Mitigating Large Vision Language Model Hallucinations via Entity-centric Multimodal Preference Optimization

Anonymous ACL submission

Abstract

Large Visual Language Models (LVLMs) have demonstrated impressive capabilities across tasks. However, their trustworthiness is often challenged by hallucinations. We attribute this issue to modality misalignment and the inherent hallucinations of Large Language Models (LLMs), which serve as the "brain" of LVLMs. Multimodal human preference alignment is a widely used approach to mitigate LVLM hallucinations. However, existing methods focus on response-level alignment while neglecting alignment at the image and instruction levels, leading to modality misalignment. For this, we propose Entity-centric Multimodal Preference Optimization (EMPO), which achieves better modality alignment than existing human preference alignment methods. Besides, to overcome the scarcity of high-quality multimodal preference data and help LVLMs mitigate hallucinations, we introduce a fine-grained multimodal preference data construction process that labels preferences at the entity level—all without requiring manual annotations. Experiments on two human preference datasets and five multimodal hallucination benchmarks demonstrate the effectiveness of EMPO, reducing hallucination rates by 80.4% on Object HalBench and 52.6% on MM HalBench, thereby enhancing the trustworthiness of LVLMs.

1 Introduction

Large Vision-Language Models (LVLMs) have recently demonstrated impressive capabilities in understanding and answering multimodal questions (Chen et al., 2023; Liu et al., 2023c, 2024b; Bai et al., 2023; Lu et al., 2024; Li et al., 2023a). An LVLM typically consists of a visual encoder that extracts image features and a large language model (LLM) that processes textual questions related to the image, generating accurate answers based on the provided visual context. To enhance LVLMs performance, most studies (Li et al., 2023a; Du et al., 2022; Lin et al., 2024) follow a two-step learning paradigm: (1) pretraining on large-scale image-text pairs to learn basic multimodal knowledge, and (2) fine-tuning on high-quality instruction datasets to improve responsiveness to user instructions (Liu et al., 2023c; Chen et al., 2024b; Wang et al., 2024c; Bai et al., 2023; Wang et al., 2024a). After that, LVLMs can learn to align large language models with visual encoders, enhancing their ability to comprehend and respond to user instructions involving multimodal inputs.

Figure 1: Causes of hallucinations. 1) Modality misalignment: the LVLM confuses entity semantics and provides the same answer for semantically conflicting questions. 2) LLM inherent hallucination: the response generated by the LVLM is entirely dependent on textual context, disregarding the image content.

Following the well-known hallucination problem in LLMs (Zhang et al., 2023; Li et al., 2023b; Dhuliawala et al., 2023), recent studies have also identified hallucinations in LVLMs (Li et al., 2023d; Liu et al., 2024a; Gunjal et al., 2024; Guan et al., 2024; Jiang et al., 2024b,a). Specifically, there are usually two causes of LVLM hallucinations. The first type is **modality misalignment** (Liu et al., 2024a; Lan et al., 2024), which arises from the modality gap between the visual encoder and the LLM in LVLM, resulting in mismatches between image content and semantic concepts. For instance, as shown in Figure 1, the LVLM (Liu et al., 2023c) identifies

the sign on the farmland LVLM (Liu et al., 2023c) correctly, but it mistakes its meanings as *"no parking"*, instead of *"no motor vehicles allowed"*. The second type is **LLM inherent hallucinations** (Lan et al., 2024). 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, *"car"* and *"road"* frequently co-occur in the LLM's pretraining corpus. As a result, the LVLM erroneously infers that whenever a *"road"* is present, a *"car"* must also be present, disregarding the image content. To address these issues, some recent approaches focus on reducing hallucinations through data denoising (Liu et al., 2023d; Yu et al., 2024a; Liu et al., 2023a; Hu et al., 2023), but this typically requires costly annotations. Some other studies (Yu et al., 2024c,b; Sun et al., 2023; Zhou et al., 2024; Wang et al., 2024b; Li et al., 2023c) try to align responses with human preferences, yet neglecting modality alignment between images and questions, which usually leads to mismatches between entity features and semantic concepts.

In this paper, we propose to mitigate the modality misalignment of LVLMs in two aspects: (1) **At the method level**, we propose Entity-centric Multimodal Preference Optimization (EMPO), a variant of DPO (Rafailov et al., 2024), which is an efficient human preference optimization method for LLMs. EMPO mitigates modality misalignment in LVLM by aligning human preferences across image, instruction, and response modalities. In addition, EMPO helps LVLM efficiently align image entities with semantic concepts through entity-grained preference weighting. (2) **At the data level**, we construct a fine-grained multimodal human preference dataset. As shown in Figure 2, we construct rejected examples that are similar to but different from the original data in three modalities: images, questions, and responses. These preference examples disrupt the co-occurrence relationships of entities with the original data, helping LVLM overcome LLM inherent hallucinations. Notably, our dataset expansion process does not rely on costly manual annotations and offers excellent scalability. The experimental results show that EMPO reduce hallucination rates by 80.4% on Object HalBench (Rohrbach et al., 2018) and by 52.6% on MM HalBench (Sun et al., 2023).

