V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization

Anonymous ACL submission

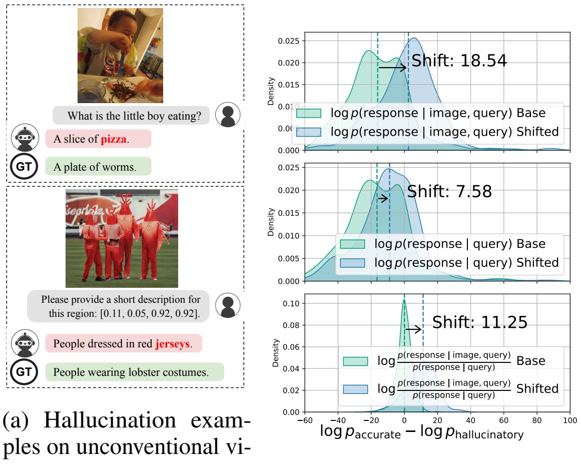
Abstract

Large vision–language models (LVLMs) suffer from hallucination, resulting in misalignment between the output textual response and the input visual content. Recent research indicates that the over-reliance on the Large Language Model (LLM) backbone, as one cause of the LVLM hallucination, inherently introduces bias from language priors, leading to insufficient context attention to the visual inputs.

We tackle this issue of hallucination by mitigating such over-reliance through preference learning. We propose Vision-guided Direct Preference Optimization (V-DPO) to enhance visual context learning at training time. To interpret the effectiveness and generalizability of V-DPO on different types of training data, we construct a synthetic dataset containing both response- and image-contrast preference pairs, compared against existing human-annotated hallucination samples. Our approach achieves significant improvements compared with baseline methods across various hallucination benchmarks. Our analysis indicates that V-DPO excels in learning from image-contrast preference data, demonstrating its superior ability to elicit and understand nuances of visual context.

1 Introduction

Recent advancements in Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023; Chiang et al., 2023; OpenAI, 2023) have catalyzed the evolution of Large Vision–Language Models (LVLMs) (Liu et al., 2023c,b; Dai et al., 2023; Anil et al., 2023) in understanding and reasoning across visual and textual modalities. Despite their impressive performance on various vision–language tasks, existing LVLMs still struggle with the issue of *hallucination*, where the model outputs are not factually grounded in the input visual contents (Rohrbach et al., 2018; Li et al., 2023; Gunjal et al., 2024; Liu et al., 2024). Hallucination in LVLMs refers to non-existing or erroneous descriptions of visual contents, such as objects, attributes, and relationships, which is especially challenging to understanding unconventional images, as shown in Figure 1a.


What is the little boy eating?

A slice of pizza.

A plate of worms.

Please provide a short description for this region: [0.11, 0.05, 0.92, 0.92].

People dressed in red jerseys.

People wearing lobster costumes.

(a) Hallucination examples on unconventional visual contents.

(b) Distribution gaps.

Figure 1: (a) Hallucination examples in visual question answering and region descriptions and (b) the model discriminative ability on the accurate and hallucinatory samples represented by difference in log-likelihoods.

The phenomenon of hallucination in LVLMs can be attributed to the integration of pre-trained LLMs in the architecture. Recent works reveal that this issue is closely tied to insufficient context attention, where the model prioritizes language patterns and focuses on partial tokens rather than fully grounding the generated content in both visual and textual context (Lee et al., 2023; Wang et al., 2024). To mitigate the over-reliance on language priors, many efforts have been devoted to decoding optimization with penalties on over-trust candidates (Huang et al., 2023) or a focus on visual uncertainty (Chen et al., 2024). However, these methods require increased inference time and specific infrastructure designs (Lee et al., 2023), obstructing their generalizability and scalability across diverse data domains and sizes. In contrast, our study explores training strategies to alleviate the over-reliance on language priors via preference learning, enhancing visual understanding to mitigate hallucination in LVLMs.

Given the difference in the likelihoods between accurate and hallucinatory samples on vision-conditioned $p(\text{response} \mid \text{image}, \text{query})$ and textual-only $p(\text{response} \mid \text{query})$ distributions, Figure 1b illustrates the shifts of this difference after aligning the model with hallucination-free data via preference learning. Before alignment, the textual-only distributions dominate the model decision on determining accurate samples as preferred compared to hallucinatory ones, reflected by the distributions (in green) of the same shape for both probabilities. This dominance in pairwise preference illustrates the over-reliance on language priors in LVLMs, which is especially crucial for unseen images in training (*e.g.*, Figure 1a), limiting the model generalizability across different data. Motivated by this challenge, we propose Vision-guided Direct Preference Optimization (V-DPO), a vision-specific variant of Direct Preference Optimization (DPO) (Rafailov et al., 2023), to employ visual guidance during preference learning for hallucination mitigation in LVLMs. We adapt Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) to integrate the visual guidance into the optimization target, inspired by its effectiveness in improving the specificity of model generations tailored for specific contents (Sanchez et al., 2023; Kornblith et al., 2023). To assess the generalizability of V-DPO, especially on unconventional contents, we construct a synthetic dataset containing both response-contrast and image-contrast preference pairs, compared against existing human-annotated preferences such as RLHF-V (Yu et al., 2023). Our approach exhibits significant and stable performance improvements through extensive experiments on various hallucination benchmarks. Further analysis of the distribution shifts from training demonstrates the effectiveness of V-DPO in mitigating the over-reliance on language priors on both image- and response-contrast data.

2 Related Work

Hallucination has emerged as a significant challenge to model reliability and generalizability in LVLM development. To alleviate hallucinated content, existing works can be divided as following two directions. The first focuses on post-processing approaches, including post-hoc corrections (Zhou et al., 2023; Yin et al., 2023; Lee et al., 2023) and specialized decoding (Huang et al., 2023; Chen et al., 2024). However, these methods often require increased inference time, obstructing their generalizability and scalability across diverse data domains and sizes (Bai et al., 2024).

The second line of work attempts to collect hallucination-aware data to mitigate hallucination in LVLMs through preference optimization leaning toward hallucination-free outputs. For example, Sun et al. (2023) and Yu et al. (2023) adapt the Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) paradigms in LLMs, respectively, to align LVLMs with hallucination-aware human preferences. Zhao et al. (2023) and Sarkar et al. (2024) propose data augmentation pipelines to construct (accurate, hallucinatory) preference pairs for contrastive tuning. Our work mitigates hallucination in the context of preference optimization with not only augmented data including both response- and image-contrast preference pairs, but also a vision-specific optimization target to enhance visual understanding.

3 Background and Motivations

We explore strategies to enhance visual understanding in LVLM preference optimization. Our framework starts from a supervised fine-tuned (SFT) model, obtained by jointly training a visual encoder and a pre-trained LLM via visual instruction tuning (Liu et al., 2023c). Specifically, we incorporate visual guidance by integrating Classifier-Free Guidance (CFG) into vanilla DPO.

3.1 Preference Optimization for LVLMs

We consider a policy LVLM π_θ parameterized by θ. For a vision-conditioned text generation task, given an input image $v \sim \mathcal{I}$ and a textual query $x \sim \mathcal{P}$, we optimize for the KL-constrained reward maximization objective:

$$\begin{aligned}\max_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}, y\sim\pi}\Big[& r(v,x,y) \\ &-\beta \mathbb{D}_{\mathrm{KL}}\left[\pi(y \mid v,x) \parallel \pi_{\mathrm{ref}}(y \mid v,x)\right]\Big]\end{aligned} \tag{1}$$

under reward function $r(v,x,y)$ and reference model π_{ref}. DPO solves the optimal policy as:

$$\pi_r(y \mid v,x) = \frac{\pi_{\mathrm{sft}}(y \mid v,x)\exp\left(\frac{1}{\beta} r(v,x,y)\right)}{Z(v,x)} \tag{2}$$

for all image–query pairs $(v, x) \sim \mathcal{I} \times \mathcal{P}$, where $Z(v, x) = \sum_y \pi_{\text{sft}}(y \mid v, x) \exp\left(\frac{1}{\beta} r(v, x, y)\right)$ is the partition function.

