2) *Entity Replacement*: Use SD2 to replace the chosen entities with incorrect but high-frequency entities, helping the LVLM overcome entity co-occurrence hallucinations (Du et al., 2022).

Instruction Preference Data We employ GPT4o-mini (Achiam et al., 2023) to adapt the original instructions in positions of the selected entities above, as well as their related attributes and relationships, thereby constructing rejected instructions q_l. Consistent with the findings of HA-DPO (Zhao et al., 2023), we observe that the distribution of GPT-modified instructions differs from the vanilla instructions, resulting in a decline in performance. We analyze that rejected samples serve as hard negative samples (Kalantidis et al.,

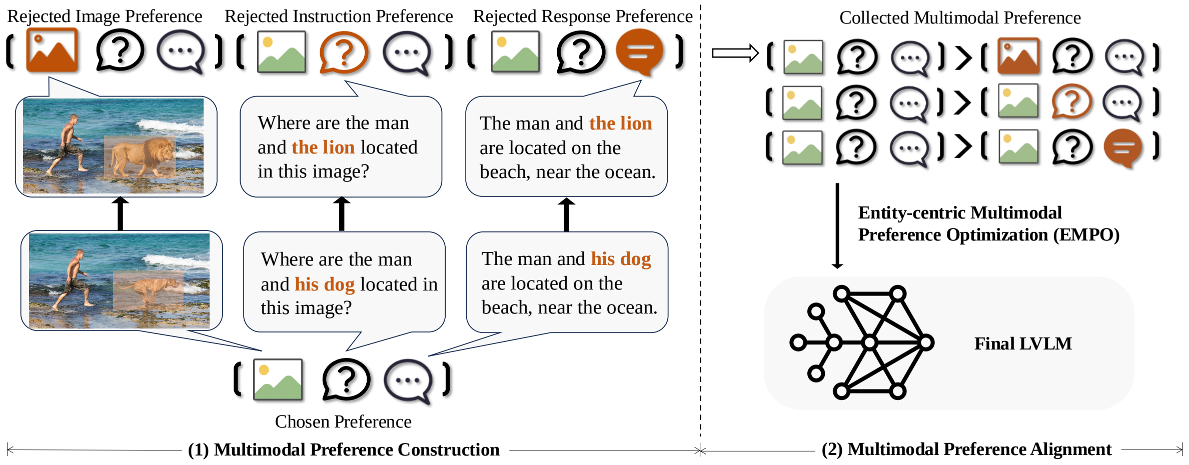

Rejected Image Preference
Rejected Instruction Preference
Rejected Response Preference
Collected Multimodal Preference
Where are the man and the lion located in this image?
The man and the lion are located on the beach, near the ocean.
Where are the man and his dog located in this image?
The man and his dog are located on the beach, near the ocean.
Chosen Preference
Entity-centric Multimodal Preference Optimization (EMPO)
Final LVLM
(1) Multimodal Preference Construction
(2) Multimodal Preference Alignment

Figure 2: Illustration of our framework. (1) At the data level, we construct a fine-grained preference alignment dataset across three aspects: image, instruction condition, and model response. (2) At the method level, we propose entity-centric multimodal preference optimization for aligning image contents with semantic concepts.

2020) because they are sufficiently similar to chosen samples while highlighting hallucination-related factual errors, which enhances LVLM's attention to entity-centric key facts. We use word2vec (Mikolov et al., 2013) and rule-based matching to ensure that the rejected instruction q_l and chosen instruction q_w have different semantic meanings but the same syntactic structure.

Response Preference Data We propose a general multimodal preference data construction method applicable to any instruction dataset, demonstrating it by construct response preferences from two open-source datasets. The first is POVID (Zhou et al., 2024) containing 17,332 entries, where we collect the rejected image preference sample (v_l) and the rejected instruction preference sample (q_l) from the previous paragraphs, using them as LVLM input to generate incorrect responses as the rejected response preference sample (y_l). The second is RLAIF-V (Yu et al., 2024c), containing 81,342 entries, where we employed MiniCPM-V2.5 (Yao et al., 2024) to compare two candidate answers generated by LLaVA-1.5 (Li et al., 2024), establishing preference rankings between responses in four iterations. Human evaluation conducted by experts on randomly sampled entries confirms the quality of these datasets. Due to space constraints, quality control and cost analysis details are included in Appendix A.

3.3 Entity-centric Multimodal Preference Optimization

In the context of LVLMs, aligning human preference includes three aspects: image, instruction, and response. To enhance LVLM's attention to image features and mitigate hallucinations caused by over-reliance on text modality, we align image and text modalities through explicit preference optimization of image instruction conditions and model responses. We define three optimization objectives: $\mathcal{L}_{\mathrm{DPOv}}$ for improving visual entity recognition, $\mathcal{L}_{\mathrm{DPOq}}$ for enhancing instruction following, and $\mathcal{L}_{\mathrm{DPOr}}$ directly align with human preference:

$$\begin{aligned}\mathcal{L}_{\mathrm{DPOv}} = -\log\sigma(\beta\log\frac{\pi_\theta\left(y \mid v_w, q\right)}{\pi_{\mathrm{ref}}\left(y \mid v_w, q\right)} \\ -\beta\log\frac{\pi_\theta\left(y \mid v_l, q\right)}{\pi_{\mathrm{ref}}\left(y \mid v_l, q\right)}),\end{aligned} \tag{3}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{DPOq}} = -\log\sigma(\beta\log\frac{\pi_\theta\left(y \mid v, q_w\right)}{\pi_{\mathrm{ref}}\left(y \mid v, q_w\right)} \\ -\beta\log\frac{\pi_\theta\left(y \mid v, q_l\right)}{\pi_{\mathrm{ref}}\left(y \mid v, q_l\right)}),\end{aligned} \tag{4}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{DPOr}} = -\log\sigma(\beta\log\frac{\pi_\theta\left(y_w \mid v, q\right)}{\pi_{\mathrm{ref}}\left(y_w \mid v, q\right)} \\ -\beta\log\frac{\pi_\theta\left(y_l \mid v, q\right)}{\pi_{\mathrm{ref}}\left(y_l \mid v, q\right)}),\end{aligned} \tag{5}$$

where w and l represent chosen and rejected preferences respectively, and $v = v_w, q = q_w, y = y_w$.

