

# DATA-AUGMENTED PHRASE-LEVEL ALIGNMENT FOR MITIGATING OBJECT HALLUCINATION

Anonymous authors

Paper under double-blind review

## ABSTRACT

Despite their significant advancements, Multimodal Large Language Models (MLLMs) often generate factually inaccurate information, referred to as hallucination. In this work, we address object hallucinations in MLLMs, where information is generated about an object not present in the input image. We introduce Data-augmented Phrase-level Alignment (DPA), a novel loss which can be applied to instruction-tuned off-the-shelf MLLMs to mitigate hallucinations, while preserving their general vision-language capabilities. To fine-tune MLLMs with DPA, we first generate a set of ‘hallucinated’ and ‘correct’ response pairs through generative data augmentation by selectively altering the ground-truth information of the correct responses at a phrase level. The DPA loss is then used to train MLLMs to reduce the likelihood of hallucinated phrases compared to the correct ones. Our thorough evaluation on various benchmarks confirms the effectiveness of DPA in mitigating hallucination while retaining the out-of-the-box performance of the MLLMs on general tasks. For instance, MLLMs finetuned with DPA, which we refer to as Hallucination Attenuated Language and Vision Assistant (HALVA), improve F1 by up to 13.4% on hallucination visual question-answering and reduce the hallucination rate by up to 4.2% on image description tasks.

## 1 INTRODUCTION

Recent advancements in Large Language Models (LLMs) (Chowdhery et al., 2023; Anil et al., 2023; Raffel et al., 2020; Touvron et al., 2023a;b; Team et al., 2023; Brown et al., 2020) have laid the foundation for the development of highly capable multimodal LLMs (MLLMs) (Team et al., 2023; Liu et al., 2024; 2023c; Dai et al., 2023; Li et al., 2023c; Achiam et al., 2023). MLLMs can process additional modalities such as image or video, while retaining language understanding and generation capabilities. Despite their impressive performance across a variety of tasks, the issue of *object hallucination* in MLLMs presents a significant challenge to their widespread and reliable use (Wang et al., 2023b; Hu et al., 2023; Rohrbach et al., 2018; Bai et al., 2024). Object hallucination refers to generated language that includes descriptions of objects or their attributes that are not present in, or cannot be verified by, the given input. We illustrate a few examples of object hallucinations in Figure 1, where on the left LLaVA-v1.5<sub>13B</sub> inaccurately describes a ‘toothpick’ in an image of utensils (knife, spoon, fork) as these items frequently appear together, while it missed identifying ‘Legos’ due to their rare occurrence with utensils. On the right, LLaVA-v1.5<sub>13B</sub> incorrectly confirms the presence of a ‘tie’ for the image of a ‘wedding cake’. This is likely due to two reasons: first, the frequent co-occurrence of wedding attire such as ‘ties’ and ‘wedding cakes’, and second, MLLMs tend to answer ‘Yes’ for most instructions presented due to positive instruction bias in the training data (Liu et al., 2023b; Bai et al., 2024).

Prior work have attempted to address object hallucination in one of three key stages: inference (Deng et al., 2024a; Yin et al., 2023; Leng et al., 2023; Lee et al., 2023; Zhou et al., 2023; Biten et al., 2022),

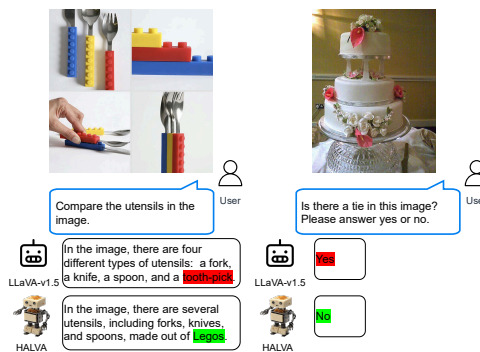


Figure 1: Examples of object hallucinations.

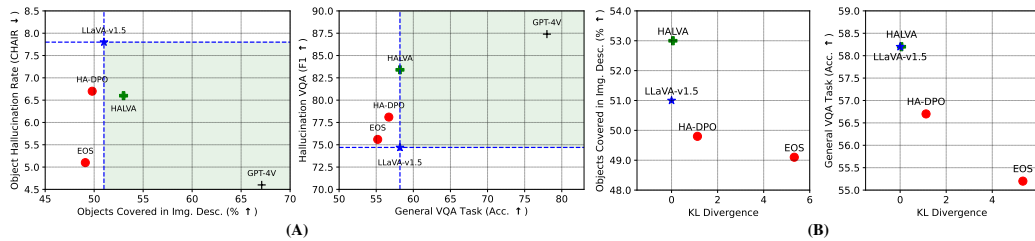


Figure 2: (A): A high-level overview comparing the performance of HALVA (the finetuned model with DPA) with existing finetuning methods in mitigating object hallucination, and their ability on general vision-language tasks. (B): Unlike HALVA, the existing finetuning approaches (e.g., HA-DPO and EOS) substantially diverge from their base model (LLaVA-v1.5<sub>7B</sub>).

pretraining (Sun et al., 2023; Jiang et al., 2023; Liu et al., 2023b), and finetuning (Zhao et al., 2023b; Yue et al., 2024). Inference-based methods aim to mitigate hallucinations during text generation, either through specialized decoding (Leng et al., 2023; Deng et al., 2024a; Zhu et al., 2024) or through iterative corrections (Lee et al., 2023; Wu et al., 2024; Zhou et al., 2023), among others. One of the key limitations of such approaches is that they can substantially increase inference time and cost, and often require modifications to the serving infrastructure (Lee et al., 2023; Bai et al., 2024). Pretraining techniques, such as negative instruction tuning or contrastive learning, have also been used to mitigate object hallucination (Liu et al., 2023b; Jiang et al., 2023). The main limitation of such approaches is that they require massive training data (>500K samples) and can not be applied to off-the-shelf MLLMs. Finally, finetuning-based approaches attempt to mitigate object hallucination through preference optimization (Zhao et al., 2023b) or human feedback (Sun et al., 2023; Yu et al., 2023a), among others (Ben-Kish et al., 2023; Yue et al., 2024).

We note that hallucinations typically occur locally and can be pinpointed to specific words or phrases, such as ‘tooth-pick’ in Figure 1. This is in contrast to other alignment problems such as helpfulness, where it is difficult to identify if a particular word contributes to the overall helpfulness (or lack thereof) in a response. Existing alignment methods (e.g., DPO (Zhao et al., 2023b)) do not leverage this and instead attempt to mitigate hallucinations using a sequence-level loss. Such sequence level loss provides a coarse and noisy signal, making it less effective and causing the model to degenerate from its initial state, leading to a deterioration in general vision-language capabilities (see Figure 2).

Our goal is to achieve a fine-grained mechanism to mitigate hallucinations that allows to tackle hallucinations while not hurting the general capabilities of the model without adding to inference time or requiring substantial re-training. To this end, we first use generative data augmentation (Qin et al., 2022; Zheng et al., 2024) to construct a training set of ‘hallucinated’ and ‘correct’ response pairs, by selectively altering the ground-truth phrases in the correct responses, while keeping the overall structure intact. Next, to reduce the likelihood of hallucinations, we introduce a training objective called *Data-augmented Phrase-level Alignment (DPA)*, to finetune MLLMs using the constructed correct and hallucinated response pairs. Our proposed DPA loss consists of two terms: the first term computes the relative log-probability of the hallucinated tokens compared to the correct ones, and the second term calculates the token-wise KL divergence using a frozen reference model. Accordingly, the MLLM is trained to minimize the likelihood of hallucinated tokens while keeping the divergence minimal. As a result, while DPA is effective in mitigating hallucination it closely retains the general capabilities of the base MLLM. We refer to MLLMs trained with our proposed DPA loss as *Hallucination Attenuated Language and Vision Assistant (HALVA)*. We perform rigorous evaluations on hallucination benchmarks, showing the benefits of our method in mitigating hallucination in both generative and discriminative vision-language tasks. While the primary goal of this work is to mitigate object hallucinations, we take a further step to also evaluate on general vision-language hallucination benchmarks. The results show that DPA also provides benefits toward other forms of vision-language hallucinations that may arise due to visual illusions among others. Finally, to ensure that the proposed DPA does not adversely affect the general capabilities of MLLMs, we evaluate HALVA on popular vision-language benchmarks. Our extensive studies confirm the effectiveness of the proposed method in mitigating object hallucinations while retaining or improving the performance in general vision-language tasks.

In summary, our main contribution is DPA, a novel method to finetune MLLMs for mitigating object hallucination in vision-language tasks. Unlike existing finetuning-based hallucination mitigation meth-

ods, DPA works at a phrase-level and penalizes the tokens where hallucination occurs and not across all the tokens. Such localized and fine-grained feedback reduces object hallucination while retaining the general performance of MLLMs. We open-source the code, checkpoints, and the generated hallucinated and correct response pairs used in training, at <https://anonymous.4open.science/r/HALVA>.

## 2 METHOD: DATA-AUGMENTED PHRASE-LEVEL ALIGNMENT (DPA)

Consider an MLLM, denoted as  $\pi_\theta$ , trained in an auto-regressive manner to predict an output  $y$  for a given vision-language instruction  $x = \{x_v, x_q\}$ , where  $x_v$  is an image and  $x_q$  is the corresponding instruction. During inference, the generated sequence  $s$  of length  $T_s$  is represented as  $\{t_1, t_2, \dots, t_{T_s}\}$ , where each  $t_i$  represents a language token. The sequence  $s$  is said to contain hallucinations if the occurrence of  $t_i$  is not grounded in, or cannot be verified from, the input  $x$ . If the data used to train  $\pi_\theta$  comprises frequent appearance of certain concepts (e.g., objects, object-attribute pairs), the MLLM may generate responses based on learned spurious correlations while ignoring the given inputs (Zhou et al., 2023; Bai et al., 2024; Rohrbach et al., 2018; Li et al., 2023d). Here, we present our strategy to mitigate object hallucinations that may occur due to such co-occurrences.

**Generative data augmentation.** We discuss our strategy to construct ‘hallucinated’ and ‘correct’ response pairs through generative data augmentation. Let  $y^c$  and  $y^h$  be a correct and hallucinated response, respectively, to a vision-language instruction  $\{x_v, x_q\}$ . We design a generative data-augmentation setup to generate  $y^h$  by selectively altering the ground-truth concepts in  $y^c$ , thus introducing hallucinated concepts that are not present in the vision input  $x_v$ . Note that there is no overlap between the correct and the induced hallucinated concepts. Formally, we generate  $y^h$ , by replacing the ground-truth set  $o$  containing the true concepts in  $y^c$ , with the hallucinated set  $o'$ , where  $o' \in \mathbb{O}$  and  $o' \notin x_v$ . Here,  $\mathbb{O}$  is a set containing hallucinated concepts. We define  $\mathbb{O} = \{(o_i, c_i) \mid o_i \in U \text{ and } c_i \subseteq U\}$ , where  $o_i$  is a concept (e.g., object, attribute, or action),  $c_i$  is a subset of concepts that co-occur with  $o_i$ , and  $U$  represents the universal set of all possible concepts of objects and object-related attributes. See an example in Figure 3.