Rearranging Eq. 2, we get the ground-truth reward model with the corresponding optimal policy. Given a response-contrast preference dataset $\mathcal{D}_y = \{v^{(k)}, x^{(k)}, y_w^{(k)}, y_l^{(k)}\}_{k=1}^N$ where y_w is preferred over y_l, DPO uses Bradley–Terry model (Bradley and Terry, 1952) to derive the objective as:

$$\mathcal{L}^{y}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(v,x,y_w,y_l)\sim\mathcal{D}_y} \log \sigma(\beta u^{y_w,y_l}_{\pi_\theta}) \tag{3}$$

where $u^{y_w,y_l}_{\pi_\theta} = \log \frac{\pi_\theta(y_w|v,x)}{\pi_{\text{ref}(y_w|v,x)}} - \log \frac{\pi_\theta(y_l|v,x)}{\pi_{\text{ref}(y_l|v,x)}}$ indicates the implicit reward corresponding to π_θ.

Enlightened by *contrast sets* (Gardner et al., 2020; Shen et al., 2023), we construct an image-constrast dataset $\mathcal{D}_v = \{v_w^{(k)}, v_l^{(k)}, x^{(k)}, y^{(k)}\}_{k=1}^M$ to enhance visual understanding. With $u^{v_w,v_l}_{\pi_\theta} = \log \frac{\pi_\theta(y|v_w,x)}{\pi_{\text{ref}}(y|v_w,x)} - \log \frac{\pi_\theta(y|v_l,x)}{\pi_{\text{ref}}(y|v_l,x)}$, we have:

$$\mathcal{L}^{v}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(v_w,v_l,x,y)\sim\mathcal{D}_v} \log \sigma(\beta u^{v_w,v_l}_{\pi_\theta}) \tag{4}$$

3.2 Classifier-Free Guidance in LLMs

CFG was originally proposed in the context of conditioned diffusion models (Dhariwal and Nichol, 2021). Given a noisy image y and a class condition c, the model predicts probability likelihood $\hat{p}$ for the conditioned step-wise sample $\hat{\pi}_\theta(y \mid c) \propto \pi_\theta(y) \cdot \pi_\phi(c \mid y)^\gamma$, where $\gamma > 0$ controls the guidance strength from the classifier π_ϕ. Ho and Salimans (2022) observe that the guidance can be offered without a classifier:

$$\hat{\pi}_\theta(y \mid c) \propto \pi_\theta(y) \cdot \pi_\theta(c \mid y)^\gamma \propto \frac{\pi_\theta(y \mid c)^\gamma}{\pi_\theta(y)^{\gamma-1}} \tag{5}$$

Given a textual completion $\mathbf{y} = \{y_i\}_{i=1}^N$ and a conditional prompt or image c, we can extend CFG to autoregressive models as $\hat{\pi}_\theta(\mathbf{y} \mid c) \propto \frac{\pi_\theta(\mathbf{y}|c)^\gamma}{\pi_\theta(\mathbf{y})^{\gamma-1}} \propto \prod_{i=1}^N \frac{\pi_\theta(y_i|y_{<i},c)^\gamma}{\pi_\theta(y_i|y_{<i})^{\gamma-1}}$. Previous works show that CFG increases the specificity of the generation to be more pertinent toward the prompt (Sanchez et al., 2023) or image (Kornblith et al., 2023). Enlightened by this insight, we apply CFG in LVLM preference optimization to enhance the importance of visual context. This employment is non-trivial considering the dynamics in the training process, which we will detail next.

4 Vision-Guided Preference Learning

In this work, we focus on mitigating hallucinations in LVLMs caused by insufficient context attention to visual information. We propose Vision-guided Direct Preference Optimization (V-DPO) to enhance visual understanding on both response- and image-contrast preference data.

4.1 Vision-Guided DPO

Our V-DPO approach builds on the insight that CFG-modified distribution produces more condition-specific generation than vanilla decoding. As we will detail next, our core contribution originates from a vision-specific term in the reward maximization objective of DPO.

V-DPO Objective. We start with the definition of visual guidance in the context of LVLMs. Following Eq. 5, we apply CFG to vision-conditioned text generation:

$$\hat{\pi}_\theta(y \mid v, x) \propto \pi_\theta(y \mid x) \left(\frac{\pi_\theta(y \mid v, x)}{\pi_\theta(y \mid x)}\right)^\gamma \tag{6}$$

where $\frac{\pi_\theta(y|v,x)}{\pi_\theta(y|x)}$ is the guidance from the visual context v to increase the specificity of the response y toward the image, given the input query x. We integrate this term as an additional target to optimize in Eq. 1. Our result vision-enhanced reward maximization objective is then:

$$\begin{aligned} &\max_\pi \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}, y\sim\pi} \Big[r(v, x, y) \\ &-\beta \mathbb{D}_{\text{KL}} \left[\pi(y \mid v, x) \parallel \pi_{\text{ref}}(y \mid v, x)\right] \\ &+\alpha \mathbb{D}_{\text{KL}} \left[\pi(y \mid v, x) \parallel \pi(y \mid x)\right] \Big] \end{aligned} \tag{7}$$

where $\alpha > 0$ controls the weight of the visual guidance to optimize. Solving the optimal solution π_r to the above objective, we have:

$$\begin{aligned} &\pi_r(y \mid v, x)^\gamma / \pi_r(y \mid x)^{\gamma-1} \\ =&\pi_r(y \mid v, x) \left(\frac{\pi_r(y \mid v, x)}{\pi_r(y \mid x)}\right)^{\gamma-1} \\ \propto&\frac{1}{Z(v, x)} \pi_{\text{sft}}(y \mid v, x) \exp\left(\frac{1}{\beta} r(v, x, y)\right) \end{aligned} \tag{8}$$

where $\gamma = 1 - \frac{\alpha}{\beta}$. Unlike inference-time CFG, we decrease $\gamma < 1$; *i.e.*, increasing $\alpha > 0$, to strengthen the guidance of visual context during training. We detail the complete derivations in Appendix A. Although only a proportional relationship holds here (as $\pi_r(y \mid v, x)^\gamma / \pi_r(y \mid x)^{\gamma-1}$ is an unnormalized probability distribution), we can still obtain the reward difference of a preference pair using the Bradley–Terry model. Similar to Eqs. 3 and 4, we derive our policy objective as:

$$\mathcal{L}_{\text{VDPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(w,l)\sim\mathcal{D}} \log \sigma(\beta u^{w,l}_{\pi_\theta}) \tag{9}$$