Introducing image conditions brings about more complex preference combinations, making it challenging to use vanilla DPO to assign credit to desirable key facts, leading to reward hacking (Pan et al., 2024). We propose assigning token weights to key entities in three aspects to solve this problem. Specifically, we allocate higher weights to critical features in the image, instruction, and response, thereby enhancing the LVLM's focus on entity features and enabling it to distinguish hallucinated tokens from non-hallucinated ones better.

It is noteworthy that assigning token weights does not incur additional effort, as the positions of high-weight tokens are already determined during the construction of preference data. The formula for assigning weights to model outputs is as follows:

$$\log \pi(y \mid v, q) = (1-\alpha) \sum_{y_i \notin y_e} \log p\left(y_i \mid v, q, y_{<i}\right) + \alpha \sum_{y_i \in y_e} \log p\left(y_i \mid v, q, y_{<i}\right), \tag{6}$$

where α is a weighting hyperparameter, y_i is the i-th token of response y, with larger α indicating greater token influence on preference. In this way, emphasizing hallucination-related entities strengthens human preference feedback to the LVLM, thereby enhancing its factual accuracy.

The overall multimodal preference optimization objective combines all three aspects:

$$\mathcal{L}_{\text{EMPO}} = \mathcal{L}_{\text{DPOv}} + \mathcal{L}_{\text{DPOq}} + \mathcal{L}_{\text{DPOr}}, \tag{7}$$

where the LVLMs are optimized to fully align hallucination-related key facts with human preference, reducing its hallucinations.

4 Experiments

4.1 Experimental settings

Following Yu et al., we evaluate the models from two aspects: trustworthiness for hallucination and helpfulness for general capability.

For the trustworthiness, we use three benchmarks: (1) **CHAIR** (Yu et al., 2024b), a widely used hallucination detection benchmark, which evaluates the multimodal object hallucinations by comparing the generated entities with manually labeled entities in the COCO (Lin et al., 2014a). We both report the sentence-level and the entity-level hallucination rate (denoted as CHAIR_s and CHAIR_i), respectively. (2) **MMHal-Bench** (Sun et al., 2023), which uses GPT-4 to evaluate model outputs with human responses from two aspects: hallucination rate (Hall.) and information richness (Score). (3) **AMBER** (Wang et al., 2023), which evaluates multimodal hallucinations based on 15220 yes-or-no questions. We report accuracy (Acc.) and F1 score on discriminative tasks.

For the helpfulness, we use two benchmarks: (1) **LLaVA Bench** (Li et al., 2024) assesses multimodal understanding and reasoning capabilities, with overall accuracy reported across all tasks. (2) **MME** (Fu et al., 2023) evaluates LVLMs on ten perception and four cognition subtasks, with reported scores for both categories (Per. and Cog.).

Baselines. We compare our method with state-of-the-art baselines of various types, including general baselines with strong performance and baselines designed to mitigate hallucinations.

First,*Vanilla LVLM baselines.* We use the open-source LVLMs, i.e., LLaVA-1.5 (Li et al., 2024), Qwen-VL (Bai et al., 2023), and LLaVA-Next (Liu et al., 2024b) VCD (Leng et al., 2024) for comparison. Besides, we also include GPT-4V (OpenAI, 2023) as a closed-source baseline.

Second, *Fine-tuned LVLMs baselines.* We include seven fine-tuned LVLMs aiming at mitigating hallucinations: (1) Silkie (Li et al., 2023c), which fine-tunes LVLMs using diverse instruction and feedback from GPT-4V. (2) LLaVA-RLHF (Sun et al., 2023), which extends human feedback alignment from text-only models to the multimodal domain. (3) HA-DPO (Zhao et al., 2023), which proposes the first multimodal DPO algorithm. (4) mDPO (Wang et al., 2024b), which optimizes image preferences rather than language preferences to avoid over-optimization issues. (5) POVID (Zhou et al., 2023), which fine-tunes VLLMs using model-generated preference data that targets differences between image and text. (6) RLHF-V (Yu et al., 2024b), which eliminates hallucination of VLLMs using high-quality human feedback to improve precise behavior boundaries. (7) RLAIF-V (Yu et al., 2024c), which automatically synthesizes preference data and trains the model using iterative DPO.

Training Datasets. We use the following preference datasets for training: (1) **POVID** (Zhou et al., 2024) incorporating 17,000 randomly sampled examples from the LLaVA-Instruct-150K dataset (Liu et al., 2023b). Its hallucinated responses are produced by using GPT-4V (OpenAI, 2023), which introduces potential errors in areas like object co-occurrence. (2) **RLAIF-V** (Yu et al., 2024c) is an open-source feedback dataset, including 4,000 instructions from 7 sources, such as MSCOCO (Lin et al., 2014b), Google Landmark v2 (Weyand et al., 2020), and VQA-v2 (Goyal et al., 2017). Each RLAIF-V instruction pairs with multiple open-source LVLM-generated responses, with more capable LVLMs determining response preferences.