We approximate  $\mathbb{O}$  for hallucinated concepts that are both closed set ( $\mathbb{O}_{cc}$ ) and open-set ( $\mathbb{O}_{oc}$ ). We prepare  $\mathbb{O}_{cc}$  based on the co-occurring concepts in a large object-centric dataset. For  $\mathbb{O}_{oc}$  we sample hallucinated concepts by directly prompting an LLM. In addition to generating descriptive responses, we also use a small set of Yes-or-No questions based on an existing visual question-answering dataset, for which we generate  $y^h$  by simply inverting  $y^c$ . This yields the correct and hallucinated response pairs  $\{y^c, y^h\}$ , which we subsequently use in DPA. Additional details of generative data augmentation, including the templates for generating correct and hallucinated responses, as well as end-to-end examples of the entire augmentation process, are presented in Appendix D.3.

**Proposed phrase-level loss.** Given an off-the-shelf trained MLLM susceptible to hallucinations, our objective is to minimize the likelihood of generating hallucinated tokens using the correct and hallucinated response pairs  $\{y^c, y^h\}$  obtained through generative data-augmentation. To this end, we define an alignment objective based on the relative probabilities of correct and hallucinated phrases.

Let’s take an example with a correct response  $y^c$  as ‘A young man in a white shirt’ and its corresponding hallucinated response  $y^h$  as ‘A young woman in a black dress’. Let  $y_i^h$  denote the  $i$ -th hallucinated phrase in  $y^h$  and  $y_i^c$  be the corresponding correct phrase in  $y^c$ . In this example, the hallucinated phrases are ‘woman’ and ‘black dress’, while their corresponding correct phrases are ‘man’ and ‘white

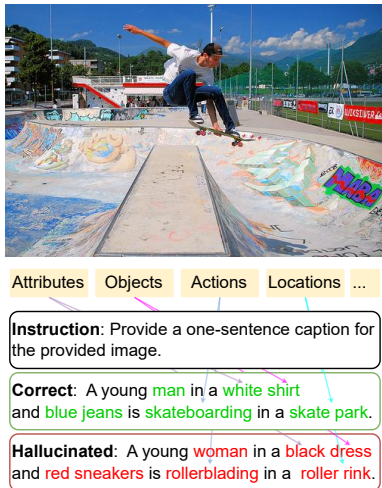


Figure 3: An example of correct and hallucinated response pairs constructed through our generative data-augmentation. The hallucinated responses are generated by selectively altering the true concepts in the correct response. For instance, we alter ‘objects’: shirt  $\rightarrow$  dress, & jeans  $\rightarrow$  sneakers; ‘attributes’: white  $\rightarrow$  black, & blue  $\rightarrow$  red; ‘actions’: skateboarding  $\rightarrow$  rollerblading; and other object-related information such as ‘location’: skate park  $\rightarrow$  roller rink. Best viewed in color.

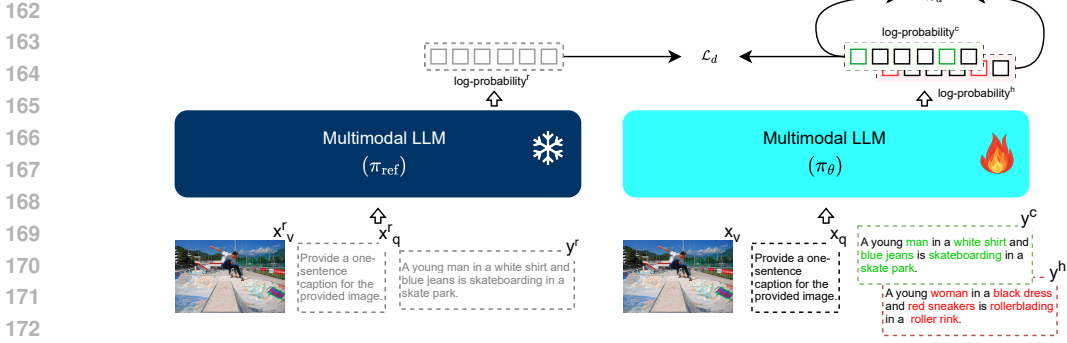


Figure 4: **Overview of our method:** Given a vision-language instruction and its correct and hallucinated response pair, the alignment objective ( $\mathcal{L}_a$ ) reduces the log-likelihood of hallucinated tokens compared to the correct ones. Also, a token-wise KL divergence regularizer ( $\mathcal{L}_d$ ) is employed using a reference model ( $\pi_{\text{ref}}$ ), to restrict the divergence of the MLLM ( $\pi_\theta$ ) during DPA training.

shirt'.  $y^h$  can be expressed as a sequence of tokens  $T_h = \{t_1^h, t_2^h, \dots, t_{|T_h|}^h\}$ , according to which  $y_i^h = T_h[s_i^h : e_i^h]$ , where  $s_i^h$  and  $e_i^h$  are the start and end indices of  $y_i^h$  with  $1 \leq s_i^h \leq e_i^h \leq |T_h|$ . Accordingly, we can compute the probability of hallucinated phrase  $y_i^h$  as  $\prod_{j=s_i^h}^{e_i^h} \pi_\theta(t_j^h|x, t_{<j}^h)$ .

Similarly, the probability of the correct phrase  $y_i^c$  can be expressed as:  $\prod_{j=s_i^c}^{e_i^c} \pi_\theta(t_j^c|x, t_{<j}^c)$ , where  $s_i^c$  and  $e_i^c$  are the start and end indices of  $y_i^c$ . Note that for every  $y_i^h \in y^h$  there exists a corresponding  $y_i^c \in y^c$ . To reduce the relative likelihood of hallucinated phrases compared to the correct ones, we define the alignment loss  $\mathcal{L}_a$  as:

$$\mathcal{L}_a = \frac{1}{N} \sum_{i=1}^N -\log \frac{\prod_{j=s_i^c}^{e_i^c} \pi_\theta(t_j^c|x, t_{<j}^c)}{\prod_{j=s_i^c}^{e_i^c} \pi_\theta(t_j^c|x, t_{<j}^c) + \prod_{j=s_i^h}^{e_i^h} \pi_\theta(t_j^h|x, t_{<j}^h)}, \quad (1)$$

where  $N$  represents the total number of hallucinated phrases in  $y^h$ . Note that our loss is designed to penalize the model  $\pi_\theta$  only for the hallucinated tokens rather than for all tokens in the sequence. This localized and fine-grained feedback is one of the key concepts that sets our method apart from existing preference optimization techniques (Christiano et al., 2017; Rafailov et al., 2024).

Note that simply optimizing  $\pi_\theta$  to minimize  $\mathcal{L}_a$  may cause  $\pi_\theta$  to substantially diverge from its initial state, which may hurt its ability in general vision-language tasks. To mitigate this effect, we train  $\pi_\theta$  with a KL-divergence constraint using a frozen reference model  $\pi_{\text{ref}}$ . For a given reference sample  $\{x^r, y^r\}$ ,  $y^r$  can be expressed as a sequence of tokens  $T_r = \{t_1^r, t_2^r, \dots, t_{|T_r|}^r\}$ . We formulate the token-wise KL-divergence regularization term  $\mathcal{L}_d$  as:

$$\mathcal{L}_d = \sum_{j=1}^{|T_r|} \pi_{\text{ref}}(t_j^r|x^r, t_{<j}^r) \cdot \left( \log(\pi_{\text{ref}}(t_j^r|x^r, t_{<j}^r)) - \log(\pi_\theta(t_j^r|x^r, t_{<j}^r)) \right). \quad (2)$$

Our formulation of  $\mathcal{L}_d$  serves as a token-level regularizer to restrict the model from diverging too far from its initial state, thus losing its general initial abilities. Note that  $\{x^r, y^r\}$  represent any set of vision-language instructions and their correct responses, which may or may not include  $\{x^c, y^c\}$ . Moreover, note that  $\pi_{\text{ref}}$  and  $\pi_\theta$  are initialized from the same checkpoint, therefore  $\mathcal{L}_d$  estimates the divergence of  $\pi_\theta$  from its initial state during training. It should be noted that we adopt a forward KL-divergence approach in calculating  $\mathcal{L}_d$  which is different from the reverse KL-divergence used in RLHF (Christiano et al., 2017). This choice is essential in our case, as we do not conduct rollouts of  $\pi_\theta$  during training and rely solely on responses from  $\pi_{\text{ref}}$ , ensuring that  $\pi_\theta$  focuses on high-probability tokens of the reference distribution. Finally, we train  $\pi_\theta$  to minimize the final DPA objective:

$$\mathcal{L}_{\text{dpa}} = \mathcal{L}_a + \alpha \cdot \mathcal{L}_d, \quad (3)$$

where  $\alpha$  is a coefficient to control the divergence of  $\pi_\theta$  during training. The value of  $\alpha$  is set based on ablation studies presented in Section 4.4. We present the pseudo code in Appendix A.

### 3 EXPERIMENT SETUP

**Training data.** We prepare vision-language instructions based on Visual Genome (VG) (Krishna et al., 2017), which is an object-centric image dataset consisting of a total of 108K images and their

216 annotations. Accordingly, we prepare the correct responses with both descriptive (e.g., Describe  
 217 the image in detail.) and non-descriptive (e.g., <Question>, Please answer in  
 218 one word, yes or no) instructions. Descriptive instructions include one-sentence captions,  
 219 short descriptions, and detailed descriptions of images. Moreover, the non-descriptive question-  
 220 answers are directly taken from (Zhao et al., 2023b). We prepare the correct responses using  
 221 Gemini Vision Pro (Team et al., 2023) and based on the original images and ground-truth annotations.  
 222 Subsequently, we perform generative data augmentation to obtain hallucinated responses, as described  
 223 in Section 2. Our final training set consists of a total of 21.5K vision-language instructions and their  
 224 corresponding correct and hallucinated responses.

225 **Implementation details.** We use LLaVA-v1.5 (Liu et al., 2023c) and VILA-v1.5 (Lin et al., 2024)  
 226 as our base models considering their superior performance in general vision-language tasks and the  
 227 availability of their code and models. LLaVA-v1.5 uses Vicuna-v1.5 (Chiang et al., 2023; Touvron  
 228 et al., 2023b) as the language encoder and CLIP ViT-L<sub>14</sub> (Radford et al., 2021) as the vision encoder.  
 229 VILA-v1.5 uses Vicuna-v1.5 (Chiang et al., 2023; Touvron et al., 2023b) as the language encoder  
 230 and SigLip-L-400M (Zhai et al., 2023) as the vision encoder. Note that while LLaVA-v1.5 uses  
 231 images of resolution 336 pixels, VILA-v1.5 is trained with images of resolution 384 pixels. During  
 232 training, we freeze the vision encoder and projection layers, and only train the LLM using LoRA (Hu  
 233 et al., 2021). We refer to the resulting DPA trained checkpoints as HALVA, i.e., HALVA<sub>7B</sub> based on  
 234 LLaVA-v1.5<sub>13B</sub>, HALVA<sub>13B</sub> based on LLaVA-v1.5<sub>13B</sub>, and HALVA<sub>13B/384</sub> based on VILA-v1.5<sub>13B/384</sub>.  
 235 All experiments are conducted on 4 A100-80GB GPUs. We utilize an effective batch size of 64 and  
 236 train for 1 epoch or 342 steps. The training time ranges from 1.5 to 3 hours for 7B and 13B variants.  
 237 The additional implementation details are presented in Appendix D.