Our contributions are tri-fold: (1) We propose EMPO, the entity-centric multimodal preference optimization framework to mitigate the hallucination of LVLMs by aligning entity features with semantic concepts. (2) We introduce a multimodal fine-grained human preference dataset construction process that requires no manual annotation to overcome the scarcity of high-quality multimodal preference data. (3) Comprehensive experiments conducted on five widely-used benchmarks validate the effectiveness of the EMPO.

2 Related Work

Large Vision Language Model. Inspired by the success of LLMs (Achiam et al., 2023; Wu et al., 2023a; Touvron et al., 2023; Dao and Gu, 2024), recent research on LVLMs (Zhu et al., 2023; Chen et al., 2023; Liu et al., 2023c, 2024b; Bai et al., 2023; Liu et al., 2023b; Lu et al., 2024; Zhang et al., 2024; Li et al., 2023a) construct 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 recent 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). (2) **Instruction fine-tuning** by using instruction datasets to improve its 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; MiniGPT-v2 (Chen et al., 2023) employs unique task identifiers during fine-tuning to reduce instruction ambiguity.

Hallucination in LVLMs. Hallucinations in LVLMs refer to model outputs conflict with the images, instructions, or context (Du et al., 2022; Sun et al., 2023; Jiang et al., 2024c). Recent researches (Jiang et al., 2024b; Yin et al., 2023; Yu et al., 2024a; Liu et al., 2023a; Hu et al., 2023) typically classify hallucinations from two aspects, modality misalignment and LLM inherent hallucination. To mitigate the hallucinations, some work (Liu et al., 2023d; Yu et al., 2024a; Hu et al., 2023; Liu et al., 2023a) tries to filter long-tail and entity co-occurrence data, while usually suffers from the expensive labeling cost. Li et al. (2023e); Jiang et al. (2023); Tong et al. (2024); Cao et al. (2024); Jiang et al. (2024d,b) has recognized that modal misalignment is a significant cause of hal-

lucinations but ignores the LLM inherent hallucinations. Some research effectively reduces hallucinations by using post-processing methods, such as optimizing decoding strategies (Gao et al., 2024b; Huang et al., 2024; Yang et al., 2024; Gao et al., 2024a; Leng et al., 2024) and applying post-hoc corrections (Lee et al., 2023; Zhou et al., 2023; Yin et al., 2023) but introducing additional inference cost. In contrast, our EMPO leverages entity-level semantic alignment to help LVLM overcome LLM's inherent hallucinations, without requiring expensive manual annotations or additional inference costs.

Human Preference Alignment. Human preference alignment has been shown to be an effective method for mitigating hallucinations in LLMs (Naseem et al., 2024; Jiang et al., 2024c; Huang et al., 2023; Ji et al., 2024). In the LVLM area, LLaVA-RLHF (Sun et al., 2023) firstly propose to apply human preference alignment to reduce hallucinations in LVLMs, and establishes a foundational framework for multimodal preference data construction. RLHF-V (Yu et al., 2024b) demonstrates that fine-grained human preference alignment can improve the visual localization capabilities of LVLMs. RLAIF-V (Yu et al., 2024c) scores individual text segments and utilizes open-source LVLM for preference construction. POVID (Zhou et al., 2024) uses images with Gaussian noise to induce LVLMs to generate rejected preference examples. However, these methods focus solely on preferences at the response level, neglecting preferences related to the visual and question conditions. The recent work MDPO (Wang et al., 2024b) try to address image-conditional preference alignment, but it overlooks the instruction modality preference and only makes a preliminary attempt to address image modality preference. In contrast, our proposed EMPO incorporates preferences across all three modalities—image, instruction, and response—and leverages entity-centric preferences to enable LVLMs to align image content and semantic concepts more efficiently.

3 Method

This work includes multimodal preference alignment and fine-grained dataset construction. In Section 3.1, we introduce the preliminaries of human preference alignment and the response preference form of Direct Preference Optimization (DPO) (Rafailov et al., 2024) in the multimodal domain. In Section 3.2, we elaborate on how our EMPO framework aligns entity features with semantic concepts to address the modality misalignment issue. Section 3.3 details the construction of the high-quality, fine-grained preference dataset, designed to help LVLMs overcome the inherent hallucination issues of LLMs.