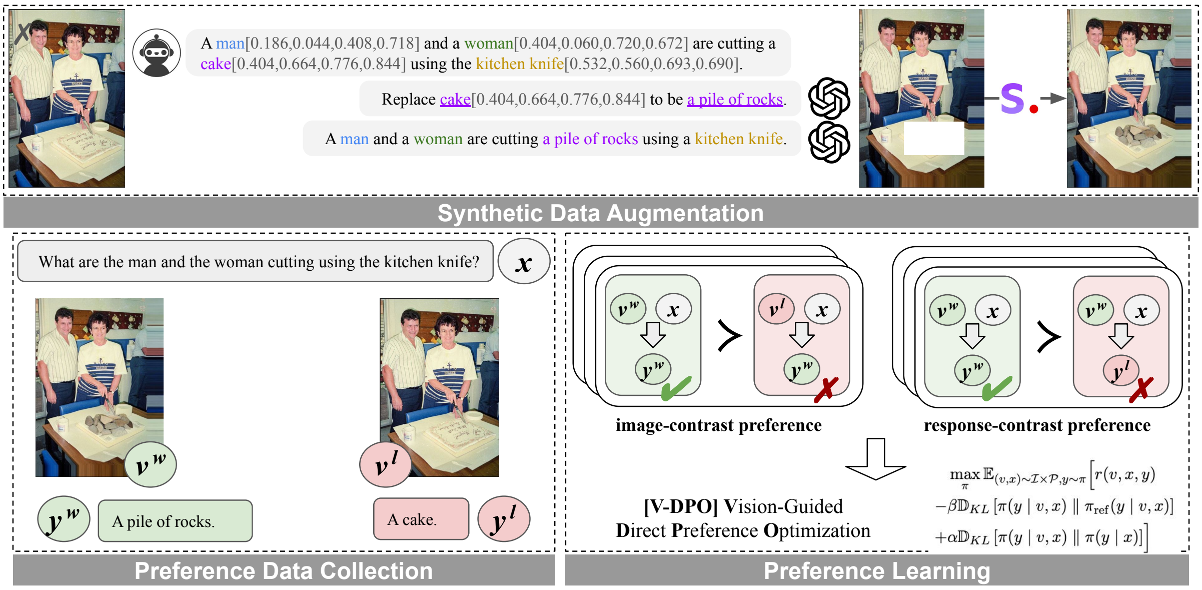

A man[0.186,0.044,0.408,0.718] and a woman[0.404,0.060,0.720,0.672] are cutting a cake[0.404,0.664,0.776,0.844] using the kitchen knife[0.532,0.560,0.693,0.690].
Replace cake[0.404,0.664,0.776,0.844] to be a pile of rocks.
A man and a woman are cutting a pile of rocks using a kitchen knife.
S.
Synthetic Data Augmentation
What are the man and the woman cutting using the kitchen knife?
x
v^w
v^l
y^w
A pile of rocks.
A cake.
y^l
Preference Data Collection
image-contrast preference
response-contrast preference
[V-DPO] Vision-Guided
Direct Preference Optimization
$\max_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}, y\sim\pi}\Big[r(v,x,y)$
$-\beta\mathbb{D}_{KL}\left[\pi(y \mid v,x) \parallel \pi_{\mathrm{ref}}(y \mid v,x)\right]$
$+\alpha\mathbb{D}_{KL}\left[\pi(y \mid v,x) \parallel \pi(y \mid x)\right]\Big]$
Preference Learning

Figure 2: Outline of our preference data construction and vision-guided preference learning framework. In the stage of Synthetic Data Augmentation, we utilize LVLMs, LLMs, and Stable Diffusion to manipulate images automatically. We formulate the generated samples into image- and response-contrast pairs for preference learning via our Vision-guided DPO approach.

where $\mathcal{D} = \mathcal{D}_y \cup \mathcal{D}_v$ and $u_{\pi_\theta}^{w,l} = f_\theta^w - f_\theta^l$, using the shorthand $f_\theta(v,x,y) = \log \frac{\pi_\theta(y|v,x)\hat{\varphi}_\theta(v,x,y)}{\pi_{\mathrm{ref}(y|v.x)}}$ with $\varphi_\theta(v,x,y) = \left(\frac{\pi_\theta(y|v,x)}{\pi_\theta(y|x)}\right)^{\gamma-1}$ controlling the strength of visual guidance.

Implementation of Visual Guidance. In Eq. 9, we disable gradient backpropagation on $\varphi_\theta(v,x,y)$ to maintain a stable textual-only distribution $\pi_\theta(\cdot \mid x)$ during training. This aims to provide a reliable reference value to calculate the visual guidance. We further discuss the choice of $\hat{\pi}(\cdot \mid x)$ in Section 5.3. Following the implementation of Liu et al. (2023c), we pass zeroes in place of the conditioning visual context to get the textual-only distribution:

$$\hat{\pi}_\theta(\cdot \mid x) = \hat{\pi}_\theta(\cdot \mid \mathbf{0}, x) \tag{10}$$

With the integration of visual guidance, we modify $\pi_\theta(y \mid v,x)$ in vanilla DPO to be a non-normalized probability distribution, $\pi_\theta(y \mid v,x)\hat{\varphi}_\theta(v,x,y)$. Empirically, this can progressively decrease the effect of visual guidance as the visual-conditioned and unconditioned distributions diverge from each other through training. To mitigate this problem, we follow Kornblith et al. (2023) to normalize it as:

$$\begin{aligned}&\pi_\theta(\cdot \mid v,x)\hat{\varphi}_\theta(v,x,\cdot)\\ &\propto \phi\left(h_\theta(v,x) + (\gamma-1)\left(\hat{h}_\theta(v,x) - \hat{h}_\theta(\mathbf{0},x)\right)\right)\end{aligned} \tag{11}$$

where h_θ are the generated logits and $\phi(\cdot)$ is the softmax function for normalization. Note that since the increase of divergence between the distribution $\pi_\theta(\cdot \mid v,x)$ and $\pi_\theta(\cdot \mid x)$ can lead to a larger exponential sum in softmax, the normalization thus gradually inflates the effect of visual guidance during training. We analyze the potential impacts of the guidance inflation in Section 5.4.

4.2 Constructing Contrast Images

As discussed in Section 3.1, we augment the preference data with image-contrast pairs to enhance visual understanding via preference learning. The construction of contrastive image pairs aims to bolster the visual understanding ability to discern nuanced visual differences between similar images. Specifically, we manipulate images by replacing conventional items with unconventional ones, considering the limited capability of LVLMs to understand weird images (Guetta et al., 2023). This section details the automatic construction process we use to collect image-contrast preference data.

Proposing Replacement Elements. Given an image from an existing dataset, we extract object-level information using LVLMs and generate detailed captions with objects grounded in respective positions in the image. Based on the layout-grounded descriptions, we employ LLMs to propose replacements for visual elements, thereby creating unexpected scenarios by leveraging their imaginative capability (Gómez-Rodríguez and Williams, 2023). Figure 2 shows an example element replacement proposed by ChatGPT. To enhance the interpretability of this automatic process, we require LLMs to supply a reasonable explanation of the replacement's unexpectedness (*cf.* Appendix C.1 for prompts and examples). We collect multiple replacements for each image, which are used to guide image generation next.

Image Editing and Filtering. Given a designated region in a source image, we use a generative model to edit via image inpainting (Lugmayr et al., 2022). Particularly, we utilize Denoising Diffusion Probabilistic Models (DDPMs) as the image inpainter, considering their superior generation quality (Dhariwal and Nichol, 2021). Empirically, the imperfections of the LLM and the generative model can result in a significant distribution gap between the generated images and the original real ones, introducing noise and bias into the synthetic data. To address this issue, we use CLIPScore (Hessel et al., 2021) to refine our data by filtering out edits that do not align well with the corresponding replacement prompts. Specifically, we approve an edited image v_i only if it achieves the highest CLIPScore with the intended textual prompt c_i in comparison with similar text–image pairs generated in our pipeline:

$$\begin{aligned} c_i &= \arg\max_{c} \text{CLIPScore}(c, v_i) \\ v_i &= \arg\max_{v} \text{CLIPScore}(c_i, v) \end{aligned} \tag{12}$$

Finally, we combine our image-contrast pairs with conventional response-contrast ones to construct our preference data for V-DPO. See Appendix C for a full construction pipeline for different types of preference data.