238 **Evaluation setup.** First, we evaluate HALVA on four object hallucination benchmarks encompassing  
 239 both generative and discriminative tasks, including CHAIR (Rohrbach et al., 2018), MME-Hall (Fu  
 240 et al., 2023), AMBER (Wang et al., 2023a), and MMHal-Bench (Sun et al., 2023). Additionally, we  
 241 perform a curiosity driven experiment to critically test the impact of our proposed DPA beyond object  
 242 hallucination, using HallusionBench (Liu et al., 2023a). Furthermore, to ensure that DPA does not  
 243 adversely affect the general language generation capabilities of MLLMs, we evaluate HALVA on  
 244 five popular vision-language benchmarks: VQA-v2 (Goyal et al., 2017), MM-Vet (Yu et al., 2023b),  
 245 TextVQA (Singh et al., 2019), MME (Fu et al., 2023) and LLaVA-Bench (Liu et al., 2024). All  
 246 evaluations are conducted thrice, and we report average scores. In the case of GPT-4-based evaluation,  
 247 the performance slightly varies due to the randomness, where we also report the standard deviations.

## 248 4 RESULTS

249 Earlier in Figure 2, we present a high-level overview of HALVA vs. existing finetuning approaches  
 250 (e.g., HA-DPO and EOS) in mitigating object hallucinations and their effect on the general vision-  
 251 language capabilities. Note that both HA-DPO and EOS are based on the same LLaVA-v1.5<sub>7B</sub> as  
 252 HALVA, ensuring a fair comparison. We consider LLaVA-v1.5<sub>7B</sub> as the lower bound and GPT-4V as  
 253 strong reference point given its performance on the standard benchmarks.  
 254

255 **Image description task.** In Figure 2 (A) Left, we compare MLLMs on image description tasks  
 256 in terms of both hallucination rate (AMBER CHAIR) and their detailedness, captured through the  
 257 number of ground-truth objects covered (AMBER Cover). Our goal is to mitigate hallucinations  
 258 while retaining or improving the richness of image descriptions compared to the base model. As  
 259 shown, HALVA captures more ground-truth objects while hallucinating less than HA-DPO. Moreover,  
 260 while EOS achieves a lower hallucination rate, it degrades the detailedness of image descriptions,  
 261 performing worse than the base model. This is an undesired artifact in MLLMs, particularly for tasks  
 262 that require detailedness such as medical imaging analysis (Wang et al., 2023b; Hu et al., 2023).

263 **Question answering task.** In Figure 2 (A) Right, we compare the performance of MLLMs on visual  
 264 question-answering tasks using both object hallucination (AMBER) and general vision-language  
 265 (TextVQA) benchmarks. As shown, both HA-DPO and EOS underperform HALVA in mitigating  
 266 object hallucination and even deteriorate general vision-language abilities compared to the base  
 267 model. These results show the shortcomings of existing approaches, which we address in this work.

268 To further understand the limitations of existing methods in greater detail, we measure divergence  
 269 from the base model in Figure 2 (B). Here we observe that unlike HALVA, both HA-DPO and EOS  
 substantially diverge from the base model, resulting in poor performance in general tasks.

Table 1: Results on **CHAIR**. ‡ and † indicate that the reported values are from (Chen et al., 2023a) and (Yue et al., 2024). \*Results are computed by us, using their official checkpoints.  $C_i$  and  $C_s$  refer to CHAIR at instance and sentence levels.

Method	$C_i$ (↓)	$C_s$ (↓)	Len.
mPLUG-Owl <sup>†</sup> <sub>7B</sub> (Ye et al., 2023a)	30.2	76.8	98.5
MultiModal-GPT <sup>†</sup> <sub>7B</sub> (Gong et al., 2023)	18.2	36.2	45.7
MiniGPT-v2 <sup>†</sup> <sub>7B</sub> (Chen et al., 2023a)	8.7	25.3	56.5
InstructBlip <sub>7B</sub> (Dai et al., 2023)	17.5	62.9	102.9
LLaVA-v1.5 <sup>†</sup> <sub>7B</sub> (Liu et al., 2023c)	15.4	50.0	100.6
EOS <sub>7B</sub> (Yue et al., 2024)	12.3	40.2	79.7
OPERA <sub>7B</sub> (Huang et al., 2023)	12.8	44.6	-
DoLA <sub>7B</sub> (Chuang et al., 2023)	13.8	47.8	-
HA-DPO <sup>†</sup> <sub>7B</sub> (Zhao et al., 2023b)	11.0	38.2	91.0
MEMVR <sub>7B</sub> (Zou et al., 2024)	13.0	46.6	99.6
AGLA <sub>7B</sub> (An et al., 2024)	14.1	43.0	98.8
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>11.7<sub>↓3.7</sub></b>	<b>41.4<sub>↓8.6</sub></b>	<b>92.2</b>
MiniGPT-4 <sup>†</sup> <sub>13B</sub> (Zhu et al., 2023)	9.2	31.5	116.2
InstructBlip <sub>13B</sub> (Dai et al., 2023)	16.0	51.2	95.6
LLaVA <sup>†</sup> <sub>13B</sub> (Liu et al., 2024)	18.8	62.7	90.7
LLaVA-v1.5 <sup>†</sup> <sub>13B</sub> (Liu et al., 2023c)	13.0	47.2	100.9
EOS <sub>13B</sub> (Yue et al., 2024)	11.4	36.8	85.1
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>12.8<sub>↓0.2</sub></b>	<b>45.4<sub>↓1.8</sub></b>	<b>98.0</b>
VILA-v1.5 <sub>13B/384</sub> (Lin et al., 2024)	9.2	33.0	183.4
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>8.4<sub>↓0.8</sub></b>	<b>30.0<sub>↓3.0</sub></b>	<b>182.6</b>

Table 2: Results on **MME-Hall**. ‡ indicating reported values from (Bai et al., 2024). \*Results are computed by us, using official checkpoints. **Red**: worse than base model.

Method	MME-Hall (†)
Cheetor <sub>7B</sub> <sup>‡</sup> (Li et al., 2023b)	473.4
LRV-Instruction <sub>7B</sub> <sup>‡</sup> (Liu et al., 2023b)	528.4
Otter <sub>7B</sub> <sup>‡</sup> (Li et al., 2023a)	483.3
mPLUG-Owl2 <sub>7B</sub> <sup>†</sup> (Ye et al., 2023b)	578.3
Lynx <sub>7B</sub> <sup>‡</sup> (Zeng et al., 2023)	606.7
Qwen-VL-Chat <sub>7B</sub> <sup>‡</sup> (Bai et al., 2023)	606.6
LLaMA-Adapter V2 <sub>7B</sub> <sup>†</sup> (Gao et al., 2023)	493.3
LLaVA-v1.5 <sub>7B</sub> (Liu et al., 2023c)	648.3
HA-DPO <sup>†</sup> <sub>7B</sub> (Zhao et al., 2023b)	<b>618.3</b>
EOS <sup>†</sup> <sub>7B</sub> (Yue et al., 2024)	<b>606.7</b>
VCD <sub>7B</sub> (Leng et al., 2023)	<b>604.7</b>
Woodpecker <sub>7B</sub> (Yin et al., 2023)	<b>366.7</b>
MEMVR <sub>7B</sub> (Zou et al., 2024)	648.3
ARA <sub>7B</sub> (Qu et al., 2024)	648.3
AGLA <sub>7B</sub> (An et al., 2024)	<b>640.0</b>
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>665.0<sub>↑16.7</sub></b>
BLIVA <sub>11B</sub> <sup>†</sup> (Hu et al., 2024)	580.0
MMICL <sub>12B</sub> <sup>†</sup> (Zhao et al., 2023a)	568.4
InstructBLIP <sub>13B</sub> <sup>†</sup> (Dai et al., 2023)	548.3
SPHINX <sub>13B</sub> <sup>†</sup> (Lin et al., 2023)	668.3
Muffin <sub>13B</sub> <sup>†</sup> (Lou et al., 2023)	590.0
RLHF-V <sub>13B</sub> (Yu et al., 2023a)	585.0
LLaVA-v1.5 <sub>13B</sub> (Liu et al., 2023c)	643.3
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>675.0<sub>↑31.7</sub></b>
VILA-v1.5 <sub>13B/384</sub> (Lin et al., 2024)	688.3
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>691.7<sub>↑3.4</sub></b>

#### 4.1 EVALUATION ON OBJECT HALLUCINATION

**CHAIR.** MLLMs can be prone to hallucinations when generating detailed image descriptions (Bai et al., 2024; Rohrbach et al., 2018; Wang et al., 2023a). To assess the impact of DPA in such scenarios, we evaluate HALVA on CHAIR, which stands for Caption Hallucination Assessment with Image Relevance (Rohrbach et al., 2018). This metric calculates the number of objects that appear in the image caption but are not present in the image. Specifically, CHAIR measures hallucination at two levels: instance-level ( $C_i$ ) and sentence-level ( $C_s$ ). During this task, HALVA is prompted with ‘Describe the image in detail’, allowing for the generation of detailed image descriptions. The results in Table 1 demonstrate that HALVA substantially reduces hallucination in image descriptions compared to the base variants. For instance, compared to LLaVA-v1.5<sub>7B</sub>, HALVA<sub>7B</sub> reduces  $C_s$  from 50.0 to 41.4, similarly, compared to VILA-v1.5<sub>13B/384</sub>, HALVA<sub>13B/384</sub> reduces  $C_s$  from 33.0 to 30.0. Furthermore, HALVA<sub>7B</sub> outperforms or matches the performance of other hallucination mitigation methods, such as OPERA (Huang et al., 2023), EOS (Yue et al., 2024), and HA-DPO (Zhao et al., 2023b). It should be noted that our proposed DPA does not negatively impact the language generation ability or expressiveness of MLLMs, unlike EOS (Yue et al., 2024), which substantially reduces the average generation length from 100 to 85 and 79 for the 13B and 7B variants, respectively. As discussed earlier in Section 4, such a degree of reduction can lead to missing key details in image descriptions and are undesirable for MLLMs. In contrast, HALVA maintains the same generation length as the base model, e.g., 98 vs. 100.9 or 182.6 vs. 183.4, while effectively reducing hallucination. However, a limitation of CHAIR (Rohrbach et al., 2018) is that it does not consider other key aspects of image descriptions, such as coverage of objects and detailedness of descriptions, when evaluating hallucination. Therefore, we also evaluate on AMBER (Wang et al., 2023a), a more recent object hallucination benchmark, which we discuss later.