3.1 Preliminaries

To ensure LVLMs effectively focus on multiple modalities while aligning with human preferences, we adopt the Direct Preference Optimization (DPO) (Rafailov et al., 2024) for training. In the context of LLMs, human preference alignment involves two parts: the instruction and the response. Given an instruction q, LLM generates multiple candidate responses, and a reward model identifies the chosen response y_w that is superior to the rejected one y_l. DPO is one of the primary methods to achieve human preference alignment. It implicitly models the reward function in Reinforcement Learning from Human Feedback (RLHF) (Yu et al., 2024b) and directly optimizes the model parameters to maximize the difference between the reward $r(q, y_w)$ for the chosen response and the reward $r(q, y_l)$ for the rejected response. Specifically, given a policy model π_θ and a reward model π_{ref}, DPO formulate the reward function as

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

where $Z(q)$ is a partition function and β is a hyperparameter that controls the deviation from the reference model. DPO directly optimizes the policy model based on this implicit reward model to align with preference data,

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

In the context of multimodal, aligning human preference always includes three modalities (image, instruction, response) (Yu et al., 2024c,b; Sun et al., 2023; Zhou et al., 2024; Wang et al., 2024b; Li et al., 2023c). DPO minimizes a new objective conditioned on the image v and instruction q,

$$\begin{aligned} \mathcal{L}_{\text{response}} = -\log \sigma(\beta \log \frac{\pi_\theta\left(y_w \mid v, q\right)}{\pi_{\text{ref}}\left(y_w \mid v, q\right)} \\ - \beta \log \frac{\pi_\theta\left(y_l \mid v, q\right)}{\pi_{\text{ref}}\left(y_l \mid v, q\right)}). \end{aligned} \quad (3)$$

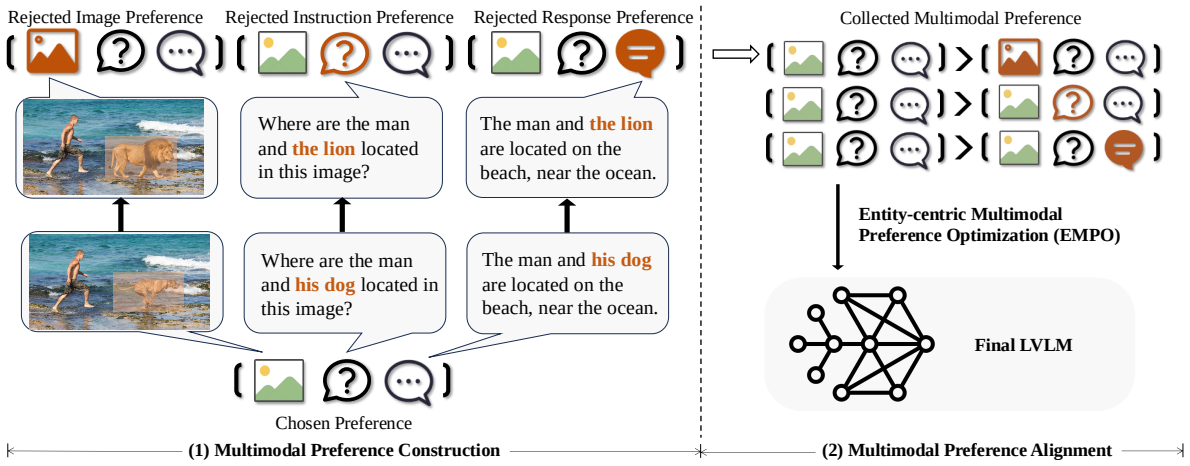

Figure 2: Illustration of our framework. (1) At the data level, we construct a fine-grained preference alignment dataset across three modalities: image, instruction, and response. (2) At the method level, we propose entity-centric multimodal preference optimization for aligning image contents with semantic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

3.2 Entity-centric Multimodal Preference Optimization

To assess the severity of hallucinations in LVLMs, we conducted a pilot experiment, evaluating inference performance on 200 preference examples from the POVID dataset (Zhou et al., 2024). Based on the results, we identified two types of errors in LVLM responses. (1) Conceptual confusion. Owing to modality misalignment, LVLMs may misinterpret semantic relationships between entities, leading to identical responses for conflicting user instructions. (2) Visual neglect. Consistent with PAI's findings (Liu et al., 2024c), when provided only textual context, LVLMs generated image-agnostic responses, indicating insufficient attention to visual content and over-reliance on textual cues due to LLM inherent hallucinations. Detailed examples are provided in Appendix A. Based on these observations, we propose the optimization objective for aligning the image and instruction modality preferences:

$$\mathcal{L}_{\text{image}} = -\log\sigma(\beta\log\frac{\pi_\theta\left(y \mid v_w, q\right)}{\pi_{\text{ref}}\left(y \mid v_w, q\right)} - \beta\log\frac{\pi_\theta\left(y \mid v_l, q\right)}{\pi_{\text{ref}}\left(y \mid v_l, q\right)}), \tag{4}$$

$$\mathcal{L}_{\text{instruction}} = -\log\sigma(\beta\log\frac{\pi_\theta\left(y \mid v, q_w\right)}{\pi_{\text{ref}}\left(y \mid v, q_w\right)} - \beta\log\frac{\pi_\theta\left(y \mid v, q_l\right)}{\pi_{\text{ref}}\left(y \mid v, q_l\right)}), \tag{5}$$

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{image}} + \mathcal{L}_{\text{instruction}} + \mathcal{L}_{\text{response}}, \tag{6}$$

where w represent chosen preference, l represent rejected preference, and v, q, y equal to v_w, q_w, y_w. $\mathcal{L}_{\text{image}}$ is the image preference loss, $\mathcal{L}_{\text{instruction}}$ is the instruction preference loss, $\mathcal{L}_{\text{response}}$ is the response preference loss from Equation 3.