5 Experiments

We now assess V-DPO across various multimodal hallucination benchmarks. To interpret how V-DPO improves visual understanding, we compare performance using various preference data.

5.1 Setup

We choose LLAVA-V1.5-7B (Liu et al., 2023b) as our initial SFT model and conduct preference learning with full fine-tuning. Our synthetic augmented data contains 5K response- and image-contrast preference pairs, compared against the human-annotated response-contrast data RLHF-V (5K) (Yu et al., 2023) of equal size.

Benchmarks. We evaluate our approach on four hallucination benchmarks: (1) POPE (Li et al., 2023) on object hallucination with discriminative tasks; (2) AMBER (Wang et al., 2023) containing both generative and discriminative tasks on object, attribute, and relation hallucination; (3) HallusionBench (Liu et al., 2023a) assessing visual illusion and knowledge hallucination with systematically structured discriminative tasks; and (4) MMHal-Bench (Sun et al., 2023) covering different question types and object topics. We also conduct general-purpose evaluation on MMBench (Xu et al., 2023) across various multimodal tasks in Appendix D.

Baselines. We compare our method against the initial SFT model and vanilla DPO as the fundamental and strengthened baselines, respectively. We also consider Hallucination-Aware Direct Preference Optimization (HA-DPO) (Zhao et al., 2023) as a variant of DPO baseline trained on 16K style-consistent hallucination sample pairs.

5.2 Main Results

We compare V-DPO with vanilla DPO methods across various hallucination benchmarks to show the effectiveness and stability of our approach.

POPE. Table 1 compares model performance (F1 score) and tendency to answer "yes" (Yes Ratio) on POPE. V-DPO outperforms the SFT and vanilla DPO baselines on random sets and more challenging tasks such as the adversarial scenario. Furthermore, V-DPO significantly increases the F1 scores from 85.98 to 86.92 and 87.22 trained on synthetic and human-annotated data, respectively, with mitigated bias in yes ratios 47.43% and 48.66%, compared to 44.22% and 47.88% of vanilla DPO. This suggests that V-DPO achieves better hallucination performance while mitigating the over-reliance on language priors with visual guidance.

AMBER. In Table 3, our approach achieves significant improvements on both AMBER's generative and discriminative tasks. For CHAIR scores, we observe an absolute improvement of 2.2 from 7.8 to 5.6 when applying V-DPO to the human-annotated data RLHF-V. Compared to vanilla DPO, we observe further improvements due to our method on most metrics in both synthetic and human-annotated scenarios. Notably, with only 5K preference pairs collected via synthetic generation, V-DPO outperforms HA-DPO trained on 16K preference pairs , with an absolute increase of 3.7

Approach	**F1 Score**				**Yes Ratio**
	F1$_{\text{R}\uparrow}$	**F1**$_{\text{P}\uparrow}$	**F1**$_{\text{A}\uparrow}$	**F1**$_{\uparrow}$	
SFT	89.69	86.83	81.80	85.98	54.20
HA-DPO	**90.25**	87.81	82.54	86.87	51.03
Synthetic Augmented Data					
DPO	88.34	87.05	83.96	86.42	44.22
V-DPO	89.57	87.62	83.77	86.92$_{\uparrow 0.94}$	47.43
RLHF-V					
DPO	89.69	87.81	84.03	87.12	47.88
V-DPO	89.90	**87.91**	**84.05**	**87.22**$_{\uparrow 1.24}$	48.66

Table 1: Result comparison (F1 score) on POPE including splits of random (R), popular (P), and adversarial (A) scenarios. We report Yes Ratio (%) to compare the biased tendency of different models.

Approach	**Accuracy**		
	qAcc$_{\uparrow}$	**fAcc**$_{\uparrow}$	**aAcc**$_{\uparrow}$
SFT	13.19	20.23	48.16
Synthetic Augmented Data			
DPO	21.97	20.52	**55.52**
V-DPO	**22.20**$_{\uparrow 9.01}$	**21.10**	55.31
RLHF-V			
DPO	16.70	20.81	51.31
V-DPO	17.36$_{\uparrow 4.17}$	19.94	51.63

Table 2: Results on HallusionBench. qAcc and fAcc assess the accuracy of answering a question and understanding a figure, paired with different images and questions, respectively.

Approach	**Generative**				**Discriminative**				**AMBER Score**$_{\uparrow}$
	CHAIR$_{\downarrow}$	**Cover**$_{\uparrow}$	**Hal**$_{\downarrow}$	**Cog**$_{\downarrow}$	**F1**$_{\text{E}\uparrow}$	**F1**$_{\text{A}\uparrow}$	**F1**$_{\text{R}\uparrow}$	**F1**$_{\uparrow}$	
SFT	7.8	51.0	36.4	4.2	64.6	65.6	62.4	74.7	83.5
HA-DPO	6.7	49.8	30.9	3.3	88.1	66.1	**68.8**	78.1	85.7
Synthetic Augmented Data									
DPO	7.3	50.2	33.6	3.7	**95.2**	75.1	60.9	83.1	87.9
V-DPO (Ours)	6.6$_{\downarrow 1.2}$	49.1$_{\downarrow 1.9}$	30.8$_{\downarrow 5.6}$	3.1$_{\downarrow 1.1}$	95.1	**76.1**	61.1	**83.5**$_{\uparrow 8.8}$	**88.4**$_{\uparrow 4.9}$
RLHF-V									
DPO	5.7	49.7	**27.3**	**2.6**	90.7	72.6	64.6	80.9	87.6
V-DPO (Ours)	**5.6**$_{\downarrow 2.2}$	49.7$_{\downarrow 1.3}$	**27.3**$_{\downarrow 9.1}$	2.7$_{\downarrow 1.5}$	91.5	73.7	64.1	81.6$_{\uparrow 5.9}$	88.0$_{\uparrow 4.5}$

Table 3: Result comparison on AMBER. For generative tasks, we use CHAIR (Rohrbach et al., 2018), Cover (coverage of ground-truth objects), Hal (hallucination rate), and Cog (Cognition) as evaluation metrics. We report the performance of discriminative tasks using F1 scores, including splits of existence (E), attribute (A), and relation (R). The holistic AMBER Score (Wang et al., 2023) is calculated by $(100 - \text{CHAIR} + \text{F1})/2$. We compare with HA-DPO (Zhao et al., 2023) backboned with the same SFT model, LLaVA-v1.5-7B (Liu et al., 2023b).

in AMBER score. This indicates the effect of visual guidance in enhancing visual understanding for hallucination mitigation.

HallusionBench. In Table 2, we use qAcc, fAcc, and aAcc to assess performance on the question-, figure-, and individual-level tasks, respectively[1]. We observe a significant improvement in qAcc of V-DPO trained on the synthetic data, with an absolute increase of 9.01% in the accuracy, compared to 4.17% when using RLHF-V for training. One possible explanation for this gap is that the synthetic data mitigates reliance on language priors more efficiently via image-contrast preference learning.

MMHal-Bench. We conduct GPT-4[2] evaluation on MMHal-Bench. Table 4 presents the hallucination rates and overall scores of the outputs from different models. We observe substantial performance improvements in both synthetic and human-annotated preference data scenarios. Furthermore, we perform meso-analysis on splits of different question types in Figure 3. Compared to vanilla DPO, V-DPO is especially effective in answering *comparison* and *environment* questions. Different types of preference data also contribute to the performance gains differently, where our synthetic data shows a superior effect in tackling challenging tasks such as *adversarial* and *relation* questions.