**MME-Hall.** We evaluate HALVA on discriminative tasks using MME (Fu et al., 2023). Specifically, we utilize the hallucination subset of MME, which consists of four object-related subtasks: existence, count, position, and color, referred to as MME-Hall. The full score of each category is 200, making the maximum total score 800. The results presented in Table 2 demonstrate that HALVA substantially improves performance compared to the base model. For instance, HALVA<sub>13B</sub> achieves a score of 675.0, resulting in a performance gain of 31.7 points with respect to the base model LLaVA-v1.5<sub>13B</sub>. Moreover, as presented in Table 2, existing methods including finetuning (e.g., HA-DPO, EOS) and

inference-based (e.g., VCD, Woodpecker) approaches are ineffective in mitigating hallucinations across such broad categories and worsen the performance compared to their base model. The detailed results of MME-Hall are presented in Appendix C.

Table 3: Results on AMBER. † indicates that the reported values are from (Wang et al., 2023a). \*Results are computed by us, using their checkpoint. Red: worse than base model.

Method	Generative Task				Discriminative Task (F1↑)			Overall
	CHAIR (↓)	Coverage (↑)	Hall. Rate (↓)	Cognition (↓)	Existence	Attribute	Relation	
mPLUG-Owl <sup>†</sup> <sub>7B</sub> (Ye et al., 2023a)	21.6	50.1	76.1	11.5	17.2	22.9	6.2	18.9
LLaVA <sup>†</sup> <sub>7B</sub> (Liu et al., 2024)	11.5	51.0	48.8	5.5	8.4	48.6	58.1	32.7
MiniGPT-4 <sup>†</sup> <sub>7B</sub> (Zhu et al., 2023)	13.6	63.0	65.3	11.3	80.0	43.7	52.7	64.7
mPLUG-Owl2 <sup>†</sup> <sub>7B</sub> (Ye et al., 2023b)	10.6	52.0	39.9	4.5	89.1	72.4	54.3	78.5
InstructBLIP <sup>†</sup> <sub>7B</sub> (Dai et al., 2023)	8.8	52.2	38.2	4.4	89.0	76.3	67.6	81.7
LLaVA-v1.5 <sup>†</sup> <sub>7B</sub>	7.8	51.0	36.4	4.2	83.3	64.6	65.6	74.7
HA-DPO <sup>*</sup> <sub>7B</sub> (Zhao et al., 2023b)	6.7	49.8	30.9	3.3	88.1	66.1	68.8	78.1
EOS <sup>*</sup> <sub>7B</sub> (Yue et al., 2024)	5.1	49.1	22.7	2.0	82.8	67.4	69.2	75.6
Woodpecker <sup>*</sup> <sub>7B</sub> (Yin et al., 2023)	6.9	48.9	30.4	3.6	81.7	53.5	41.5	67.0
<b>HALVA<sub>7B</sub> (Ours)</b>	6.6 <sub>±1.2</sub>	53.0 <sub>±2.0</sub>	32.2 <sub>±4.2</sub>	3.4 <sub>±0.8</sub>	93.3	77.1	63.1	83.4 <sub>±8.7</sub>
RLHF-V <sub>13B/448</sub> Yu et al. (2023a)	6.8	46.1	27.4	2.5	95.7	80.0	71.8	87.1
LLaVA-v1.5 <sub>13B</sub> (Liu et al., 2023c)	6.6	51.9	30.5	3.3	78.5	70.2	45.0	73.1
<b>HALVA<sub>13B</sub> (Ours)</b>	6.4 <sub>±0.2</sub>	52.6 <sub>±0.7</sub>	30.4 <sub>±0.1</sub>	3.2 <sub>±0.1</sub>	92.6	81.4	73.5	86.5 <sub>±13.4</sub>
VILA-v1.5 <sub>13B/384</sub> (Lin et al., 2024)	9.9	63.3	56.1	4.8	87.5	77.8	66.7	82.2
<b>HALVA<sub>13B/384</sub> (Ours)</b>	9.1 <sub>±0.8</sub>	63.9 <sub>±0.6</sub>	54.2 <sub>±1.9</sub>	4.0 <sub>±0.8</sub>	93.9	82.6	75.9	87.9 <sub>±5.7</sub>
GPT-4V <sup>†</sup> (Achiam et al., 2023)	4.6	67.1	30.7	2.6	94.5	82.2	83.2	87.4

**AMBER.** To evaluate performance on both generative and discriminative tasks, we use AMBER (Wang et al., 2023a), which measures hallucination using several metrics. For generative tasks, AMBER assesses the frequency of hallucinated objects in image descriptions, similar to (Rohrbach et al., 2018). Moreover, AMBER evaluates hallucination in three additional aspects of generative abilities: the number of ground-truth objects covered in the description, the hallucination rate, and the similarity of hallucinations in MLLMs to those observed in human cognition. Discriminative tasks are categorized into three broad groups: existence, attribute, and relation, each assessed using F1 scores. For additional details on these evaluation metrics, we refer the reader to (Wang et al., 2023a).

The results presented in Table 3 demonstrate that HALVA outperforms the base model by a large margin, in both generative and discriminative tasks. For instance, HALVA<sub>7B</sub> reduces hallucination in caption generation from 7.8 to 6.6, while increasing the coverage of ground-truth objects in the descriptions from 51% to 53%. This confirms that our method reduces hallucination without compromising the descriptive power of MLLMs. On the other hand, while HA-DPO and EOS report slightly lower hallucination rates, the number of ground-truth objects covered is reduced to 49.8% and 49.1%, respectively. This indicates a degradation in the overall performance of these MLLMs on general tasks. Similar shortcomings are also noticed when using inference-based correction methods such as Woodpecker (Yin et al., 2023), where the object coverage is reduced by 2.1% compared to the base model. Woodpecker also performs poorly on discriminative tasks as it fails to capture key concepts from short responses of LLaVA-v1.5 which it aims to correct. Moreover, our proposed DPA substantially enhances performance on discriminative tasks, for both 7B and 13B variants. For instance, HALVA<sub>7B</sub> improves the F1-score on the attribute category from 64.6% to 80.0%. Additionally, HALVA<sub>13B</sub> improves the F1 score on relation-based tasks from 45.0% to 73.5%. Overall, HALVA<sub>7B</sub> outperforms both HA-DPO and EOS on discriminative tasks by a large margin, achieving a 5.3% and 7.8% higher F1 score respectively. Furthermore, HALVA<sub>13B</sub> and HALVA<sub>13B/384</sub> perform better or on par with GPT-4V on discriminative tasks, i.e., F1-score of 86.5 by HALVA<sub>13B</sub>, 87.9 by HALVA<sub>13B/384</sub>, and 87.4 by GPT-4V.

**MMHal-Bench.** We also conduct LLM-assisted hallucination evaluation to rigorously test for potential hallucinations in generated responses that might not be captured when validated against a limited ground-truth information, as done in (Rohrbach et al., 2018). We utilize MMHal-Bench (Sun et al., 2023), which evaluates hallucination across 12 object-topics, including object attributes, presence of adversarial objects, and spatial relations, among others. Following (Sun et al., 2023), we use GPT-4 (Achiam et al., 2023) as the judge to rate the responses on a scale of 0 to 6, with respect to standard human-generated answers and other ground-truth information of the images. The results presented in Table 4 demonstrate that HALVA considerably improves performance with respect to

Table 4: Results on **MMHal-Bench**. †, ‡, and \*\* indicate that the reported values are from (Sun et al., 2023), (Jiang et al., 2023), and (Yu et al., 2024). \* Results are computed by us, using their official checkpoint. **Red**: worse than base model.

Method	Overall Score (†)	Hall. Rate (‡)
Kosmos-2 <sup>†</sup> (Peng et al., 2023)	1.69	0.68
IDEFIC <sup>†</sup> <sub>9B</sub> (Laurençon et al., 2024)	1.89	0.64
InstructBLIP <sup>†</sup> <sub>7B</sub> (Dai et al., 2023)	2.10	0.58
LLaVA <sup>†</sup> <sub>7B</sub> (Liu et al., 2024)	1.55	0.76
VCD <sup>**</sup> <sub>7B</sub> (Leng et al., 2023)	2.12	0.54
OPERA <sub>7B</sub> (Huang et al., 2023)	2.33	0.50
LURE <sub>7B</sub> (Zhou et al., 2023)	1.64	0.60
LLaVA-SFT <sub>7B</sub> (Sun et al., 2023)	1.76	0.67
LLaVA-RLHF <sub>7B</sub> (Sun et al., 2023)	2.05	0.68
LLaVA-v1.5 <sub>7B</sub> (Liu et al., 2023c)	2.11 <sup>±0.05</sup>	0.54 <sup>±0.01</sup>
HACL <sub>7B</sub> (Jiang et al., 2023)	2.13	<b>0.50</b>
HA-DPO <sup>†</sup> <sub>7B</sub> (Zhao et al., 2023b)	<b>1.97</b>	<b>0.60</b>
EOS <sup>†</sup> <sub>7B</sub> (Yue et al., 2024)	<b>2.03</b>	<b>0.59</b>
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>2.25<sup>±0.09</sup></b>	<b>0.54<sup>±0.01</sup></b>
LLaVA <sup>†</sup> <sub>13B</sub> (Liu et al., 2024)	1.11	0.84
InstructBLIP <sup>†</sup> <sub>13B</sub> (Dai et al., 2023)	2.14	0.58
RLHF-V <sub>13B/448</sub> (Yu et al., 2023a)	-	0.52
LLaVA-SFT <sub>13B</sub> (Sun et al., 2023)	2.43	0.55
LLaVA-RLHF <sub>13B</sub> (Sun et al., 2023)	2.53	0.57
LLaVA-v1.5 <sub>13B</sub> (Liu et al., 2023c)	2.37 <sup>±0.02</sup>	0.50 <sup>±0.00</sup>
CODE <sub>13B</sub> (Kim et al., 2024)	2.49	0.51
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>2.58<sup>±0.07</sup></b>	<b>0.45<sup>±0.02</sup></b>
VILA-v1.5 <sub>13B/384</sub> (Lin et al., 2024)	2.58 <sup>±0.02</sup>	0.46 <sup>±0.01</sup>
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>2.58<sup>±0.06</sup></b>	<b>0.45<sup>±0.01</sup></b>
GPT4V (Achiam et al., 2023)	3.49	0.28

LLaVA-v1.5. Furthermore, we observe that our approach is more effective in mitigating hallucination than existing RLHF, SFT, or DPO-based methods. For example, HALVA<sub>7B</sub> achieves a score of 2.25 surpassing the 7B variants of RLHF, DPO, and SFT-based methods, which report scores of 2.05, 1.97, and 1.76, respectively. Moreover, HALVA<sub>13B</sub> reduces the hallucination rate to 0.45, compared to 0.57 for LLaVA-RLHF. Note that as LLaVA-RLHF and LLaVA-SFT use the same language and vision encoders as HALVA (Vicuna-V1.5 and ViT-L/14), ensuring a fair direct comparison. The detailed results for the individual categories are presented in Appendix C.