In addition, to improve LVLMs' focus on entity features and address annotation ambiguity and low learning efficiency in the original DPO method (Wu et al., 2023b; Yu et al., 2024b), we assign higher weights to key entities in the image, instruction, and response modalities:

$$\log\pi(y \mid v, q) = (1-\alpha)\sum_{y_i \notin y_e}\log p\left(y_i \mid v, q, y_{<i}\right) + \alpha\sum_{y_i \in y_e}\log p\left(y_i \mid v, q, y_{<i}\right), \tag{7}$$

where α is a weighting hyperparameter, y_i is the i-th token of the response y. Larger α indicates that the corresponding token has a greater influence on preference. In this way, hallucination-related entities are emphasized, helping the LVLM receive stronger human preference feedback and ensuring its factual accuracy. The computation of entity weights is detailed in Section 3.3.

Overall, the LVLM optimized with our proposed EMPO can fully align entity features and semantic concepts, thereby mitigating hallucinations. As shown in Figure 3, the inference attention more effectively focuses on key information in the image and instruction tokens after training.

3.3 Fine-grained Preference Dataset Construction

As shown in Figure 1, the LVLM is affected by the LLM inherent hallucination, and outputs frequently co-occurring entities while overlooking the

actual content of the image. Moreover, most existing multimodal preference datasets contain only the response modality, which does not satisfy our EMPO optimization needs.

To fill this gap, we propose a fine-grained preference data construction method that compels LVLMs to focus on entity semantic. As shown in Figure 2, we remove and replace entities in three modalities to construct rejected preference samples. These preference examples disrupt the co-occurrence relationships of entities in pretraining corpus, helping LVLM learn fine-grained differences between chosen and rejected samples.

Image Preference Data Inspired by the phenomenon that LVLMs may generate non-existent objects, we introduce two strategies to construct image preference rejected samples q_l: entity cropping and entity replacement. Specifically, we first employ GPT4o-mini (Achiam et al., 2023) to identify entities in both the instruction and response. Next, we use an object detection model to locate these entities and classify them with GPT4o-mini. Finally, we apply a diffusion model (Rombach et al., 2022) to either remove 30% of the entities or substitute them with visually plausible alternatives, thereby generating an edited image as rejected image sample v_l. These selected entities will be weighted as described in Section 3.2. (1) **Entity Cropping**: Use a diffusion model to delete the chosen entities. The images with deleted entities serve as rejected preference samples to reduce the occurrence of non-existent entities generated by the LVLM. (2) **Entity Replacement**: Use a diffusion model to replace the chosen entities with incorrect but high-frequency entities, helping the LVLM overcome entity co-occurrence hallucinations. The prompts used for entity identification via GPT4o-mini are described in Appendix C.

Instruction Preference Data We employ GPT4o-mini (Achiam et al., 2023) to adapt the original instructions in terms of the selected entities in Section 3.3, thereby constructing rejected instructions q_l. We observe that the distribution of GPT-modified instructions differs from that of the original instructions, resulting in a decline in performance (Zhao et al., 2023). To address this problem, we also use GPT4o-mini to rewrite the chosen instructions, ensuring the rewritten instructions q_w retain the same meaning. The prompts used for constructing both chosen and rejected samples are described in the Appendix C.

Response Preference Data Our response preference data are constructed based on two existing datasets. The first is POVID (Zhou et al., 2024). We collect the rejected image preference sample v_l and the rejected instruction preference sample q_l from the two paragraphs above and use them as LVLM input to generate incorrect responses as the rejected response preference sample y_l. However, the final response preference triple still consists of y_w, q_w, and y_l. The second dataset is RLAIF-V (Yu et al., 2024c), in which we use MiniCPM-V2.5 (Yao et al., 2024) to compare two candidate answers generated by LLaVA-1.5 (Li et al., 2024), thereby establishing preference rankings. Notably, the complete preference construction process for RLAIF-V requires four iterations, whereas POVID achieves this in a single iteration.

4 Experiments

4.1 Experimental setups

Datasets. We conduct experiments on two datasets to demonstrate the generalizability of our method. These datasets differ in their construction methods, data volume, and data distribution.

(1) **POVID** (Zhou et al., 2024) dataset incorporates 17,000 examples randomly sampled from the LLaVA-Instruct-150K dataset (Liu et al., 2023c). Covering various types of tasks such as image captioning, simple Visual instruction responseing (VQA), and logical reasoning, POVID generates preferred responses by modifying the original responses using GPT-4V (OpenAI, 2023). These hallucinated responses introduce potential errors in areas like object co-occurrence, logical relationships between entities, and attribute descriptions.