5.3 Ablation Study

We conduct analyses to investigate the effect of visual guidance in V-DPO. We consider ablations on the γ-controlled strength of visual guidance, the calculation of vision-unconditioned distribution, and guidance inflation from normalization.

Strength of Visual Guidance. Figure 4 illustrates the performance changes on AMBER with different values of the visual guidance weight γ. Specifically, we maintain the same $\beta = 0.1$ as in DPO (Rafailov et al., 2023) to avoid substantial divergence from the initial model during training and increase $\alpha > 0$ to enhance the strength of visual guidance. When $\gamma = 1$, it becomes vanilla DPO without additional enhancement on visual guidance. As γ decreases (*i.e.*, α increases), the performance first increases in both training scenarios. However, V-DPO is more sensitive to the guid-

[1] The GPT-4 evaluation was performed in June 2024.

[2] We obtained these results (gpt-4-0613) also in June 2024.

Approach	Hal$_{\downarrow}$	Score$_{\uparrow}$
SFT	0.62	1.97
Synthetic Augmented Data		
DPO	0.59	2.12
V-DPO	**0.53**$_{\downarrow 0.09}$	**2.36**$_{\uparrow 0.39}$
RLHF-V		
DPO	0.60	2.08
V-DPO	0.56$_{\downarrow 0.06}$	2.16$_{\uparrow 0.19}$

Table 4: MMHal-Bench results on hallucination rate (Hal) and overall GPT-4 score.

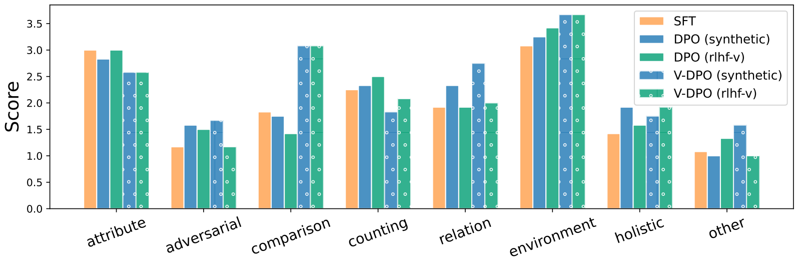


Figure 3: Meso-analysis on MMHal-Bench results comparing performance in different splits of question types.

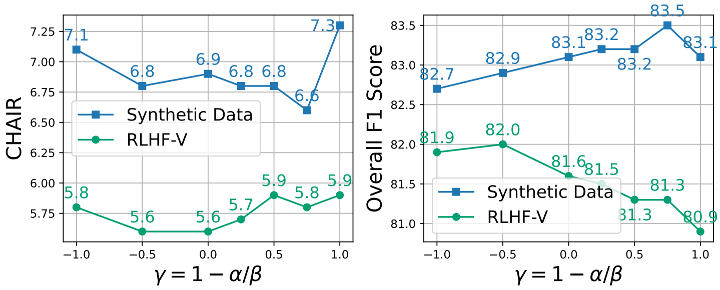


Figure 4: Performance curves (CHAIR$_{\downarrow}$ and F1$_{\uparrow}$) on AMBER with the change of the visual guidance weight γ.

Approach	CHAIR$_{\downarrow}$	F1$_{\uparrow}$
Synthetic Augmented Data		
V-DPO	6.6	83.5
w/ static-lm	6.3$_{\downarrow 0.3}$	83.7$_{\uparrow 0.2}$
w/ normalization	6.2$_{\downarrow 0.4}$	83.1$_{\downarrow 0.4}$
RLHF-V		
V-DPO	5.6	81.6
w/ static-lm	5.2$_{\downarrow 0.4}$	82.4$_{\uparrow 0.8}$
w/ normalization	5.5$_{\downarrow 0.1}$	80.4$_{\downarrow 1.2}$

Table 5: Ablation study on the choice of vision-unconditioned distribution and normalization for V-DPO.

ance control on synthetic preference data, where a small γ such as $\gamma = 0.0$[3] can lead to substantial divergence from the initial model, resulting in performance degradation in hallucination tasks. One possible cause of this degradation is the integration of image-contrast data, which may deviate greatly from the initial SFT model generation distributions, increasing the instability of V-DPO given a higher guidance weight. Empirically, we suggest employing data-specific visual guidance control with $\gamma = (0.75, 0.00)$ for (synthetic-, human-annotated) scenarios, respectively.

Vision-Unconditioned Distribution Calculation. In Eq. 10, we estimate the vision-unconditioned distribution by replacing the visual representations with zeroes. However, as we only utilize vision-conditioned data for preference learning, the vision-unconditioned distribution can become unreliable due to distribution shifts during training (Figure 1b). To interpret the potential influence of the distribution shifts, we use the initial SFT model to calculate the vision-unconditioned distribution instead (*i.e.*, "w/ static-lm" in Table 5). The static textual-only probabilities improve the model performance across both generative and discriminative tasks. This indicates the importance of maintaining reliable vision-unconditioned distribution to integrate appropriate visual guidance during training, shedding light on incorporating textual-only preference data to refine the vision-unconditioned distribution.

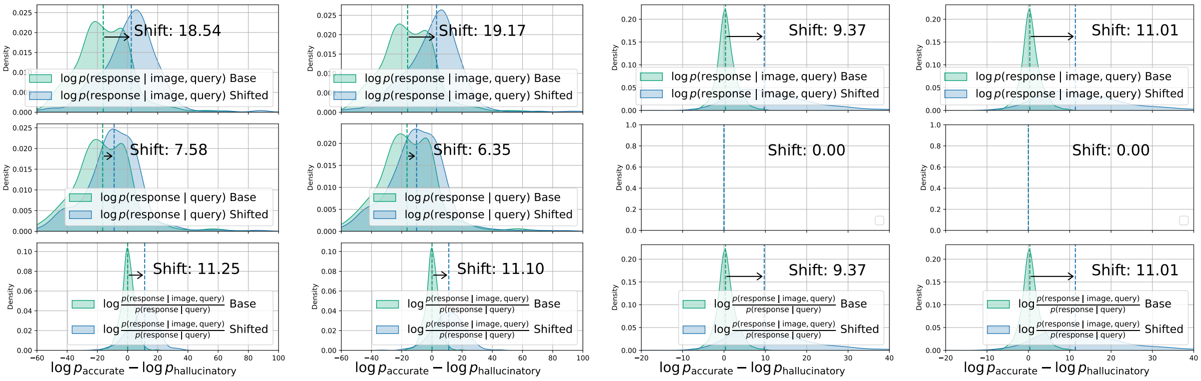
Guidance Inflation with Normalization. As discussed in Section 4.1, we can normalize the vision-enhanced distribution to inflate the guidance effect. Table 5 shows the model performance after this normalization. Notably, the guidance inflation further mitigates hallucination in generative tasks, achieving lower CHAIR scores (*e.g.*, 6.2 and 5.5 compared to 6.6 and 5.6) in both data scenarios. However, it may lead to performance drops in discriminative tasks where the result generation distribution is more sensitive to the modified target in preference optimization.

5.4 Further Analysis

We now investigate the distribution shifts in V-DPO and analyze the qualitative results on MMHal-Bench. Finally, we use the non-hallucination benchmark MMBench to assess the stability of our approach in general tasks in Appendix D.