#### 4.2 EVALUATION ON HALLUCINATION BENCHMARKS BEYOND OBJECT HALLUCINATION

To further stress-test DPA on other forms of vision-language hallucinations that are not restricted to objects and may occur due to visual illusions, we evaluate performance on HallusionBench (Liu et al., 2023a). The results presented in Table 5 demonstrate that our proposed method directly benefits other forms of vision-language hallucinations as well. HALVA<sub>7B</sub>, HALVA<sub>13B</sub>, and HALVA<sub>13B/384</sub> improve the overall accuracy by 1.86%, 2.16%, and 1.21%, respectively, compared to their base models. Moreover, DPA mitigates Yes/No bias in MLLM responses. Specifically, HALVA<sub>13B/384</sub> reduces Yes/No bias from 0.19 to 0.02. Detailed results on HallusionBench are in Appendix C.

#### 4.3 EVALUATION ON NON-HALLUCINATION BENCHMARKS

We further assess HALVA on general vision-language tasks using four popular benchmarks: VQA-v2 (Goyal et al., 2017), MM-Vet (Yu et al., 2023b), TextVQA (Singh et al., 2019), MME (Fu et al., 2023), and LLaVA-Bench-in-the-Wild (LLaVA-BW) (Liu et al., 2024). We follow the evaluation protocol mentioned in LLaVA-v1.5 (Liu et al., 2023c). The results presented in Table 6 show that HALVA maintains or improves performance with respect to the base models. For example, HALVA<sub>7B</sub> improves on MME, MM-Vet, and LLaVA-BW by 16.3, 1%, and 1.8% respectively, while retaining the same performance on TextVQA and VQA-v2. A similar trend is noticed in the case of HALVA<sub>13B</sub>. Unlike HALVA<sub>7B</sub>, existing finetuning methods such as HA-DPO<sub>7B</sub> and EOS<sub>7B</sub>, based on LLaVA-v1.5<sub>7B</sub>, exhibit deterioration in general tasks when tuned for hallucination mitigation.

Table 5: Results on **HallusionBench**. † indicates that the reported values are from (Liu et al., 2023a). \* Results are computed by us, using their official checkpoint.

Method	Yes/No Bias (~0)	Overall Acc. (†)
mPLUG_Owl-v1 <sup>†</sup> <sub>7.2B</sub> (Ye et al., 2023a)	0.32	43.93
MiniGPT5 <sup>†</sup> <sub>7B</sub> (Zheng et al., 2023)	0.25	40.30
MiniGPT4 <sup>†</sup> <sub>7B</sub> (Zhu et al., 2023)	0.19	35.78
InstructBLIP <sup>†</sup> <sub>7B</sub> (Dai et al., 2023)	-0.13	45.26
BLIP2 <sup>†</sup> <sub>7B</sub> (Li et al., 2023c)	0.18	40.48
mPLUG_Owl-v2 <sup>†</sup> <sub>7B</sub> (Ye et al., 2023b)	0.25	47.30
LRV-Instruction <sup>†</sup> <sub>7B</sub> (Liu et al., 2023b)	0.26	42.78
LLaVA-1.5 <sup>†</sup> <sub>7B</sub> (Liu et al., 2023c)	0.31	47.09 <sup>±0.14</sup>
LLaVA-RLHF <sup>†</sup> <sub>7B</sub> (Sun et al., 2023)	0.24	42.96
HA-DPO <sup>†</sup> <sub>7B</sub> (Zhao et al., 2023b)	0.26	48.36
EOS <sup>†</sup> <sub>7B</sub> (Yue et al., 2024)	0.29	48.72
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>0.17<sub>±0.14</sub></b>	<b>48.95<sup>±0.13</sup></b>
Qwen-VL <sup>†</sup> <sub>9.6B</sub> (Bai et al., 2023)	0.12	39.15
Open-Flamingo <sup>†</sup> <sub>9B</sub> (Awadalla et al., 2023)	0.33	38.44
BLIP2-T5 <sup>†</sup> <sub>12B</sub> (Li et al., 2023c)	0.08	48.09
RLHF-V <sup>†</sup> <sub>13B/448</sub> (Yu et al., 2023a)	0.13	47.47
LLaVA-1.5 <sup>†</sup> <sub>13B</sub> (Liu et al., 2023c)	0.26	46.94
LLaVA-1.5 <sup>†</sup> <sub>13B</sub> (Liu et al., 2023c)	0.38	46.50 <sup>±0.09</sup>
LLaVA-RLHF <sup>†</sup> <sub>13B</sub> (Sun et al., 2023)	<b>0.17</b>	46.41
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>0.20<sub>±0.06</sub></b>	<b>49.10<sup>±0.05</sup></b>
VILA-v1.5 <sup>†</sup> <sub>13B/384</sub> (Lin et al., 2024)	0.19	55.39 <sup>±0.05</sup>
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>0.02<sub>±0.17</sub></b>	<b>56.60<sup>±0.18</sup></b>
GPT4V <sup>†</sup> (Achiam et al., 2023)	0.06	65.28
Gemini Pro Vision <sup>†</sup> (Team et al., 2023)	-0.02	36.85



Table 6: Results on **general vision-language tasks**. Our method not only mitigates hallucinations but also retains or improves performance on general vision-language tasks. \*Results are computed by us, using their official checkpoint.

Method	VQA <sub>↑</sub> <sup>2</sup>	MM-Vet <sub>↑</sub>	TextVQA <sub>↑</sub>	MME <sub>↑</sub>	LLaVA-BW <sub>↑</sub>
LLaVA-v1.5 <sub>7B</sub>	78.5	31.1	58.2	1510.7	65.4
HA-DPO <sub>7B</sub>	77.6 <sup>*</sup> <sub>±0.9</sub>	30.7 <sup>*</sup> <sub>±0.4</sub>	56.7 <sup>*</sup> <sub>±1.5</sub>	1502.6 <sub>±8.1</sub>	66.2 <sub>±0.8</sub>
EOS <sub>7B</sub>	77.6 <sup>*</sup> <sub>±0.9</sub>	31.4 <sub>±0.3</sub>	55.2 <sub>±3.0</sub>	1424.4 <sup>*</sup> <sub>±102.6</sub>	65.8 <sub>±0.4</sub>
HALVA <sub>7B</sub>	78.5 <sub>±0.0</sub>	32.1 <sub>±1.0</sub>	58.2 <sub>±0.0</sub>	1527.0 <sub>±16.3</sub>	67.2 <sub>±1.8</sub>
LLaVA-v1.5 <sub>13B</sub>	80.0	36.1	61.2	1530.1	72.5
HALVA <sub>13B</sub>	80.0 <sub>±0.0</sub>	37.8 <sub>±1.7</sub>	61.2 <sub>±0.0</sub>	1544.0 <sub>±13.9</sub>	72.7 <sub>±0.2</sub>

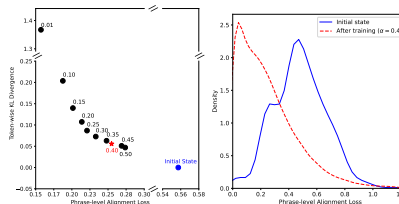


Figure 5: **Left:** Changes in the model state due to DPA training with varying  $\alpha$ . **Right:** Changes in alignment loss before and after training across all training samples. Default  $\alpha$  is 0.4 for HALVA<sub>7B</sub>.

#### 4.4 ABLATION STUDY

Recalling the final DPA objective, which combines the alignment loss ( $\mathcal{L}_a$ ) and KL divergence ( $\mathcal{L}_d$ ), defined as  $\mathcal{L}_{dpa} = \mathcal{L}_a + \alpha \cdot \mathcal{L}_d$ , we examine the change in model state with varying  $\alpha$ , as depicted in Figure 5 (Left). The  $y$  axis represents the extent to which the model diverges from its initial state during DPA training, while the  $x$  axis shows the change in the relative log-probability of the hallucinated tokens. Each data point in this figure represents the calculated alignment loss and divergence after training for different values of  $\alpha$ . The figure illustrates that with a very low  $\alpha$ , e.g. 0.01, the model substantially diverges from its initial state. As  $\alpha$  increases, the model tends to retain a state similar to the base model. We empirically find that  $\alpha=0.4$  works optimally for HALVA<sub>7B</sub>. The change in  $\mathcal{L}_a$  before and after DPA training computed over the entire training samples is presented in Figure 5 (Right). In-depth ablation studies on the proposed loss and generative data-augmentation are presented in Appendix C.

#### 4.5 QUALITATIVE ANALYSIS

A qualitative comparison of HALVA to the base model is shown in Figure 6, with additional examples in Appendix E. HALVA consistently provides more accurate image descriptions than LLaVA-v1.5. For example, in Figure 6 (A), LLaVA-v1.5 hallucinates ‘people’, ‘airport staff’, ‘passengers’ in an image of a parked airplane. In contrast, HALVA accurately describes the image with necessary details. Additionally, our method does not exhibit LLaVA-v1.5’s tendency to answer ‘Yes’ to most questions, which can contribute to hallucinations. This is shown in Figure 6 (B), where HALVA correctly answers ‘Yes’ when asked ‘Is the cloud white in the image?’ and responds with ‘No’ when asked ‘Is the cloud black in this image?’, whereas LLaVA-v1.5 answers ‘Yes’ to both cases. In another example, shown in Figure 6 (C), unlike LLaVA-v1.5, HALVA provides the correct answer to the number of people present in the image. Lastly, we present an example of hallucination caused by visual illusion in Figure 6 (D). While HALVA is not explicitly trained for such vision-language hallucinations, our approach shows some ability to mitigate it.

### 5 RELATED WORK

**Multimodal LLMs.** Vision-language models (VLMs) often align image and text features in a shared embedding space, as pioneered by CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021), followed by others (Yu et al., 2022; Chen et al., 2022; Li et al., 2022; Wang et al., 2022). This alignment is achieved through contrastive learning on large image-text datasets. VLMs show strong generalization across various tasks. Leveraging LLMs and vision encoders from VLMs like CLIP, recent MLLMs (Liu et al., 2024; Zhu et al., 2023; Team et al., 2023; Achiam et al., 2023; Dai et al., 2023; Li et al., 2023c; Peng et al., 2023; Hu et al., 2024; Dai et al., 2023; Bai et al., 2023; Chen et al., 2023b) further enhance visual perception, understanding, and reasoning. While some MLLMs are open-source, others are only accessible through APIs (Achiam et al., 2023; Team et al., 2023; Bai et al., 2023). Among the publicly available MLLMs, LLaVA (Liu et al., 2024; 2023c) and VILA Lin et al. (2024) are widely used due to its simplicity and the availability of code, models, and training data. This makes them suitable base models for demonstrating applicability of DPA to off-the-shelf MLLMs.