(2) **RLAIF-V** (Yu et al., 2024c) dataset is an open-source feedback dataset created to improve the reliability of Multimodal Large Language Models (MLLMs). It includes high-quality instructions from various sources, such as MSCOCO (Lin et al., 2014a), ShareGPT-4V (Chen et al., 2024a), MovieNet (Huang et al., 2020), Google Landmark v2 (Weyand et al., 2020), VQA v2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), and TextVQA (Singh et al., 2019), collecting about 4,000 instructions in each training round to ensure comprehensive data coverage. Each instruction is paired with different candidate responses generated by open-source LVLMs and is detail-scored.

Metrics. We evaluate the methods from two perspectives: trustworthiness and helpfulness. The

Table 1: Main experimental results. We present our experimental results from two perspectives: trustworthiness and helpfulness. To demonstrate the scalability of our proposed method, we conduct the experiments on two datasets (POVID and RLAIF-V). "Annotation" represents model reliance on preference data annotation. "-" means unknown. The best and second best results are shown in **bold** and underlined respectively. The experimental data of general baselines hallucination baselines refer to RLAIF-V (Yu et al., 2024c).

Model	Size	Annotation	Object HalBench		MMHal Bench		AMBER		LLaVA Bench	MME	
			$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Score	Hall.↓	Acc.	F1.	Overall	Per.	Cog.
GPT-4V (OpenAI, 2023)	-	-	13.6	7.3	**3.42**	28.1	**83.4**	87.4	**93.1**	**1459.4**	**426.8**
QWEN-VL (Bai et al., 2023)	10B	×	40.4	20.7	2.76	38.5	81.9	86.4	71.9	1487.6	331.6
LLaVA-NeXT (Liu et al., 2024b)	34B	×	12.6	6.4	3.31	34.4	81.4	85.4	77.7	1531.1	295.6
VCD (Leng et al., 2024)	7B	×	48.8	24.3	2.12	54.2	71.8	74.9	65.8	1512.4	289.6
Silkie (Li et al., 2023c)	10B	×	27.1	13.4	3.19	32.3	82.2	**87.6**	73.2	1539.6	397.1
LLaVA-RLHF (Sun et al., 2023)	13B	Human	38.1	18.9	2.02	62.5	79.7	83.9	61.5	-	-
HA-DPO (Zhao et al., 2023)	7B	1-iter	39.9	19.9	1.98	60.4	75.2	79.9	67.2	-	-
POVID (Zhou et al., 2024)	7B	Human	40.4	19.1	2.08	56.2	82.9	87.4	62.2	1478.5	235.4
RLHF-V (Yu et al., 2024b)	7B	Human	12.2	7.5	2.45	51.0	72.6	75.0	51.4	1340.9	292.2
RLAIF-V (Yu et al., 2024c)	7B	4-iter	**8.5**	**4.3**	2.93	32.3	81.6	86.4	64.9	1366.3	297.5
MDPO (Wang et al., 2024b)	7B	1-iter	35.7	9.8	2.39	54.0	73.4	74.7	-	-	-
LLaVA 1.5 (Li et al., 2024)	7B	×	52.3	25.5	2.36	52.7	73.5	77.7	60.6	1496.7	297.5
+ DPO (POVID DataSet)	7B	1-iter	48.9	22.4	2.15	56.0	75.1	78.9	65.0	1494.0	300.0
+ Ours (POVID DataSet)	7B	1-iter	38.1	19.3	2.58	49.1	82.7	87.1	67.0	1483.9	296.8
+ DPO (RLAIF-V DataSet)	7B	4-iter	19.13	9.32	2.70	36.6	76.8	81.5	66.4	1356.7	299.3
+ Ours (RLAIF-V DataSet)	7B	4-iter	10.0	5.2	3.14	**25.0**	82.2	87.5	69.3	1365.8	286.8

former reflects the degree of hallucination, and the latter reflects the general ability of the method.

For trustworthiness, we evaluate on three benchmarks: (1) **CHAIR** (Rohrbach et al., 2018) is a widely adopted benchmark for evaluating entity hallucination in image captioning. It identifies hallucinations by comparing the entities mentioned in the model's output with the entities manually annotated in the COCO dataset (Lin et al., 2014b). However, CHAIR is a metric designed for traditional image captioning tasks and performs poorly when assessing LVLM tasks that include instructions. To enhance the stability of the evaluation, we follow (Yu et al., 2024b) and sample 300 examples from the CHAIR dataset, using eight different prompts to improve evaluation consistency. We report the sentence-level hallucination rate (i.e., the percentage of hallucinated sentences) and the entity-level hallucination rate (i.e., the percentage of hallucinated entities). (2) **MMHal-Bench** (Sun et al., 2023) evaluates model outputs from two aspects: hallucination rate and information richness. This benchmark uses GPT-4 to compare the model's outputs with human responses and multiple entity labels, providing five-level scores. (3) **AMBER** (Wang et al., 2023) is a multi-dimensional hallucination benchmark. We report the accuracy and F1 metric on discriminative tasks.