Implementation Details. We implement EMPO based on the LLaVA-v1.5-7B (Li et al., 2024), which uses CLIP-ViT (Radford et al., 2021) as the vision module and Vicuna (Zheng et al., 2023) as the LLM backbone. We train the EMPO for 4 epochs using DeepSpeed (Lian et al., 2024), which

Model	Model	Size	Object-HalBench		MMHal-Bench		AMBER	
			CHAIR_s ↓	CHAIR_i ↓	Hall. ↓	Score ↑	Acc. ↑	F1 ↑
GPT-4V (OpenAI, 2023)▲	GPT-4V	-	13.6	7.3	28.1	3.42	83.4	87.4
Vanilla LVLMs								
QWEN-VL (Bai et al., 2023)▲	Qwen-VL-Chat	10B	40.4	20.7	38.5	2.76	81.9	86.4
LLaVA-NeXT (Liu et al., 2024b)▲	LLaVA-NeXT	34B	12.6	6.4	34.4	3.14	81.4	85.4
VCD (Leng et al., 2024)▲	LLaVA-1.5	7B	48.8	24.3	54.2	2.12	71.8	74.9
LLaVA-1.5 (Li et al., 2024)	LLaVA-1.5	7B	52.3	25.5	52.7	2.36	73.5	77.7
LLaVA-1.5 (Li et al., 2024)	LLaVA-1.5	13B	50.7	24.8	51.4	2.39	81.8	87.3
Fine-tuned LVLMs								
LLaVA-RLHF (Sun et al., 2023)▲	LLaVA-1.5	13B	38.1	18.9	62.5	2.02	79.7	83.9
Silkie (Li et al., 2023c)▲	Qwen-VL-Chat	10B	27.1	13.4	32.3	3.19	82.2	87.6
HA-DPO (Zhao et al., 2023)▲	LLaVA-1.5	7B	39.9	19.9	60.4	1.98	75.2	79.9
POVID (Zhou et al., 2024)▲	LLaVA-1.5	7B	40.4	19.1	56.2	2.08	**82.9**	87.4
RLHF-V (Yu et al., 2024b)▲	Muffin	13B	12.2	7.5	51.0	2.45	72.6	75.0
RLAIF-V (Yu et al., 2024c)▲	LLaVA-1.5	7B	8.5	4.3	29.2	3.06	76.8	84.5
MDPO (Wang et al., 2024b)♢	LLaVA-1.5	7B	35.7	9.8	54.0	2.39	73.4	74.7
DPO (POVID DataSet)	LLaVA-1.5	7B	48.9	22.4	56.0	2.15	75.1	78.9
DPO (RLAIF-V DataSet)	LLaVA-1.5	7B	19.13	9.32	36.6	2.70	76.8	81.5
EMPO (POVID DataSet)	LLaVA-1.5	7B	38.1	19.3	49.1	2.58	82.7	87.1
EMPO (RLAIF-V DataSet)	LLaVA-1.5	7B	**7.16**	**3.44**	**25.6**	**3.21**	82.4	**87.7**

Table 1: Main experimental results. The best and second-best results are highlighted in **bold** and underlined, respectively. ▲ denotes the results are reported by RLAIF-V (Yu et al., 2024c), and ♢ denotes the results from the original papers.

Model	LLaVA-Bench	MME	
	overall ↑	Cog. ↑	Per. ↑
LLaVA-1.5	60.6	297.5	1496.7
+ DPO	66.4$_{+9.57\%}$	299.3$_{+0.61\%}$	1356.7$_{-9.35\%}$
+ EMPO	69.3$_{+14.36\%}$	302.8$_{+1.78\%}$	1389.8$_{-7.14\%}$

Table 2: General capability evaluation results. Red indicates improvements after using EMPO, green indicates performance decline.

is an open-source library by Microsoft for efficient distributed training. We set a hyperparameter α of 0.7 and β of 0.5, an image resolution of 336*336, a learning rate of 5e-7, and a batch size of 8. The training is conducted on 8 A100 GPUs, taking 4 hours on the POVID dataset and approximately 12 hours on the RLAIF-V dataset.

4.2 Main Results

As shown in Table 1: (1) EMPO can substantially reduce hallucinations in the baseline model LLaVA-1.5-7B. EMPO, trained on either POVID or RLAIF-V, reduces LLaVA-1.5-7B's object hallucination rates on Object HalBench by 24.8% and 85.9%, respectively. Additionally, EMPO increases the accuracy (F1) on the AMBER benchmark by 12.5% (12.9%) relative percentage points and achieves continuous improvement in terms of overall Hallucination rate on the MMHal dataset. (2) EMPO consistently outperforms DPO across all three benchmarks and two training datasets, with better hallucination reduction effects. This indicates that the proposed EMPO method can effectively improve modal alignment. (3) EMPO achieves state-of-the-art performance in trustworthiness among open-source models, even outperforming commercial models like GPT-4V.

4.3 Ablation Studies

Our framework consists of two key components: Multi-modal Preference Alignment and Fine-grained Entity Weighting. As shown in Table 4, to verify the contribution of each component in our framework, we conduct comprehensive ablation studies on the RLAIF-V dataset.

Ablation of Multi-modal Preference Alignment. We first examine the necessity of aligning with human preferences across the *image*, *instruction*, and *response* modalities, respectively. Removing any modality preference markedly increases hallucination: +22.9%, +9.5%, and +11.9% for image, instruction, and response, respectively, but still outperforms vanilla DPO. The full three-modal alignment performs best, showing that jointly modeling visual and textual signals captures human intent more comprehensively.

Ablation of Fine-grained Entity Weighting. To evaluate the effect of explicitly emphasizing key entities, as shown in Table 4, removing entity weights in Formula 6 increases hallucination by 2.6%. Moreover, we study the impact of weighting

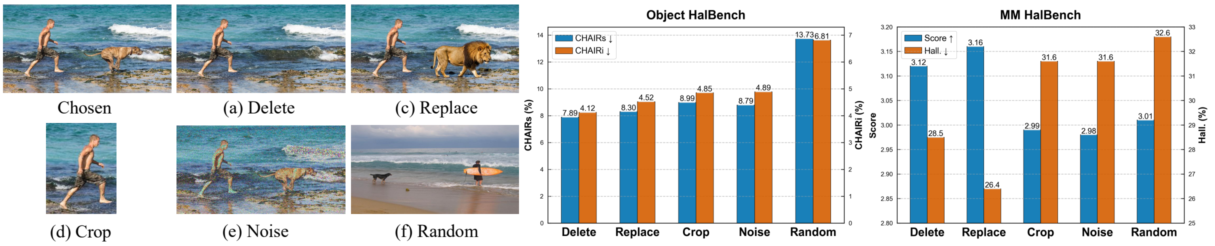

Chosen
(a) Delete
(c) Replace
(d) Crop
(e) Noise
(f) Random

Figure 3: Examples of different rejected image construction strategies and the hallucination rates.