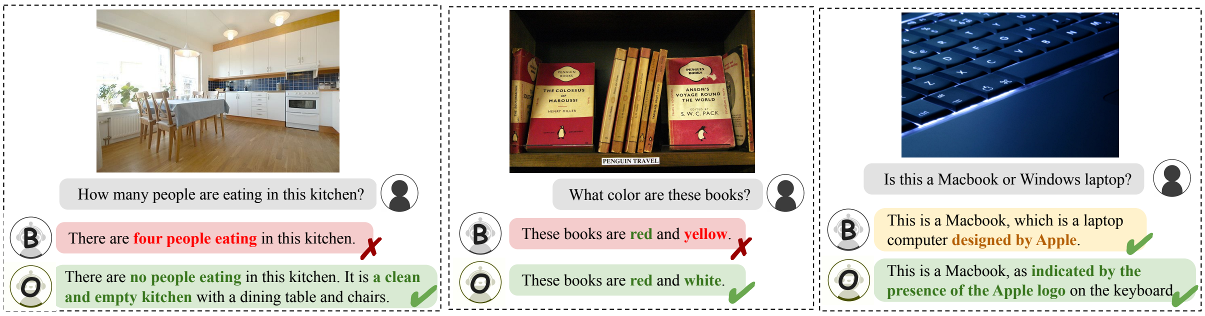
Shifts of Distribution Gaps in V-DPO. Our ablation study (§ 5.3) shows that preference learning can also shift the distribution gaps between

[3] $\gamma - 1 = -1$ in Eq. 8



(a) response-contrast DPO (b) response-contrast V-DPO (c) image-contrast DPO (d) image-contrast V-DPO

Figure 5: Comparison between V-DPO and vanilla DPO on the shifts of distribution gaps. Rows from top to bottom illustrate the distributions of vision-conditioned generation, textual-only-conditioned generation, and the difference between the two generations, respectively. Note that the shifts equal to 0 in the textual-only case on image-contrast data, as the samples within a preference pair have the same textual context with each other.


How many people are eating in this kitchen?

B: There are **four people eating** in this kitchen. ✗

O: There are **no people eating** in this kitchen. It is **a clean and empty kitchen** with a dining table and chairs. ✓

What color are these books?

B: These books are **red** and **yellow**. ✗

O: These books are **red** and **white**. ✓

Is this a Macbook or Windows laptop?

B: This is a Macbook, which is a laptop computer **designed by Apple**. ✓

O: This is a Macbook, as **indicated by the presence of the Apple logo** on the keyboard. ✓

Figure 6: Qualitative examples of different predictions of the SFT baseline (**B**) and our V-DPO approach (**O**). We **bolded** keywords indicating the accuracy and informativeness of visual understanding.

accurate and hallucinatory samples in the vision-unconditioned case. In Figure 5, we show how V-DPO shifts the distributions across different preference data. Our V-DPO approach is more effective than vanilla DPO in enhancing the ability to determine image-contrast hallucination samples, with a shift of 11.01, compared with 9.37 in DPO, as measured by the log-likelihood pairwise preference data differences. For the response-contrast scenario, V-DPO also increases the discriminability with a shift of 19.17. Furthermore, we observe a smaller shift of 6.35 in V-DPO in the textual-only distributions compared with that of 7.58 in DPO, indicating the effectiveness of our approach to mitigate the over-reliance on language priors with visual guidance.

Qualitative Analysis on MMHal-Bench. We conduct qualitative analysis to investigate how V-DPO eliminates hallucination in the generated responses. Figure 6 compares the different generations of V-DPO and the baseline on three examples from MMHal-Bench. The first example, from the adversarial split, shows the significant efficacy of our approach in mitigating the language priors, which may provide a plausible but incorrect answer to the question; *i.e.* "four people eating". In the third example, the model learns to justify its answer "Macbook" according to the specific visual clue of the "Apple logo" in the image. This indicates that our approach enhances visual understanding to elicit related details in the images, improving the informativeness of the generations.

6 Conclusion

We propose V-DPO, utilizing Classifier-Free Guidance (CFG) to integrate visual guidance in LVLM preference learning. Integrating visual guidance into the training process enhances visual context understanding via preference optimization, improving the accuracy and specificity of model generations. Extensive experiments on various preference data demonstrate the generalizability of V-DPO. We hope our work sheds light on visual guidance for more general tasks in LVLM alignment.

Limitations

The main limitations of our work come from two parts. The first one, regarding the V-DPO approach, is the unexplored domains where the language priors are important to guide LVLMs to provide correct answers. For example, preference pairs that prioritize the fluency of the generated text are not considered in our data construction. Future work may explore more general scenarios where both visual and textual modalities are important to elicit the preferred responses. The second one, related to the construction of our synthetic dataset, is the noise and bias introduced by the automatic generation pipeline which may cause performance degradation during preference optimization. For future work, we may consider a more reliable and scalable way to conduct data filtering and reweighting to refine the quality of synthetic augmented data.

Ethics Statement

This work mainly focuses on enhancing visual understanding via preference optimization to mitigate hallucination in LVLMs. One potential ethical concern may come from the data collection process for our synthetic preference pair construction. As the image manipulation process is conducted collaboratively among LVLMs, LLMs, and Stable Diffusion models, systematic bias may be introduced into the generated data. In this case, usage of our synthetic augmented data should be constrained within research-only targets. We leave it to future work to mitigate the bias in model-generated data to further improve the quality of our preference data.

References

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *CoRR*, abs/2404.18930.

Ralph Allan Bradley and Milton E Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. HALC: object hallucination reduction via adaptive focal-contrast decoding. *CoRR*, abs/2403.00425.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Prafulla Dhariwal and Alexander Quinn Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8780–8794.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 1307–1323. Association for Computational Linguistics.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A confederacy of models: a comprehensive evaluation of llms on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 14504–14528. Association for Computational Linguistics.

Nitzan Bitton Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! A vision-and-language benchmark of synthetic and compositional images. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 2616–2627. IEEE.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. 2023. Guiding image captioning models toward more specific captions. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 15213–15223. IEEE.

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

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *CoRR*, abs/2311.07362.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v(ision), llava-1.5, and other multi-modality models. *CoRR*, abs/2310.14566.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11451–11461. IEEE.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. 2023. Stay on topic with classifier-free guidance. *CoRR*, abs/2306.17806.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arık, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *arXiv preprint arXiv:2405.18654*.

Lingfeng Shen, Sihao Chen, Linfeng Song, Lifeng Jin, Baolin Peng, Haitao Mi, Daniel Khashabi, and Dong Yu. 2023. The trickle-down impact of reward (in-)consistency on RLHF. *CoRR*, abs/2309.16155.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented RLHF. *CoRR*, abs/2309.14525.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. 2024. VIGC: visual instruction generation and correction. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 5309–5317. AAAI Press.

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. *CoRR*, abs/2311.07397.

Cheng Xu, Xiaofeng Hou, Jiacheng Liu, Chao Li, Tianhao Huang, Xiaozhi Zhu, Mo Niu, Lingyu Sun, Peng Tang, Tongqiao Xu, Kwang-Ting Cheng, and Minyi Guo. 2023. Mmbench: Benchmarking end-to-end multi-modal dnns and understanding their hardware-software implications. In *IEEE International Symposium on Workload Characterization, IISWC 2023, Ghent, Belgium, October 1-3, 2023*, pages 154–166. IEEE.

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. *CoRR*, abs/2310.16045.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2023. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *CoRR*, abs/2312.00849.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. From recognition to cognition: Visual commonsense reasoning. *CoRR*, abs/1811.10830.

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

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. *CoRR*, abs/2310.00754.