For helpfulness, we use two benchmarks: (1) **LLaVA Bench** (Liu et al., 2023c; Li et al., 2024) is a widely adopted benchmark for evaluating multimodal dialogue, detailed description, and complex reasoning capabilities. (2) **MME** (Fu et al., 2023) is a comprehensive benchmark specifically designed to evaluate LVLMs across ten perception subtasks and four cognition subtasks.

Baselines. We compared our model with state-of-the-art baselines of different types, including general baselines with strong performance, baselines that mitigate hallucinations, baselines that train LLaVA with human preference optimization, and a proprietary baseline. The primary baseline is training LLaVA with vanilla DPO.

(1) **General baselines.** We use LLaVA 1.5 (Li et al., 2024), Qwen-VL (Bai et al., 2023), and LLaVA-Next (Liu et al., 2024b) as general baseline representatives. These models have been pre-trained on large-scale multimodal data and fine-tuned on instruction datasets, demonstrating strong multimodal understanding capabilities. Moreover, We chose GPT-4V (OpenAI, 2023) as a reference to compare the performance gap between open-source models and closed-source commercial models.

(2) **Hallucination baselines.** Silkie et al. (Li et al., 2023c) construct a feedback dataset using GPT-4V, featuring various instructions and feedback sources. VCD (Leng et al., 2024)mitigates statistical bias in LVLMs by comparing probability distributions between the original and hallucinated inputs during decoding. LLaVA-RLHF (Sun

et al., 2023) transfers human feedback reinforcement learning from the text domain to the multimodal domain to align modal information.

(3) **Human Preference Learning based baselines** HA-DPO (Zhao et al., 2023) is the first work to apply DPO in the multimodal domain. mDPO (Wang et al., 2024b) avoids the problem of over-optimizing language preferences by optimizing image preferences. POVID (Zhou et al., 2023) proposes a method to adjust the image and text modalities in VLLMs using AI-generated preference differences.. RLHF-V (Yu et al., 2024b) uses high-quality, detailed human feedback to help large models learn precise behavior boundaries and eliminate hallucinations. RLAIF-V (Yu et al., 2024c) utilizes open-source models to generate high-quality preference data and resolves the offline issues of the DPO (Qi et al., 2024) algorithm by producing preferences through multiple iterative cycles.

Implementation Details. We implement EMPO based on the LLaVA-v1.5-7B framework (Li et al., 2024). The model uses CLIP-ViT (Radford et al., 2021) as the vision module and Vicuna (Zheng et al., 2023) (fine-tuned from LLaMA (Touvron et al., 2023)) as the LLM backbone. We trained the model for 4 epochs using deepspeed, which is an open-source library by Microsoft for efficient distributed training. We set a hyperparameter α of 0.9 and β of 0.5, an image resolution of 336, a learning rate of 5e-7, and a batch size of 8. The training was conducted on 8 A100 GPUs, taking 4 hours on the POVID dataset and approximately 12 hours on the RLAIF-V dataset.

4.2 Experimental Results

The main experimental results are reported in Table 1, from which we observe that: (1) EMPO is comparable to state-of-the-art performance in trustworthiness among open-source models, even outperforming commercial models like GPT-4V on some metrics. Using either POVID or RLAIF-V data, EMPO reduces the object hallucination rate of LLaVA 1.5 on Object HalBench by 26.2% and 80.4%, respectively. The reduction in hallucination rates is consistent across multiple benchmarks, including Object HalBench, MMHal-Bench, and AMBER. (2) EMPO reduces the trustworthiness of LVLM without significantly impairing its helpfulness. Specifically, EMPO results in a performance decrease of 7.9% on MME but an increase of 14.4% on LLaVA-Bench. (3) EMPO can reduce baseline hallucinations in both the POVID and RLAIFV datasets, which are entirely different, indicating that EMPO has excellent scalability.

Table 2: Ablation Studies on RLAIF-V datasets. "w/o image/instruction /response" indicates removing modality preferences for image, instruction, and response respectively. "w/o weighting" indicates removing the weights on key entities.

Model	Object HalBench		MMHalBench	
	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Score	Hall.↓
Ours	10.0	5.2	**3.14**	**25.0**
w/o visual	10.8	5.4	2.78	34.4
w/o instruction	**8.7**	**4.2**	2.96	31.2
w/o response	9.7	4.9	2.91	31.0
w/o weighting	8.9	4.6	3.05	28.1

4.3 Detailed Analysis

We analyze the following research questions: (1) How different components contribute to the performance of EMPO and illustrate how EMPO enhances overall performance. (2) Can EMPO align different modalities, enabling LVLM to align image content with semantic concepts, focusing on the correct entities? (3) Can EMPO help LVLM overcome the LLM inherent hallucinations? (4) Does EMPO achieve better performance across different preference datasets?