Table 3: Ablation results on different rejected instruction & response sampling strategies.

Strategy	**Object-HalBench**		**MMHal-Bench**	
	CHAIR$_s$ ↓	CHAIR$_i$ ↓	Hall. ↓	Res ↑
Random Instruction	7.61	4.10	39.9	2.70
Edit Instruction (EMPO)	**7.16**	**3.44**	**25.6**	**3.21**
POVID Response	38.10	19.30	49.1	2.58
RLAIFV Response (EMPO)	**7.16**	**3.44**	**25.6**	**3.21**

Table 4: Ablation results on different components. We indicates "w/o image / instruction / response" denotes removing the corresponding modality preference; "w/o weighting" removes the weights on key entities.

Model	**Object-HalBench**		**MMHal-Bench**	
	CHAIR$_s$ ↓	CHAIR$_i$ ↓	Hall. ↓	Score ↑
DPO	19.13	9.32	36.6	2.70
EMPO	**7.16**	**3.44**	**25.6**	**3.21**
w/o image	10.8	5.4	34.4	2.78
w/o instruction	8.5	4.1	30.9	2.96
w/o response	11.2	6.3	32.1	2.87
w/o weighting	7.45	3.88	28.1	3.12

coefficient α in Formula 6 that balances the entity-centric term and generic preference-alignment loss. Figure 4 shows that α=0.7 yields the lowest hallucination rate on both ObjectHalBench and MMHal-Bench, indicating that assigning higher weights to entity-related tokens can effectively mitigate hallucination, while the overall semantic coherence of the response cannot be ignored.

4.4 Analysis

General Capability Analysis. Previous studies show that preference learning may impair models' general understanding capabilities (Xiao et al., 2024; Lan et al., 2024). We evaluate LVLMs' general capabilities after EMPO training using recognized evaluation datasets: LLaVA-Bench and MME's Perception and Cognition assessments. Table 2 shows EMPO improves LVLM performance on LLaVA-Bench and MME (cog.) by 14.36% and 1.78% respectively. On MME (Perception), EMPO shows a slight 7.14% decrease versus the baseline model, though it still outperforms DPO. These results indicate that while EMPO reduces hallucination in LVLMs, it maintains the model's general understanding capabilities.

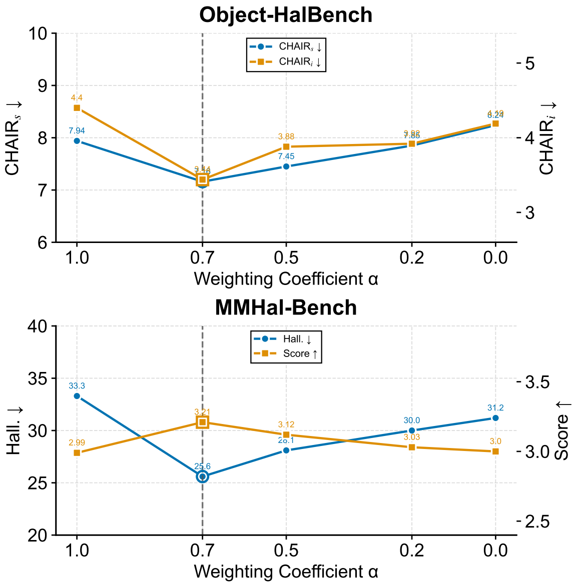
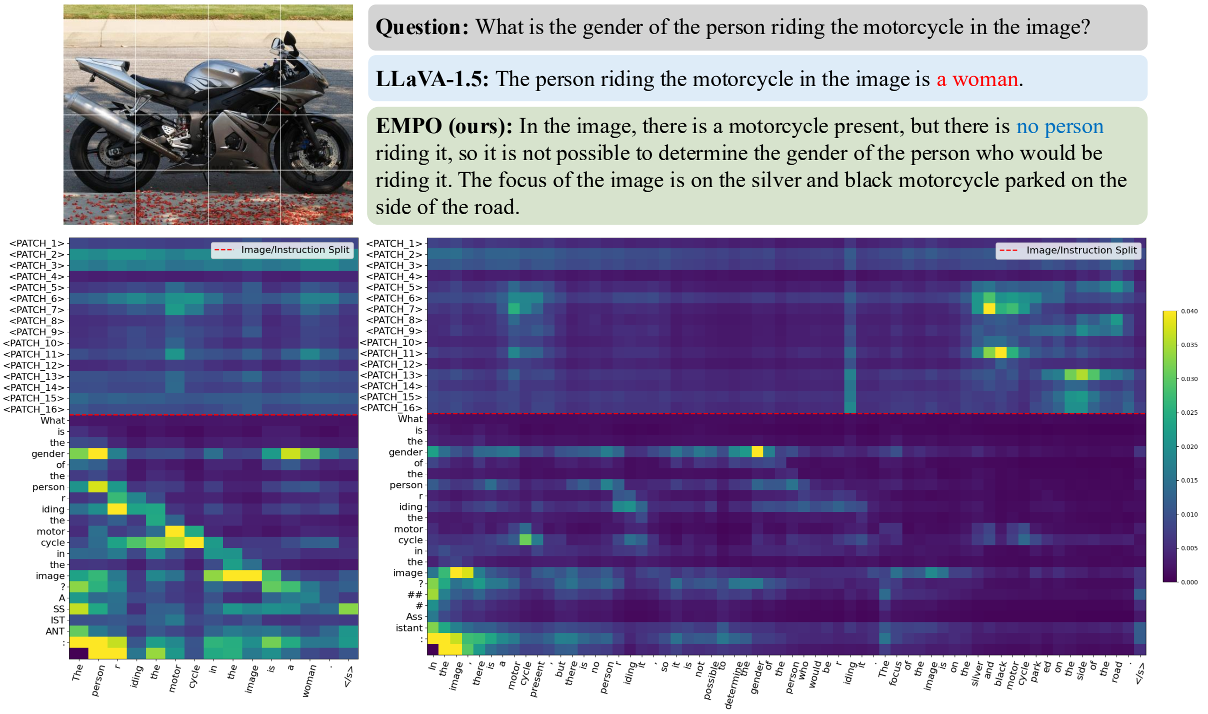


Figure 4: Hyper-parameter ablation on the α.