**Hallucination in MLLMs.** Multimodal hallucination generally refers to the misrepresentation of verifiable information in relation to the given input. This phenomenon has been primarily studied in

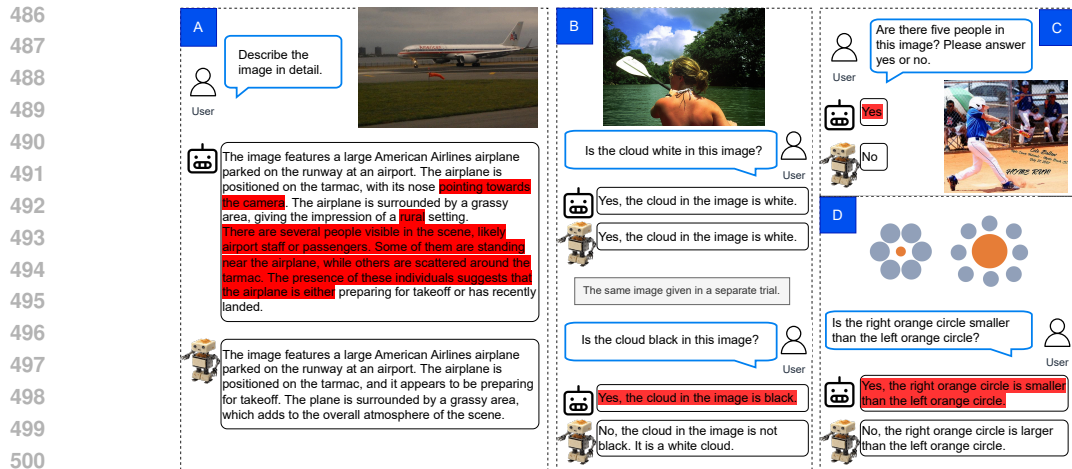


Figure 6: Qualitative comparisons between HALVA [17] and LLaVA-v1.5 [16]. Our proposed DPA effectively mitigates hallucination under different setups: (A) detail image description, (B) visual question-answering, (C) Yes-or-No answer, (D) visual illusion. Hallucinations are highlighted in red. More examples, comparing with LLaVA-v1.5 and VILA-v1.5, are in Appendix E.

the context of object hallucination (Rohrbach et al., 2018; Bai et al., 2024; Zhou et al., 2023; Sun et al., 2023; Biten et al., 2022). Prior work to mitigate this issue can be categorized into three phases: pretraining, where techniques include using balanced instruction-tuning data with equal positive and negative examples (Liu et al., 2023b) or generating and correcting image-instruction pairs on-the-fly (Wang et al., 2024); inference, with methods involving specialized decoding strategies (Leng et al., 2023; Deng et al., 2024a; Zhu et al., 2024) or iterative corrections using offline models to detect and correct hallucinations at inference time (Zhou et al., 2023; Yin et al., 2023); and finetuning, with approaches relying on human feedback (Sun et al., 2023; Yu et al., 2023a) to train reward models or employing preference optimization techniques (Zhao et al., 2023b; Yu et al., 2023a; 2024; Pi et al., 2024; Zhou et al., 2024; Deng et al., 2024b). While finetuning methods are a more efficient direction as they do not require training from scratch (unlike pretraining-based methods) nor changes in the serving infrastructure (unlike inference-based methods), existing finetuning approaches may deteriorate the performance of the base model on general vision-language tasks (Figure 2). To address this, we introduce DPA, which is effective in mitigating object hallucination on a broad set of vision-language tasks while retaining or improving the general abilities of the base model. In contrast to (Gunjal et al., 2024) that explores training a reward model to provide sub-sequence level feedback for preference optimization training, we introduce a fine-grained objective function that can be directly used to finetune multimodal LLMs for hallucination mitigation.

## 6 CONCLUDING REMARKS

We introduce data-augmented phrase-level alignment to mitigate object hallucination in MLLMs. Our approach uses generative data augmentation to create pairs of hallucinated and correct responses by selectively altering ground-truth phrases in the correct responses. These pairs are then used to train MLLMs with our proposed DPA loss, which reduces the relative log-likelihood of hallucinated tokens compared to correct ones. Our extensive study demonstrates the effectiveness of DPA in mitigating various forms of object hallucinations, including those related to existence and attributes, as well as hallucinations arising from visual illusions or complex charts. Additionally, unlike existing fine-tuning-based solutions, DPA effectively mitigates hallucination across diverse vision-language tasks while maintaining or even enhancing performance on general vision-language tasks.

**Limitations.** In this work, we focused on mitigating *object hallucinations* in MLLMs. However, MLLMs also suffer from other forms of hallucinations that may occur due to modality misalignment or over-reliance on language while ignoring other input modalities, among others. While we showed some promising results on generalization to other forms of hallucination, a rigorous exploration of those directions is left for future work. Finally, we believe our method may have applications in other areas as well. For example, it might be adapted to mitigate bias and harmful language generation, among others. We leave this exploration for future research.

## REFERENCES

- 540  
541  
542 Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,  
543 Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report.  
544 *arXiv preprint arXiv:2303.08774*, 2023.
- 545 Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Guang Dai, Ping  
546 Chen, and Shijian Lu. Agla: Mitigating object hallucinations in large vision-language models with  
547 assembly of global and local attention. *arXiv preprint arXiv:2406.12718*, 2024.
- 548 Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos,  
549 Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv*  
550 *preprint arXiv:2305.10403*, 2023.
- 551 Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe,  
552 Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for  
553 training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.
- 554 Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,  
555 and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization,  
556 text reading, and beyond. 2023.
- 557 Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou.  
558 Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*,  
559 2024.
- 560 Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mocha:  
561 Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*,  
562 2023.
- 563 Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach:  
564 Reducing object hallucination in image captioning. In *WACV*, pp. 1381–1390, 2022.
- 565 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,  
566 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are  
567 few-shot learners. *NeurIPS*, 33:1877–1901, 2020.
- 568 Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman  
569 Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large  
570 language model as a unified interface for vision-language multi-task learning. *arXiv preprint*  
571 *arXiv:2310.09478*, 2023a.
- 572 Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing  
573 multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023b.
- 574 Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian  
575 Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual  
576 language-image model. *arXiv preprint arXiv:2209.06794*, 2022.
- 577 Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi  
578 Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial  
579 multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.
- 580 Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,  
581 Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An  
582 open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>.
- 583 Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam  
584 Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm:  
585 Scaling language modeling with pathways. *JMLR*, 24(240):1–113, 2023.
- 586 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep  
587 reinforcement learning from human preferences. *NeurIPS*, 30, 2017.
- 588  
589  
590  
591  
592  
593

- 594 Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola:  
595 Decoding by contrasting layers improves factuality in large language models. *arXiv preprint*  
596 *arXiv:2309.03883*, 2023.
- 597 Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,  
598 Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language  
599 models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023.
- 600 Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large  
601 vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024a.
- 602 Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang.  
603 Enhancing large vision language models with self-training on image comprehension. *arXiv preprint*  
604 *arXiv:2405.19716*, 2024b.
- 605 Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu  
606 Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation  
607 benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.
- 608 Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu,  
609 Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model.  
610 *arXiv preprint arXiv:2304.15010*, 2023.
- 611 Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu,  
612 Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for  
613 dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.
- 614 Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in  
615 vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pp.  
616 6904–6913, 2017.
- 617 Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision  
618 language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp.  
619 18135–18143, 2024.
- 620 Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang,  
621 and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint*  
622 *arXiv:2106.09685*, 2021.
- 623 Mingzhe Hu, Shaoyan Pan, Yuheng Li, and Xiaofeng Yang. Advancing medical imaging with  
624 language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*, 2023.
- 625 Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal  
626 llm for better handling of text-rich visual questions. In *AAAI*, volume 38, pp. 2256–2264, 2024.
- 627 Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming  
628 Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models  
629 via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023.
- 630 Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung,  
631 Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with  
632 noisy text supervision. In *ICML*, pp. 4904–4916. PMLR, 2021.
- 633 Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaying Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang,  
634 Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large  
635 language model. *arXiv preprint arXiv:2312.06968*, 2023.
- 636 Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. Code: Contrasting self-generated  
637 description to combat hallucination in large multi-modal models. *arXiv preprint arXiv:2406.01920*,  
638 2024.
- 639 Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie  
640 Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language  
641 and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017.

- 648 Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov,  
649 Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open  
650 web-scale filtered dataset of interleaved image-text documents. *NeurIPS*, 36, 2024.
- 651
- 652 Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal  
653 hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023.
- 654 Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong  
655 Bing. Mitigating object hallucinations in large vision-language models through visual contrastive  
656 decoding. *arXiv preprint arXiv:2311.16922*, 2023.
- 657
- 658 Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and  
659 Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*,  
660 2023a.
- 661 Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-  
662 Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow  
663 interleaved vision-language instructions. *arXiv preprint arXiv:2308.04152*, 2023b.
- 664 Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-  
665 training for unified vision-language understanding and generation. In *ICML*, pp. 12888–12900.  
666 PMLR, 2022.
- 667
- 668 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-  
669 training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*,  
670 2023c.
- 671 Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object  
672 hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023d.
- 673
- 674 Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On  
675 pre-training for visual language models. In *CVPR*, pp. 26689–26699, 2024.
- 676 Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi  
677 Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for  
678 multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023.
- 679
- 680 Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi  
681 Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context  
682 reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models.  
683 *arXiv preprint arXiv:2310.14566*, 2023a.
- 684 Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating  
685 hallucination in large multi-modal models via robust instruction tuning. In *ICLR*, 2023b.
- 686
- 687 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction  
688 tuning. *arXiv preprint arXiv:2310.03744*, 2023c.
- 689 Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36,  
690 2024.
- 691 Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint*  
692 *arXiv:1711.05101*, 2017.
- 693
- 694 Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin.  
695 Muffin: Curating multi-faceted instructions for improving instruction following. In *ICLR*, 2023.
- 696
- 697 Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu  
698 Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint*  
699 *arXiv:2306.14824*, 2023.
- 700 Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strength-  
701 ening multimodal large language model with bootstrapped preference optimization. *arXiv preprint*  
*arXiv:2403.08730*, 2024.