Ablation Studies As shown in Table 2, we conduct ablation studies to evaluate the effectiveness of different components in our approach, including the fine-grained preference dataset and entity-centric preference alignment. We report the average changes across four hallucination evaluation metrics to comprehensively validate the effects of the ablation experiments, rather than relying on a single metric. (1) we test the necessity of aligning with human preferences across three modalities: image, instruction, and response. Removing the image/instruction/response modality significantly increases the hallucination rate by 22.9%/9.5%/11.9%. The complete three-modal preference alignment exhibits the best performance, indicating that integrating image, instruction, and response enables a more comprehensive capture of human preferences. (2) We also examine the function of fine-grained preferences. Removing the weight of the key entities increased the hallucination rate by 2.6%. Therefore, We highlight that fine-grained preferences can help LVLM better locate entities.

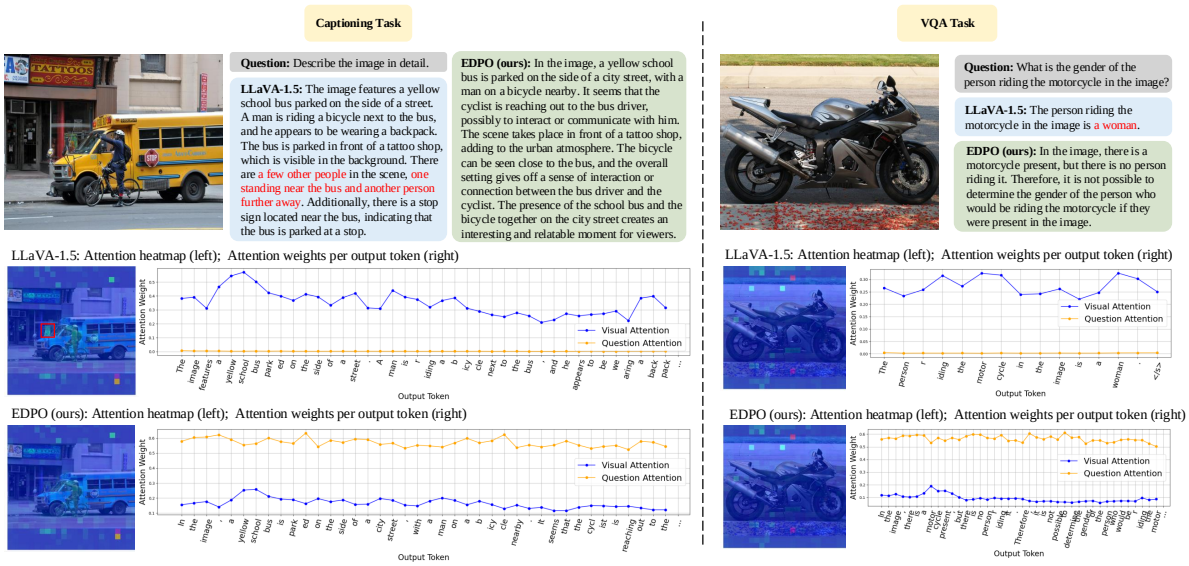

Figure 3: Illustration of hallucination correction by our proposed EMPO at different tasks. Hallucinated tokens from LVLM's regular decoding are highlighted in red. The red box region in the attention heatmap is labeled with the hallucination entity removed by EMPO. The blue line represents the total attention from each output token to the image tokens, while the orange line represents the total attention from each output token to the question tokens.

Modality Alignment Analysis As shown in Figure 3, We compare the performance of EMPO and LLaVA 1.5 on the tasks of Image Captioning and Visual Question Answering (VQA). The visual attention heatmap and blue line show the variation in the image attention weights assigned to output tokens during LVLM inference, while the orange line shows the variation in the attention weights given to the instruction. The image captioning example demonstrates that LLaVA 1.5 focused on an incorrect image feature (the red region in the attention heatmap), resulting in the output of a non-existent entity (a few other people). Our EMPO corrects this, indicating a strong consistency between the image content and semantic concepts. The VQA example shows that LLaVA 1.5 was influenced by the prompt and produced an affirmative answer even when the corresponding image feature was absent. In contrast, EMPO overcomes the inherent hallucination of the LLM. Furthermore, our EMPO assigns higher attention weights to both the image and the instruction, suggesting that it focuses on these inputs better than LLaVA 1.5. One possible explanation is that by comparing real data with generated negative data and mitigating the internal hallucination patterns, EMPO redirects the LVLM's attention, causing it to pay more attention to the image and question tokens. Due to space constraints, more inference examples for captioning and VQA tasks will be presented in Appendix B.