A Deriving V-DPO Objective

Given the maximization objective to optimize in Eq. 7, we have:

$$
\begin{aligned}
&\max_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P},y\sim\pi}\Big[r(v,x,y) - \beta\mathbb{D}_{\mathrm{KL}}\left[\pi(y \mid v,x) \parallel \pi_{\mathrm{ref}}(y \mid v,x)\right] + \alpha\mathbb{D}_{\mathrm{KL}}\left[\pi(y \mid v,x) \parallel \pi(y \mid x)\right]\Big] \\
&\qquad\qquad = \max_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}}\mathbb{E}_{y\sim\pi(y|v,x)}\left[r(v,x,y) - \beta\log\frac{\pi(y \mid v,x)}{\pi_{\mathrm{ref}}(y \mid v,x)} + \alpha\log\frac{\pi(y \mid v,x)}{\pi(y \mid x)}\right] \\
&\qquad\qquad = \min_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}}\mathbb{E}_{y\sim\pi(y|v,x)}\left[\log\frac{\pi(y \mid v,x)}{\pi_{\mathrm{ref}}(y \mid v,x)} - \frac{\alpha}{\beta}\log\frac{\pi(y \mid v,x)}{\pi(y \mid x)} - \frac{1}{\beta}r(v,x,y)\right] \\
&\qquad\qquad = \min_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}}\mathbb{E}_{y\sim\pi(y|v,x)}\left[\log\frac{\pi(y \mid v,x)^{1-\frac{\alpha}{\beta}}/\pi(y \mid x)^{-\frac{\alpha}{\beta}}}{\frac{1}{Z(v,x)}\pi_{\mathrm{ref}}(y \mid v,x)\exp\left(\frac{1}{\beta}r(v,x,y)\right)} - \log Z(v,x)\right] \\
&\qquad\qquad = \min_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}}\mathbb{E}_{y\sim\pi(y|v,x)}\left[\log\frac{\pi(y \mid v,x)\left(\frac{\pi(y|v,x)}{\pi(y|x)}\right)^{-\frac{\alpha}{\beta}}}{\frac{1}{Z(v,x)}\pi_{\mathrm{ref}}(y \mid v,x)\exp\left(\frac{1}{\beta}r(v,x,y)\right)} - \log Z(v,x)\right] \\
&\qquad\qquad = \min_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}}\mathbb{E}_{y\sim\pi(y|v,x)}\left[\log\frac{\pi(y \mid v,x)\left(\frac{\pi(y|v,x)}{\pi(y|x)}\right)^{\gamma-1}}{\frac{1}{Z(v,x)}\pi_{\mathrm{ref}}(y \mid v,x)\exp\left(\frac{1}{\beta}r(v,x,y)\right)} - \log Z(v,x)\right]
\end{aligned}
\tag{13}
$$

where we set $\gamma = 1 - \frac{\alpha}{\beta}$ and the partition function is:

$$
Z(v,x) = \sum_{y}\pi_{\mathrm{sft}}(y \mid v,x)\exp\left(\frac{1}{\beta}r(v,x,y)\right).
$$

Following Rafailov et al. (2023), we define:

$$
\pi^{*}(y \mid v,x) = \frac{1}{Z(v,x)}\pi_{\mathrm{ref}}(y \mid v,x)\exp\left(\frac{1}{\beta}r(v,x,y)\right)
$$

as a valid normalized probability distribution. Different from vanilla DPO, we have the non-normalized term $\pi(y \mid v,x)\left(\frac{\pi(y|v,x)}{\pi(y|x)}\right)^{\gamma-1}$ in our V-DPO objective, which cannot be directly optimized to be $\pi^{*}(y \mid v,x)$. Rearranging Eq. 13 with normalization, we have:

$$
\min_{\pi} \mathbb{E}_{(v,x)\sim\mathcal{I}\times\mathcal{P}}\mathbb{E}_{y\sim\pi(y|v,x)}\left[\log\frac{\frac{1}{W_{\pi}(v,x)}\pi(y \mid v,x)\left(\frac{\pi(y|v,x)}{\pi(y|x)}\right)^{\gamma-1}}{\frac{1}{Z(v,x)}\pi_{\mathrm{ref}}(y \mid v,x)\exp\left(\frac{1}{\beta}r(v,x,y)\right)} - \log Z(v,x) + \log W_{\pi}(v,x)\right]
\tag{14}
$$

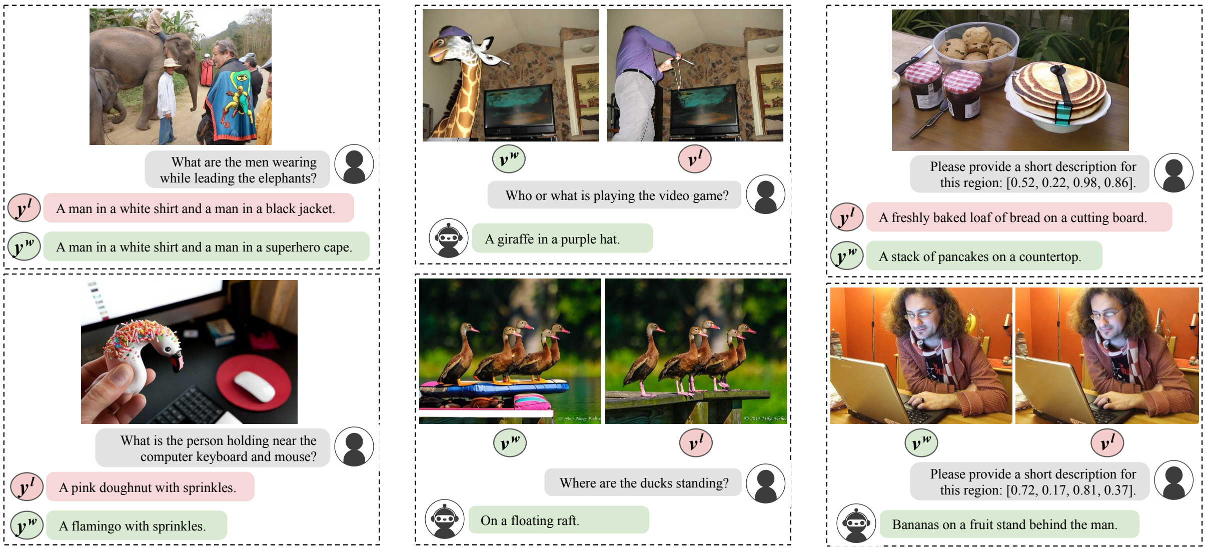
where the partition function:

$$
W_{\pi}(v,x) = \sum_{y\sim\pi(y|v,x)}\pi(y \mid v,x)\left(\frac{\pi(y \mid v,x)}{\pi(y \mid x)}\right)^{\gamma-1}
$$

depends on the policy π. Therefore, we cannot directly solve the normalized vision-enhanced probability distribution using $\pi^{*}(y \mid v,x)$. As $\gamma < 1$, $W_{\pi}(v,x)$ decreases when the vision-conditioned distribution diverges from the textual-only one. As the LVLM is aligned with the LLM backbone, we can make the following proposition:

Proposition 1. *$\exists M < \infty$, for any $y \sim \pi(y \mid v,x)$, the ratio of $\frac{\pi(y|x)}{\pi(y|v,x)}$ is bounded by M*

Proposition 1 holds, according to the practical observation that the LVLM mainly fits well on the seen image data during training while maintaining a similar distribution with the textual-only generation when


What are the men wearing while leading the elephants?

y^l A man in a white shirt and a man in a black jacket.

y^w A man in a white shirt and a man in a superhero cape.

What is the person holding near the computer keyboard and mouse?

y^l A pink doughnut with sprinkles.

y^w A flamingo with sprinkles.

v^w v^l

Who or what is playing the video game?

A giraffe in a purple hat.

v^w v^l

Where are the ducks standing?