- 702 Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang.  
703 Understanding and improving robustness of vision transformers through patch-based negative  
704 augmentation. *NeurIPS*, 35:16276–16289, 2022.
- 705 Xiaoye Qu, Qiyuan Chen, Wei Wei, Jishuo Sun, and Jianfeng Dong. Alleviating hallucination in  
706 large vision-language models with active retrieval augmentation. *arXiv preprint arXiv:2408.00555*,  
707 2024.
- 708 Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal,  
709 Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual  
710 models from natural language supervision. In *ICML*, pp. 8748–8763. PMLR, 2021.
- 711 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea  
712 Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*,  
713 36, 2024.
- 714 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi  
715 Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text  
716 transformer. *JMLR*, 21(140):1–67, 2020.
- 717 Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity:  
718 Breaking the gpu memory wall for extreme scale deep learning. In *SC*, pp. 1–14, 2021.
- 719 Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia  
720 Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training. In  
721 *USENIX ATC*, pp. 551–564, 2021.
- 722 Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object  
723 hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.
- 724 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy  
725 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- 726 Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh,  
727 and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pp. 8317–8326, 2019.
- 728 Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan,  
729 Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with  
730 factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023.
- 731 Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu  
732 Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable  
733 multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- 734 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée  
735 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and  
736 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023a.
- 737 Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay  
738 Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation  
739 and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.
- 740 Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li,  
741 Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. In *AAAI*, volume 38,  
742 pp. 5309–5317, 2024.
- 743 Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang,  
744 and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation.  
745 *arXiv preprint arXiv:2311.07397*, 2023a.
- 746 Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou,  
747 Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a  
748 simple sequence-to-sequence learning framework. In *ICML*, pp. 23318–23340. PMLR, 2022.
- 749
- 750
- 751
- 752
- 753
- 754
- 755

- 756 Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interac-  
757 tive computer-aided diagnosis on medical image using large language models. *arXiv preprint*  
758 *arXiv:2302.07257*, 2023b.
- 759  
760 Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical  
761 closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint*  
762 *arXiv:2402.11622*, 2024.
- 763 Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu,  
764 Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with  
765 multimodality. *arXiv preprint arXiv:2304.14178*, 2023a.
- 766  
767 Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and  
768 Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality  
769 collaboration. *arXiv preprint arXiv:2311.04257*, 2023b.
- 770 Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing  
771 Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language  
772 models. *arXiv preprint arXiv:2310.16045*, 2023.
- 773  
774 Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu.  
775 Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*,  
776 2022.
- 777 Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu,  
778 Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment  
779 from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023a.
- 780  
781 Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He,  
782 Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for  
783 super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024.
- 784 Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang,  
785 and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv*  
786 *preprint arXiv:2308.02490*, 2023b.
- 787  
788 Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an  
789 eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024.
- 790 Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and  
791 Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv*  
792 *preprint arXiv:2307.02469*, 2023.
- 793  
794 Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language  
795 image pre-training. In *ICCV*, pp. 11975–11986, 2023.
- 796  
797 Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng  
798 Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with  
799 multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023a.
- 800  
801 Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond  
802 hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv*  
803 *preprint arXiv:2311.16839*, 2023b.
- 804  
805 Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmenta-  
806 tion. *NeurIPS*, 36, 2024.
- 807  
808 Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation  
809 via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023.
- 810  
811 Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal,  
812 and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models.  
813 *arXiv preprint arXiv:2310.00754*, 2023.

810 Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in  
811 vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.  
812  
813 Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigt-4: En-  
814 hancing vision-language understanding with advanced large language models. *arXiv preprint*  
815 *arXiv:2304.10592*, 2023.  
816 Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations  
817 in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*,  
818 2024.  
819 Xin Zou, Yizhou Wang, Yibo Yan, Sirui Huang, Kening Zheng, Junkai Chen, Chang Tang, and  
820 Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination  
821 mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863



## Appendix

The organization of the appendix is as follows:

- Appendix **A**: Pseudo code
- Appendix **B**: Distinction between ours DPA and DPO-based hallucination mitigation methods
- Appendix **C**: Additional experiments and results
- Appendix **D**: Implementation details
- Appendix **E**: Qualitative results

### A DPA PSEUDO CODE

Our proposed DPA is fairly straightforward to implement. Below, we provide a PyTorch-based pseudo code. Please note that this is a minimal implementation to present the key steps of our algorithm. Some of the intermediary and rudimentary steps (e.g., ignoring padded inputs during loss calculation) are intentionally omitted for brevity. The code will be made publicly available.

```

880 import torch
881 import torch.nn.functional as F
882
883 def forward(self, **inputs):
884     """x: vision-language input
885     y_pos: correct response of x
886     y_neg: hallucinated response of x constructed through gen. data aug.
887     x_ref, y_ref: reference input-output pair to calculate divergence
888     """
889
890     batch_size = x.shape[0]
891
892     # forward pass with correct and hallucinated responses
893     pos_logits = self.model(x, y_pos)
894     neg_logits = self.model(x, y_neg)
895
896     # calculate log-probabilities
897     pos_logps, pos_labels = self.log_softmax(pos_logits, y_pos)
898     neg_logps, neg_labels = self.log_softmax(neg_logits, y_neg)
899
900     # accumulate log-probabilities of
901     # correct and hallucinated tokens at phrase level
902     pos_logps = self.accumulate_logps(pos_logps)
903     neg_logps = self.accumulate_logps(neg_logps)
904
905     # phrase-level alignment loss
906     alignment_loss = torch.log(1 + torch.exp(neg_logps - pos_logps))
907     alignment_loss = alignment_loss.mean()
908
909     # forward pass with the reference samples
910     logits = self.model(x_ref, y_ref)
911     with torch.no_grad():
912         reference_logits = self.reference_model(x_ref, y_ref)
913
914     # calculate probability
915     proba = F.softmax(logits, dim=-1)
916     reference_proba = F.softmax(reference_logits, dim=-1)
917
918     # token-wise KL divergence
919     divergence = (reference_proba*(reference_proba.log()-proba.log()))
920     divergence = divergence.sum()/batch_size
921
922     # final loss
923     loss = alignment_loss + self.alpha*divergence
924
925     return loss

```

## B DISTINCTION BETWEEN OURS DPA AND DPO-BASED HALLUCINATION MITIGATION METHODS

Several existing and concurrent works, such as HA-DPO (Zhao et al., 2023b), RLHF-V (Yu et al., 2023a), and RLAIIF (Yu et al., 2024), have introduced hallucination mitigation techniques for MLLMs, that are derived from DPO (Rafailov et al., 2024). Following, we discuss the differences between our proposed DPA and DPO.

We write both DPA (ours) and the DPO (Rafailov et al., 2024) objectives using the same notations, which are as follows:  $\pi_\theta$  as the model being trained;  $\pi_{\text{ref}}$  as the frozen reference model;  $x$  as the input;  $y^c$  and  $y^h$  as correct and hallucinated responses;  $\mathcal{D}$  as training samples. We express  $y^h$  as a sequence of tokens  $T_h = \{t_1^h, t_2^h, \dots, t_{|T_h|}^h\}$  and denote the  $i$ -th hallucinated phrase  $y_i^h = T_h[s_i^h : e_i^h]$ , where  $s_i^h$  and  $e_i^h$  are the start and end indices of  $y_i^h$  with  $1 \leq s_i^h \leq e_i^h \leq |T_h|$ . Similarly,  $y^c$  is expressed as a sequence of tokens  $T_c = \{t_1^c, t_2^c, \dots, t_{|T_c|}^c\}$ , and we denote the  $i$ -th correct phrase  $y_i^c = T_c[s_i^c : e_i^c]$ , where  $s_i^c$  and  $e_i^c$  are the start and end indices of  $y_i^c$  with  $1 \leq s_i^c \leq e_i^c \leq |T_c|$ .  $N$  is the total number of hallucinated phrases in  $y^h$ ;  $\alpha$  and  $\beta$  are loss coefficients to control the influence of the reference model in training. For the sake of simplicity, we assume that  $\{x_c, y_c\}$  are reused as reference sample in DPA. Therefore, as discussed in Section 2, the final DPA loss can be expressed as:

$$\begin{aligned} \mathcal{L}_{dpa}(\pi_\theta; \pi_{\text{ref}}) = & - \mathbb{E}_{(x, y^c, y^h) \sim \mathcal{D}} \left[ \frac{1}{N} \sum_{i=1}^N - \log \underbrace{\frac{\prod_{j=s_i^c}^{e_i^c} \pi_\theta(t_j^c | x, t_{<j}^c)}{\prod_{j=s_i^c}^{e_i^c} \pi_\theta(t_j^c | x, t_{<j}^c) + \prod_{j=s_i^h}^{e_i^h} \pi_\theta(t_j^h | x, t_{<j}^h)}}_{\text{phrase-level alignment loss}} \right. \\ & \left. + \alpha \cdot \underbrace{\sum_{j=1}^{|T_c|} \pi_{\text{ref}}(t_j^c | x, t_{<j}^c) \cdot \left( \log \left( \pi_{\text{ref}}(t_j^c | x, t_{<j}^c) \right) - \log \left( \pi_\theta(t_j^c | x, t_{<j}^c) \right) \right)}_{\text{token-wise KL divergence}} \right] \end{aligned}$$

On the other hand, the training objective of DPO is:

$$\mathcal{L}_{dpo}(\pi_\theta; \pi_{\text{ref}}) = - \mathbb{E}_{(x, y^c, y^h) \sim \mathcal{D}} \left[ \log \sigma \left( \beta \log \frac{\pi_\theta(y^c | x)}{\pi_{\text{ref}}(y^c | x)} - \beta \log \frac{\pi_\theta(y^h | x)}{\pi_{\text{ref}}(y^h | x)} \right) \right]$$

Note that in our proposed DPA ( $\mathcal{L}_{dpa}$ ), given  $\{x, y^c, y^h\}$ , we calculate the phrase-level alignment loss based on the log-probabilities of the tokens in the hallucinated phrases and not on all the tokens of a sequence. Additionally, the KL-regularizer is applied at the token-level to closely retain the vision-language capabilities of the base model. In DPO ( $\mathcal{L}_{dpo}$ ), however, given  $x, y^c, y^h$ , the reward margin between the correct and hallucinated responses is maximized to increase the log-likelihood of the correct response while reducing that of the hallucinated response. Despite the fact that the loss formulation of DPO is different from ours DPA, one fundamental difference is that their loss is calculated at a sequence level, i.e., penalizing all the tokens of a hallucinated response. Intuitively, the training objective of DPA provides more localized and fine-grained feedback unlike DPO (Rafailov et al., 2024) and other existing alignment techniques (Christiano et al., 2017; Schulman et al., 2017). This makes DPA unique and effective compared to existing and concurrent works.

Accordingly, the nature of the correct and hallucinated responses used in DPO-based methods and our DPA also differ. To illustrate this we present one side-by-side comparison using a training sample from HA-DPO (Zhao et al., 2023b) and ours in Figure S1, which shows that while HA-DPO make changes at the sequence level, we apply changes at the word or phrase-level to construct the negative responses. In particular, unlike, HA-DPO, we selectively alter the ground-truth information in the correct description, while keeping the rest of the response intact.