Scalabiliy As shown in Table 1, EMPO significantly reduces baseline hallucinations across both the POVID and RLAIFV datasets, which differ fundamentally in structure and content. This clear ability to perform effectively on datasets with distinct characteristics highlights the strong adaptability and scalability of EMPO. Moreover, as shown in Figure 3, its consistent performance on real test samples suggests that it can address real-world challenges involving heterogeneous data. The ability to generalize its performance to diverse datasets further underscores its potential applicability across a wide range of domains, proving its robustness in handling varied data distributions.

5 Conclusion

This paper addresses the LVLM hallucination problem from two perspectives: modality misalignment and LLM inherent hallucination. At the method level, we propose an entity-centric multimodal preference optimization method to help LVLM align entity features with semantic concepts, enhancing its trustworthiness. On the data side, we construct fine-grained preference data to assist LVLM in overcoming the inherent hallucination of LLMs. Extensive experiments across multiple benchmarks demonstrate that our method effectively reduces LVLM hallucinations while preserving its comprehensive capabilities. For future work, we plan to further explore the construction of preferences that involve more complex interconnected entities.

Limitations

The limitation of this study is that its investigation into hallucination issues is confined to entities. Currently, our dataset is constructed solely by deleting and replacing entities; we have not delved into entity attributes or inter-entity relationships, nor have we explored hallucinations caused by non-entity factors. We propose the following directions for future research: (1) Explore the construction of preferences based on entity attributes and inter-entity relationships. (2) Investigate hallucinations caused by non-entity factors and construct corresponding preference samples.

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, et al. 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, 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, 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*.

Tri Dao and Albert Gu. 2024. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*.

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, et al. 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. 2024a. Fact: Teaching mllms with faithful, concise and transferable rationales. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 846–855.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. 2024b. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9096–9105.

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, et al. 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*, volume 38, pages 18135–18143.

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.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer.

Shijia Huang, Jianqiao Zhao, Yanyang Li, and Liwei Wang. 2023. Learning preference model for llms via automatic preference data generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9187–9199.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. *Advances in Neural Information Processing Systems*, 36.

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.

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*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pretraining 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, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

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. 2023c. 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. 2023d. 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, et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

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.

Tahira Naseem, Guangxuan Xu, Sarathkrishna Swaminathan, Asaf Yehudai, Subhajit Chaudhury, Radu Florian, Ramón Fernandez Astudillo, and Asim Munawar. 2024. A grounded preference model for llm alignment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 151–162.

OpenAI. 2023. Gpt-4v system card.

Biqing Qi, Pengfei Li, Fangyuan Li, Junqi Gao, Kaiyan Zhang, and Bowen Zhou. 2024. Online dpo: Online direct preference optimization with fast-slow chasing. *arXiv preprint arXiv:2406.05534*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 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.

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.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. 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.

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. 2024a. Vigc: Visual instruction generation and correction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5309–5317.

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, et al. 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.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023a. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023b. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033.

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, et al. 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, et al. 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, et al. 2024c. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, et al. 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, et al. 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, et al. 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 Experimental Supplement

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 often struggle to accurately interpret semantic relationships between entities, leading to concept confusion. For example, the models frequently generated identical or highly similar responses to user instructions that were semantically conflicting or conceptually distinct. This 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 tended to generate image-agnostic responses that disregarded 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 applicability of LVLMs in real-world tasks that require both textual and visual reasoning.

B Example Appendix

The section to show EMPO examples.

C Prompt Appendix

The section is to describe the prompt for identifying entities and the prompt for rewriting chosen instruction and rejected instruction.

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
```

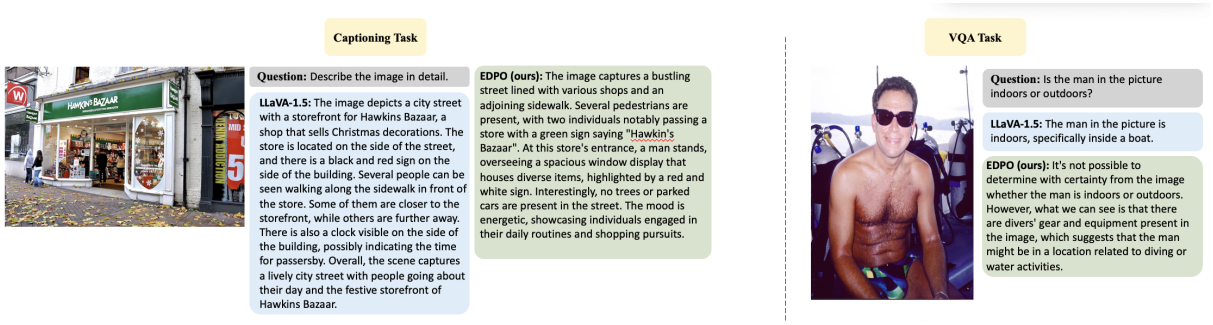

Figure 4: More Examples

```
  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.
'''
```

The prompt for rewriting 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.'''
'''
```