Impact of Rejection Construction Strategy. The construction strategy of preference samples significantly impacts the performance of preference alignment algorithms (Jiang et al., 2024c). Below, we compare different strategies with our vanilla method in three modalities.

For **rejected images**, we test five strategies: (1) Delete: using Stable-Diffusion-2 to remove key entities; (2) Replace: using Stable-Diffusion-2 to replace key entities with same-type entities; (3) Crop: randomly cropping the chosen image; (4) Noise: adding Gaussian noise; (5) Random: selecting a random training set image. Figure 3 shows *Delete* and *Replace* outperform others. These strategies maintain structural similarity with chosen images while introducing significant object representation differences, serving as hard negative samples (Kalantidis et al., 2020) and enabling finer-grained modal feature alignment. We finally adopt a combination of the Delete and Replace as rejected image construction strategies.

For **rejected instructions**, we test two strategies: (1) Edit: randomly replacing/deleting entity-related


Question: What is the gender of the person riding the motorcycle in the image?

LLaVA-1.5: The person riding the motorcycle in the image is a woman.

EMPO (ours): In the image, there is a motorcycle present, but there is no person riding it, so it is not possible to determine the gender of the person who would be riding it. The focus of the image is on the silver and black motorcycle parked on the side of the road.

Image/Instruction Split
<PATCH_1>
<PATCH_2>
<PATCH_3>
<PATCH_4>
<PATCH_5>
<PATCH_6>
<PATCH_7>
<PATCH_8>
<PATCH_9>
<PATCH_10>
<PATCH_11>
<PATCH_12>
<PATCH_13>
<PATCH_14>
<PATCH_15>
<PATCH_16>
What
is
the
gender
of
the
person
r
iding
the
motor
cycle
in
the
image
?
A
SS
IST
ANT
:
The person r iding the motor cycle in the image is a woman . </s>
##
#
Ass
istant
In the image , there is a motor cycle present , but there is no person r iding it , so it is not possible to determine the gender of the person who would be r iding it . The focus of the image is on the silver and black motor cycle park ed on the side of the road . </s>
0.040
0.035
0.030
0.025
0.020
0.015
0.010
0.005
0.000

(a) LLaVA-1.5

(b) EMPO (Ours)

Figure 5: Attention weight heatmaps of outputs to image and instruction tokens from LLaVA-1.5 and EMPO. The hallucinated tokens are highlighted in red. The image is tokenized into 4×4 patches (above the red dash line).

nouns, attributes, and relationships. (2) Random: selecting a different random instruction from the training set. As shown in Table 3, Edit achieves a lower hallucination rate than Random, so we select Edit as our primary strategy.

For **rejected responses**, we experiment with: (1) POVID (Zhou et al., 2024): using LVLM-generated responses from rejected images. (2) RLAIF-V (Yu et al., 2024c): constructing preference pairs using an evaluation model. Table 3 shows RLAIF-V significantly outperforms POVID, thus we adopt it for our final EMPO implementation, yet to verify our approach's robustness across different data configurations, we report the results of both instruction datasets compared against DPO in Section 4.2.

Modality Alignment Visualization. Figure 5 presents attention heatmaps illustrating how model output tokens (horizontal axis) attend to input image patches (above the red line) and instruction tokens (below the red line) during inference. In the baseline LLaVA-1.5 (Figure 5(a)), the attention heatmap reveals that while generating the hallucinated entity *"woman"*, the model incorrectly focuses on specific image patches (e.g., <PATCH_6>, <PATCH_11>) and the word *"gender"* in the question. This misaligned attention pattern exemplifies how the model grounds erroneous assertions in visual features, leading to hallucination. In contrast, our EMPO-enhanced model (Figure 5(b)) demonstrates significantly improved modality alignment. When inferring *"no person"* and describing *"silver and black motor"*, the model's attention correctly concentrates on image patches showing only the motorcycle (e.g., <PATCH_6>, <PATCH_7>, <PATCH_11>), effectively verifying the absence of a rider. This precise attention allocation shows how EMPO helps LLaVA-1.5 correctly attend to key facts in both the image and user instruction, fostering stronger semantic consistency between visual inputs and generated text to avoid hallucination.

5 Conclusion

This paper addresses the LVLM hallucination problem from two perspectives: modality misalignment and LLM inherent hallucination. At the method level, we propose a comprehensive multimodal preference optimization method to help LVLM align entity features with semantic concepts, enhancing its trustworthiness. For data side, we introduce a general method for constructing multimodal preference data. Experiments on multiple benchmarks show our method significantly reduces hallucinations while maintaining LVLM capabilities. For future work, we will explore commonsense knowledge in multimodal domains and investigate hallucinations in long-term interactive environments like multi-turn dialogue.

Limitations

A limitation of this paper is that the investigation into hallucinations was restricted to entity-centric hallucinations. Although entity-centric hallucinations constitute a major component of multimodal hallucinations, non-entity-related hallucinations such as common sense knowledge and long context memory loss are also important aspects for optimizing LVLM effectiveness. Due to space limitations and the complexity of defining common sense and long context, we did not explore these issues in this paper. We propose the following directions for future research: (1) Exploring methods to define common-sense knowledge in multimodal domains and its relationship with hallucinations. (2) Investigating hallucination issues in a long-term interactive environment, such as multi-turn dialogue.

Ethics Statement

This study focuses on mitigating hallucination phenomena in LVLMs to enhance their reliability and trustworthiness. We have carefully considered the ethical implications of the research and do not expect any major ethical issues to arise. This study is based on publicly available and widely used data and models; therefore, our findings may inherit the biases and limitations present in these resources.

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

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

Yuhang Cao, Pan Zhang, Xiaoyi Dong, Dahua Lin, and Jiaqi Wang. 2024. Dualfocus: Integrating macro and micro perspectives in multi-modal large language models. *arXiv preprint arXiv:2402.14767*.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer.