On a floating raft.

Please provide a short description for this region: [0.52, 0.22, 0.98, 0.86].

y^l A freshly baked loaf of bread on a cutting board.

y^w A stack of pancakes on a countertop.

v^w v^l

Please provide a short description for this region: [0.72, 0.17, 0.81, 0.37].

Bananas on a fruit stand behind the man.

(a) response-contrast vqa (b) image-contrast vqa (c) region description

Figure 7: Examples of generated preference data.

given unseen images. Based on proposition 1, we take $\min_\pi \mathbb{E} \log W_\pi(v, x)$ as a secondary target and focus on minimizing the first term in Eq. 13 and 14 to elicit an approximation of the optimal solution.

For Eq. 13, one straightforward but probably sub-optimal solution is to solve the vision-enhanced distribution with a proportional constraint with $\pi^*(y \mid v, x)$:

$$\pi(y \mid v, x) \left(\frac{\pi(y \mid v, x)}{\pi(y \mid x)} \right)^{\gamma - 1} \propto \pi^*(y \mid v, x) \tag{15}$$

For Eq. 14, we can solve the normalized probability distribution directly using $\pi^*(y \mid v, x)$:

$$\frac{1}{W_\pi(v, x)} \pi(y \mid v, x) \left(\frac{\pi(y \mid v, x)}{\pi(y \mid x)} \right)^{\gamma - 1} = \pi^*(y \mid v, x) \tag{16}$$

Hence, we complete the derivations for Eq. 7 and 8.

B Implementation Details

We tune the initial SFT model, LLaVA-v1.5-7B, using our V-DPO and the vanilla DPO approaches with the highest learning rate 1e-6 through 4 epochs on both synthetic and human-annotated data scenarios. We adopt a batch size of 64 and set $\beta = 0.1$, following the DPO paper (Rafailov et al., 2023). We employ different weights of visual guidance on the synthetic ($\gamma = 0.75$) and human-annotated ($\gamma = 0.0$) data according to their sensitivity to the control strength. All experiments are conducted with a maximum of 4×40GB GPUs (NVIDIA A100).

C More Details in Preference Data Construction

We choose the images from COCO (Lin et al., 2014), Visual-Genome (Krishna et al., 2017), Visual Commonsense Reaosning (VCR) (Zellers et al., 2018) as the seed set for our synthetic data augmentation pipeline, covering various types of visual content including daily-life scenes and drama-event or human-involved scenarios. Our result synthetic augmented data contains preference pairs, including image-contrast and response-contrast samples on visual instruction following, visual question answering, and region description tasks.

« Element Replacement »

System: You are a good assistant to help me do academic research.
User: I have an image with the caption: "A train is passing by a church.". Substitute each of the following objects with something unexpected to create a sense of discordance: train, church in the format: [what] -> [what]. Provide a brief sentence explaining each substitution.
Assistant:

« Captioning for Manipulated Images »

System: You are a good assistant to generate new captions.
User: I have an original caption and a substitution operation. Return the new caption after conducting the substitution. The original caption is: A train is passing by a church. The substitution involves changing the train to an elephant. Return the updated caption.
Assistant:

« Question Generation »

System: You are a good assistant to generate questions.
User: I have a pair of descriptions. Could you help me generate a question that will lead to different answers based on the two descriptions? Ensure that the question is suitable for both descriptions.
The first description is: A woman is cleaning her dining room.
The second description is: A robot is cleaning her dining room.
Return a question and the corresponding answers according to the two descriptions.
Assistant:

« Distractor (Answer Candidate) Generation »

System: You are a good assistant to generate possible answers.
User: Given a question, please help me to generate some reasonable answers that are common in the real life.
The question is: Where is the bear sitting?
A reasonable answer can be: In a grassy area.
An unreasonable answer can be: In a floating jelly beans. Please help me to generate several reasonable answers, and seperate each answer with "|".
Assistant:

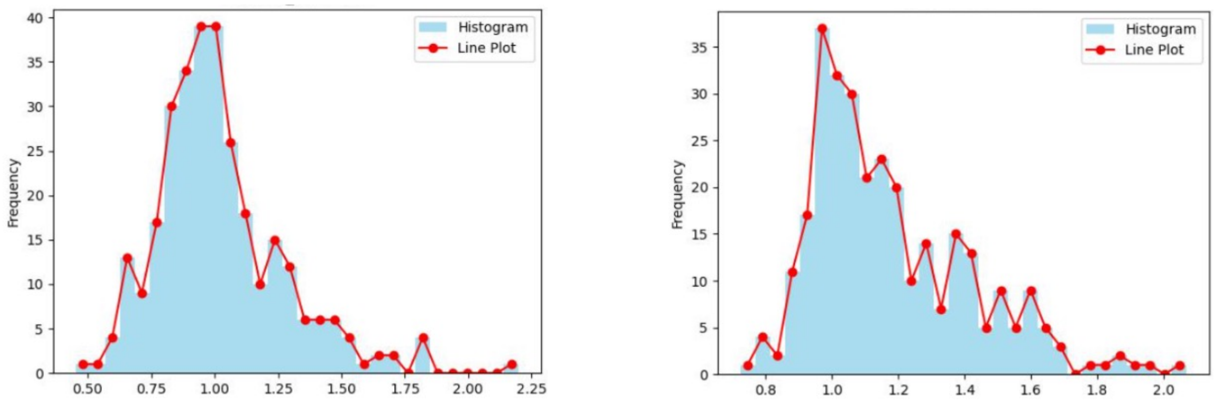
Table 6: Prompt Templates to utilize LLMs to guide the image manipulation process.

C.1 Prompts for Image Manipulation

We show the designed prompts to elicit element replacement ideas from LLMs such as ChatGPT[4] (OpenAI, 2023) in Table 6 and examples of generated preference pairs in Figures 7a to 7c.

C.2 Filtering via CLIPScore

Figure 8 shows the distributions regarding the difference in CLIPScore between positive and negative samples before filtering. We set a threshold $r = \frac{\text{CLIPScore}^w}{\text{CLIPScore}^l} \geq t = 1.5$ to approve the synthetic samples as a valid preference pair.



(a) response-contrast

(b) image-contrast

Figure 8: Distributions of CLIPScore ratios of unfiltered generated preference pairs.

[4] We used gpt-3.5-turbo-1106.

D General Evaluation on MMBench

One drawback of alignment methods is the enlarged divergence from the initial SFT model through training, potentially resulting in model performance degradation on general multimodal tasks. Table 7 assesses V-DPO on the general evaluation benchmark MMBench. While V-DPO still causes a slight drop in overall accuracy, we observe a relatively improved performance compared to the vanilla DPO on both synthetic and human-annotated data scenarios. We leave it to future work to further enhance the stability and generalizability of V-DPO across more general tasks in LVLMs.

Approach	**Level-2 Capability Accuracy**						**Overall Accuracy**$_\uparrow$
	AR$_\uparrow$	**CP**$_\uparrow$	**FP-C**$_\uparrow$	**FP-S**$_\uparrow$	**LR**$_\uparrow$	**RR**$_\uparrow$	
SFT	73.37	77.70	57.34	68.94	32.20	53.04	65.21
Synthetic Augmented Data							
DPO	**74.37**	76.35	56.64	68.94	32.20	53.91	65.03
V-DPO	**74.37**	76.01	**58.04**	68.94	31.36	**54.78**	65.12
RLHF-V							
DPO	**74.37**	76.01	57.34	68.60	31.36	53.04	64.78
V-DPO	73.87	76.69	57.34	68.60	32.20	53.04	64.95

Table 7: MMBench results