Ruibo Chen, Yihan Wu, Lichang Chen, Guodong Liu, Qi He, Tianyi Xiong, Chenxi Liu, Junfeng Guo, and Heng Huang. 2024b. Your vision-language model itself is a strong filter: Towards high-quality instruction tuning with data selection. *arXiv preprint arXiv:2402.12501*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. *arXiv preprint arXiv:2202.10936*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Minghe Gao, Shuang Chen, Liang Pang, Yuan Yao, Jisheng Dang, Wenqiao Zhang, Juncheng Li, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. 2024. Fact: Teaching mllms with faithful, concise and transferable rationales. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 846–855.

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, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. 2023. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Chaoya Jiang, Hongrui Jia, Mengfan Dong, Wei Ye, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024a. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 525–534.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024b. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. 2023. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*.

Ruili Jiang, Kehai Chen, Xuefeng Bai, Zhixuan He, Juntao Li, Muyun Yang, Tiejun Zhao, Liqiang Nie, and Min Zhang. 2024c. A survey on human preference learning for large language models. *arXiv preprint arXiv:2406.11191*.

Songtao Jiang, Yan Zhang, Chenyi Zhou, Yeying Jin, Yang Feng, Jian Wu, and Zuozhu Liu. 2024d. Joint visual and text prompting for improved object-centric perception with multimodal large language models. *arXiv preprint arXiv:2404.04514*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. 2020. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33:21798–21809.

Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. 2024. A survey of hallucination in large visual language models. *arXiv preprint arXiv:2410.15359*.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023b. Halueval: A large-scale hallucination evaluation benchmark for large language models. *arXiv preprint arXiv:2305.11747*.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. 2023c. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*.

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*.

Zhuang Li, Yuyang Chai, Terry Yue Zhuo, Lizhen Qu, Gholamreza Haffari, Fei Li, Donghong Ji, and Quan Hung Tran. 2023e. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*.

Xinyu Lian, Sam Ade Jacobs, Lev Kurilenko, Masahiro Tanaka, Stas Bekman, Olatunji Ruwase, and Minjia Zhang. 2024. Universal checkpointing: Efficient and flexible checkpointing for large scale distributed training. *Preprint*, arXiv:2406.18820.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014a. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014b. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

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. In *The Twelfth International Conference on Learning Representations*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Yanqing Liu, Kai Wang, Wenqi Shao, Ping Luo, Yu Qiao, Mike Zheng Shou, Kaipeng Zhang, and Yang You. 2023c. Mllms-augmented visual-language representation learning. *arXiv preprint arXiv:2311.18765*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

OpenAI. 2023. Gpt-4v system card.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. 2024. Feedback loops with language models drive in-context reward hacking. *arXiv preprint arXiv:2402.06627*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and 1 others. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024b. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, and 1 others. 2024c. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. 2020. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Wenyi Xiao, Zechuan Wang, Leilei Gan, Shuai Zhao, Wanggui He, Luu Anh Tuan, Long Chen, Hao Jiang, Zhou Zhao, and Fei Wu. 2024. A comprehensive survey of direct preference optimization: Datasets, theories, variants, and applications. *Preprint*, arXiv:2410.15595.

Dingchen Yang, Bowen Cao, Guang Chen, and Changjun Jiang. 2024. Pensieve: Retrospect-then-compare mitigates visual hallucination. *arXiv preprint arXiv:2403.14401*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. 2024a. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024b. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024c. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. 2025. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, and 1 others. 2024. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output. *arXiv preprint arXiv:2407.03320*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

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 Quality control and Cost analysis

A.1 Quality control

To validate the effectiveness of our image editing approach, we conducted a comprehensive evaluation using CLIP-based similarity metrics. We analyzed the similarity between edited image regions and their corresponding entity labels before and after the editing process. We employed the CLIP model (Jiang et al., 2023) to calculate the semantic similarity between image regions and textual entity labels. This approach provides an objective measure of how well the edited regions align with their intended semantic meaning. Figure 6 presents a detailed visualization of our findings.

As shown in Figure 6, our editing approach significantly improves the semantic alignment between image regions and their target entity labels. The similarity score distribution (Figure 6a) demonstrates a clear shift from predominantly low scores before editing to significantly higher scores after editing. The before-editing distribution peaks around 0.3, while the after-editing distribution centers approximately at 0.75, indicating a substantial improvement in semantic accuracy.

The entity-specific analysis (Figure 6b) reveals consistent improvements across all five entity types. Notably, the "person" and "car" categories show the most substantial gains, with median similarity increases of 0.41 and 0.39, respectively. This suggests that our editing approach is particularly effective for these common object categories. The distribution of similarity improvements (Figure 6c) further confirms the effectiveness of our approach, with a mean improvement of 0.43 across all samples. The distribution is positively skewed, indicating that while most edits yield moderate improvements, a significant portion achieves substantial similarity gains exceeding 0.6. Most importantly, when analyzing editing success rates using various similarity thresholds (Figure 6d), we observe that at a threshold of 0.5—commonly used in semantic similarity tasks—the success rate increases from 27.3% before editing to 89.7% after editing. This represents a 62.4 percentage point improvement, demonstrating the robustness of our approach.

A.2 Cost analysis

Having established robust quality safeguards, we now consider quantifying the computational requirements and monetary costs of each processing step. Extracting image entities costs about \$0.002 per entry; generating all 17,332 image entries on 8 A100 GPUs takes roughly 4–6 hours. Instruction-level entity extraction and rewriting add \$0.002 and \$0.003 per entry, respectively. In total, the labeling cost is \$0.007 per entry plus 8.3 seconds of compute time on a single A100. By contrast, manual annotation would cost approximately \$0.30 per entry, over 40× higher.

B Hallucination Cause Analysis Experiment

To better understand the severity of hallucinations in LVLMs, we conducted a pilot experiment to evaluate their inference performance. Specifically, we assessed the models on 200 preference examples selected from the POVID dataset (Zhou et al., 2024). Through this analysis, we identify two prominent types of errors in LVLM responses, which highlight critical limitations in their reasoning and multimodal understanding capabilities.

1. Concept Confusion: We observe that LVLMs struggle to accurately interpret semantic relationships between entities, leading to concept confusion. The models frequently generate identical or highly similar responses to user instructions that were semantically conflicting or conceptually distinct, which suggests that LVLMs may fail to fully grasp the fine-grained differences between related but distinct concepts, resulting in responses that lack precision and contextual appropriateness.

2. Visual Neglect: When provided with only textual context (i.e., without accompanying visual input), the models tend to generate image-agnostic responses that disregard the potential relevance of visual information. This behavior indicates an over-reliance on textual cues and insufficient attention to visual content, which we attribute to the influence of LLM-induced hallucinations. Such hallucinations appear to bias the models toward text-based reasoning, even in scenarios where visual understanding is critical. This is also in line with the previous work PAI (Liu et al., 2024c)

These findings highlight the challenges LVLMs face in achieving robust multimodal understanding and highlight the need for improved mechanisms to mitigate hallucinations. Addressing these issues is essential for enhancing the reliability and appli-

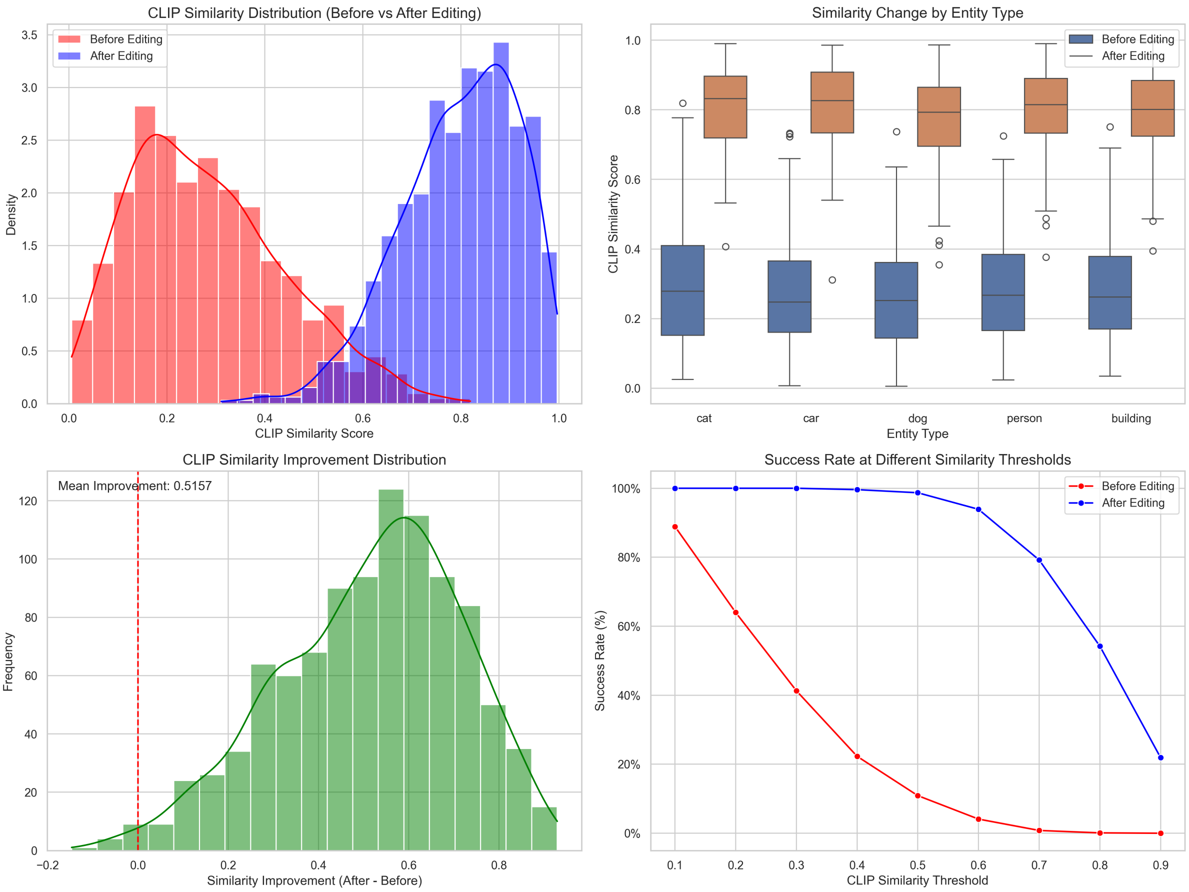


Figure 6: CLIP-based similarity analysis for image editing evaluation: (a) Distribution of similarity scores before and after editing, (b) Box plots showing similarity changes by entity type, (c) Distribution of similarity improvement scores, and (d) Success rates at different similarity thresholds.

cability of LVLMs in real-world tasks that require both textual and visual reasoning.

C Prompt Appendix

The section is to describe the GPT4o-mini prompt for identifying entities and the prompt for rewriting chosen instructions and the rejected instructions.

The prompt for rewriting the chosen instruction:

```
# prompt for rewriting chosen instruction
prompt = '''Task: Rephrase the following question
while maintaining its original meaning:

Original question: {question}

Requirements:
1. If original question was a declarative sentence,
then keep rewritten question as a declarative
sentence.
2. Ensure the rephrased question is clear, concise,
and maintains the original inquiry intent.
3. You may adjust sentence structure or wording, but
 do not change the essence of the question.
4. If necessary, slightly expand the question to
improve clarity, but keep it concise.
5. Use natural, fluent English in the rephrased
version.
Please only provide the rephrased question that
meets these criteria without any additional
explanation.
'''
```

The prompt for rewriting rejected instruction:

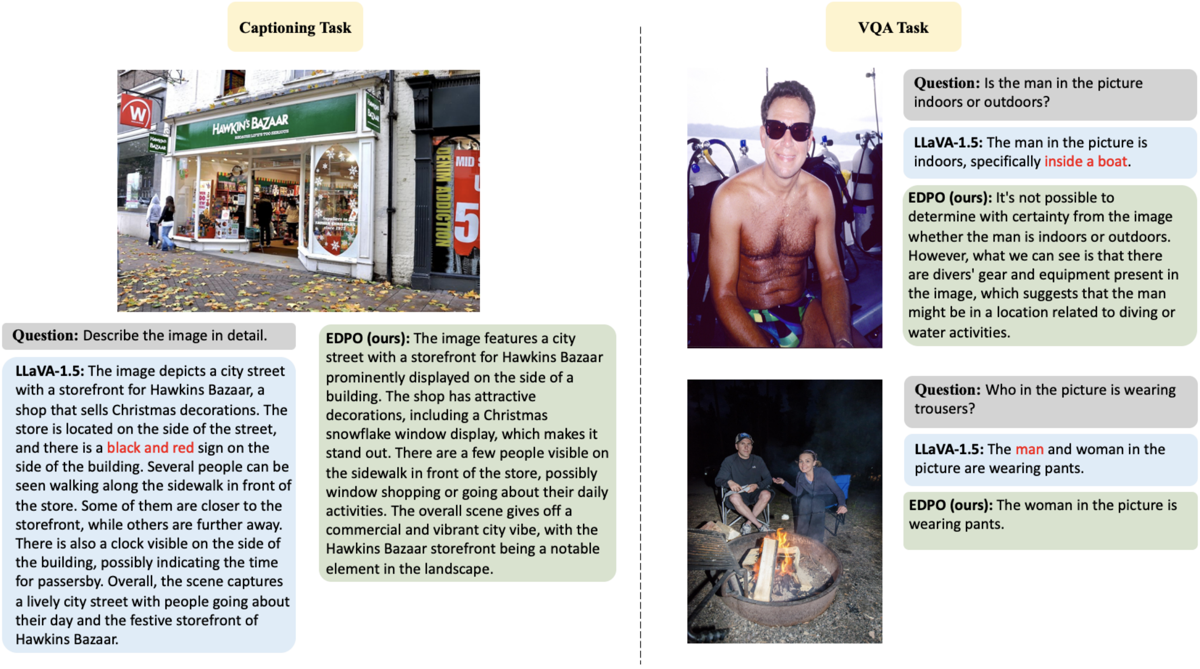
```
# prompt for rewriting rejected instruction
'''You are an expert in creative writing and
linguistic transformation. Your task is to rewrite
the given question so that its meaning is
significantly different from the original, while
maintaining the same general structure and format.
Follow these guidelines:

1. Analyze the original question's structure, tone,
and key elements.
2. Identify a different perspective or context that
could radically change the question's meaning.
3. Rewrite the question using the new perspective,
ensuring it has a distinctly different meaning.
4. Maintain the original question's format,
including any specific phrasing or sentence
structure.
5. Ensure the rewritten question is coherent,
grammatically correct, and makes sense on its own.

Original question: {question}

Rewritten question:

Provide only the rewritten question without any
additional explanation.'''
'''
```

The prompt for identifying entities:

```
# prompt for identifying entities
prompt = ’’’
You are a selective entity replacement engine. You
need to perform entity replacement on the original
text.

Core Instructions:
1. Analyze the input text to identify replaceable
entities.
2. Randomly select approximately 50% of the
identified entities for substitution.
3. Replace the chosen entities with contextually
appropriate alternatives.
4. Maintain grammatical correctness and readability.
5. Output the modified version and a summary of
changes.

Workflow:

1. Entity Identification
   - Named entities (people, places, organizations)
   - Common nouns
   - Actions/verbs
   - Descriptors/adjectives

2. Replacement Rules:
   - Maintain the original part of speech.
   - Preserve sentence structure.
   - Ensure semantic coherence.
   - Keep consistent tense and number.
   - Replace only approximately 50% of the
   identified entities to retain the original
   context and flow.

3. Input Original Text:
   {original_text}

4. Output Format:
   - Modified Text:
     [text with approximately 30% replaced entities]

   - Changes Summary:
     - [Original Entity 1] -> [Replacement Entity 1]
     - [Original Entity 2] -> [Replacement Entity 2]
     - ...

Additional Instructions:

- Entity Selection:
  - After identifying all replaceable entities,
  calculate 30% of the total number.
  - Randomly select the calculated number of
  entities to replace.
  - Ensure that the selection is random to maintain
  variability across different texts.

- Replacement Constraints:
  - Do not replace entities that are crucial for the
   understanding of the text.
  - Avoid replacing more than 30% to prevent
  altering the original meaning significantly.
  - If the total number of replaceable entities is
  small, adjust the replacement percentage
  proportionally to avoid replacing too many.
’’’
```

D Example Appendix

The section to show EMPO examples 7. We will expand with more examples later...



Question: Describe the image in detail.

LLaVA-1.5: The image depicts a city street with a storefront for Hawkins Bazaar, a shop that sells Christmas decorations. The store is located on the side of the street, and there is a black and red sign on the side of the building. Several people can be seen walking along the sidewalk in front of the store. Some of them are closer to the storefront, while others are further away. There is also a clock visible on the side of the building, possibly indicating the time for passersby. Overall, the scene captures a lively city street with people going about their day and the festive storefront of Hawkins Bazaar.

EDPO (ours): The image features a city street with a storefront for Hawkins Bazaar prominently displayed on the side of a building. The shop has attractive decorations, including a Christmas snowflake window display, which makes it stand out. There are a few people visible on the sidewalk in front of the store, possibly window shopping or going about their daily activities. The overall scene gives off a commercial and vibrant city vibe, with the Hawkins Bazaar storefront being a notable element in the landscape.

Question: Is the man in the picture indoors or outdoors?

LLaVA-1.5: The man in the picture is indoors, specifically inside a boat.

EDPO (ours): It's not possible to determine with certainty from the image whether the man is indoors or outdoors. However, what we can see is that there are divers' gear and equipment present in the image, which suggests that the man might be in a location related to diving or water activities.

Question: Who in the picture is wearing trousers?

LLaVA-1.5: The man and woman in the picture are wearing pants.

EDPO (ours): The woman in the picture is wearing pants.

Figure 7: More Examples