972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
1010  
1011  
1012  
1013  
1014  
1015  
1016  
1017  
1018  
1019  
1020  
1021  
1022  
1023  
1024  
1025



**Chosen:** The photo depicts an **exciting** moment of a snowboarder executing a mid-air jump, with the snowboard **prominently** visible underneath. The snowboarder is wearing **protective gear**, including a helmet and goggles, to ensure safety while **experiencing** the exhilarating activity. The snowy landscape with trees in the backdrop **sets** the scene, and the snowboarder takes center stage, **exhibiting impressive** skill and athleticism as they soar through the air.

**Rejected:** The picture depicts an **electrifying** moment of a snowboarder executing a mid-air jump, with the snowboard clearly visible underneath. The snowboarder, wearing a helmet and goggles, ensures safety while **relishing in** the exhilarating activity. The snowy landscape, **adorned** with trees, **serves as** the backdrop for this scene, **where** the snowboarder takes center stage, **showcasing their** skill and athleticism as they soar through the air.



**Correct:** A **snowboarder** is jumping in the air. The **snowboarder** is surrounded by snow and has a **blue sky** in the background. He has a patch of clear **blue sky** behind him. The **snowboarder** is doing a trick and has his legs bent in the air with his arms extended downward. He has a black and white **glove** on his right hand. The **snowboarder** is wearing a white vest with a black number on the back.

**Hallucinated:** A **skier** is jumping in the air. The **skier** is surrounded by snow and has a **blue water** in the background. He has a patch of clear **blue water** behind him. The **skier** is doing a trick and has his legs bent in the air with his arms extended downward. He has a black and white **hat** on his right hand. The **skier** is wearing a white vest with a black number on the back.

Figure S1: We present training samples from the DPO-based method on the left (from HA-DPO) and ours on the right, highlighting differences in the *nature of the negative samples*. While HA-DPO makes changes (highlighted in blue) at a sequence level, we apply one-to-one changes (highlighted in green and red) at the word or phrase-level to construct the negatives. The positives are referred to as ‘Chosen’ in HA-DPO, while we refer to them as ‘Correct’; and the negatives are referred to as ‘Reject’ in HA-DPO, while we refer to them as ‘Hallucinated’. Since there are no overlapping samples of descriptive responses between HA-DPO and our data, we use a sample that closely resemble each other.

## C ADDITIONAL EXPERIMENTS AND RESULTS

### C.1 ABLATION ON LOSS

Recall our final objective function, which is comprised of both alignment loss ( $\mathcal{L}_a$ ), and token-wise KL divergence ( $\mathcal{L}_d$ ) between the  $\pi_\theta$  (the model being trained) and  $\pi_{\text{ref}}$  (the reference model that is kept frozen), defined as:  $\mathcal{L}_{dpa} = \mathcal{L}_a + \alpha \cdot \mathcal{L}_d$ . First, we study the behavior of HALVA with varying  $\alpha$ . Simply put, a lower  $\alpha$  allows  $\pi_\theta$  to diverge more from  $\pi_{\text{ref}}$ , whereas a higher  $\alpha$  aligns  $\pi_\theta$  more closely with  $\pi_{\text{ref}}$ . By default, we initialize both  $\pi_\theta$  and  $\pi_{\text{ref}}$  from the same base model. Therefore, a higher  $\alpha$  would result in  $\pi_\theta$  to perform the same as the base model. Following, we analyze the impact of varying  $\alpha$  on HALVA<sub>7B</sub> and HALVA<sub>13B</sub>, while tracking their performance on the MME-Hall dataset. The results are presented in Figures S2 and S3. We observe that for HALVA<sub>7B</sub>, an  $\alpha$  of between 0.3 and 0.4 yields a better outcome, whereas the model behaves similar to the base model when  $\alpha > 0.4$ . For HALVA<sub>13B</sub> on the other hand, an  $\alpha$  in the range of 0.4 to 0.6 shows the highest performance. We present qualitative examples in Figure S4, showing the adverse effect of using a very low  $\alpha$ . By default, we use  $\alpha = 0.4$  for HALVA<sub>7B</sub>,  $\alpha = 0.5$  for HALVA<sub>13B</sub>, and  $\alpha = 0.2$  for HALVA<sub>13B/384</sub>.

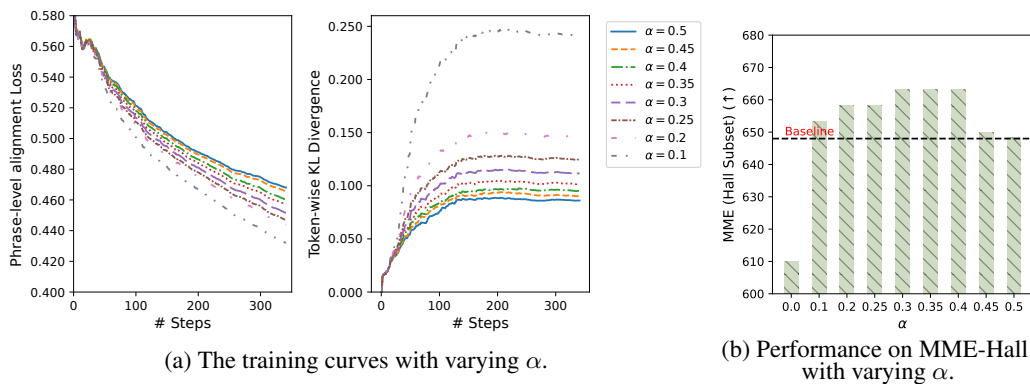


Figure S2: The training curves with varying  $\alpha$  (a) and their performance on object hallucination (b) are presented.  $\alpha$  in the range of 0.3 to 0.4 achieves optimal performance on the 7B variant.

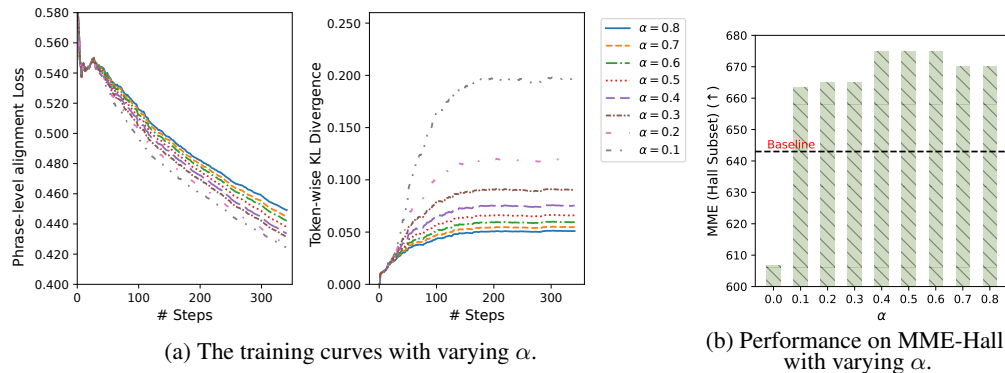


Figure S3: The training curves with varying  $\alpha$  (a) and their performance on object hallucination (b) are presented.  $\alpha$  in the range of 0.4 to 0.6 achieves optimal performance on the 13B variant.



## C.2 ABLATION ON GENERATIVE DATA AUGMENTATION

We perform an ablation study to explore the effect of different sampling strategies which have been used in generative data augmentation. As mentioned in Section 2, we generate hallucinated responses in three setups: closed-set co-occurrences (9K), open-set co-occurrences (11K), and Yes-or-No questions (1.5K). We generate a total of 21.5K samples that contains 28K unique pairs of correct and hallucinated phrases based on 5K unique hallucinated objects. We study the impact of these categories along with their varying number of samples. We perform this study on HALVA<sub>7B</sub> and use the same training hyperparameters as those obtained by tuning on the entire data. From the results presented in Table S1, three key observations are made. First, open-set hallucinated descriptions show benefits in reducing hallucinations in generative tasks, as evidenced by the superior performance on CHAIR. Second, mixing the Yes-or-No hallucinated responses reduces hallucination in discriminative tasks, leading to an F1 boost on the AMBER dataset. Finally, combining all the splits results in overall improvements or competitive performances across a broader range of tasks. We present the key statistics of all the splits in Table S2. In Figure S5, we present the training curves for different generative data augmentations, demonstrating stability during training across various data splits.

Table S1: Ablation study on sampling strategy used in generative data augmentation.  $C_i$  and  $C_s$  refer to CHAIR at instance and sentence-level; F1 refers to the F1-scores of all the discriminative tasks and HR refers to hallucination rate on generative tasks.

Data Split	CHAIR		AMBER		MME-Hall
	$C_i$ ↓	$C_s$ ↓	F1 ↑	HR ↓	Score ↑
Closed set	12.6	45.0	73.9	34.7	643.3
Open-set	<b>11.2</b>	<b>39.6</b>	73.1	33.3	643.3
Closed set + Open-set (50%)	<u>11.7</u>	41.8	79.8	<b>32.0</b>	643.3
Closed set + Open-set	12.6	43.6	74.1	34.0	<u>648.3</u>
Closed set + Open-set + Y-or-N (50%)	11.8	43.2	82.4	<u>32.2</u>	<u>641.0</u>
<b>Closed set + Open-set + Y-or-N</b>	<u>11.7</u>	<u>41.4</u>	<b>83.4</b>	<u>32.2</u>	<b>665.0</b>

Table S2: Key statistics of training samples used in DPA training.

Data Split	# Samples	# Avg. hallucinated instances per sample	Length (in words) Avg./Min./Max.
One-sentence caption	528	2.7	15/6/53
Short description	11573	6.9	42/12/128
Detailed description	8268	11.3	71/32/246
Yes-or-No (one word answer)	1510	1	1/1/1
Full	21874	8.1	49/1/246

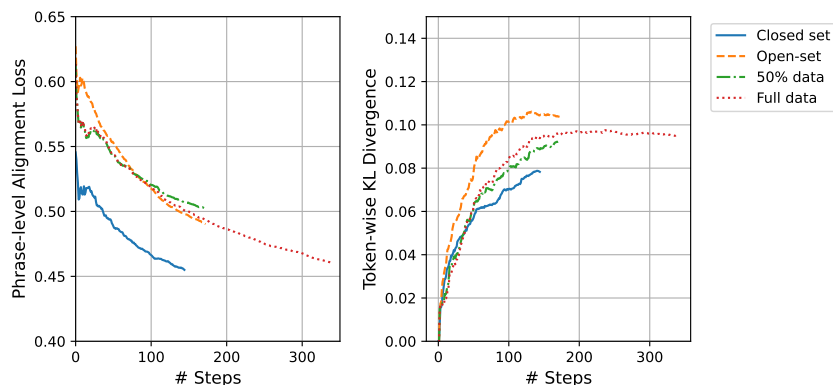


Figure S5: Training curves for different generative data augmentations using  $\alpha = 0.4$ .

Table S3: Ablation study on divergence measure using HALVA<sub>7B</sub>. **(a)** We find that using *seen* samples as the reference data for divergence measure achieve overall better performance. **(b)** Our study shows that initializing the reference model and the model being trained from the same checkpoint, achieves optimal performance.  $C_i$  and  $C_s$  refer to CHAIR at instance and sentence-level; F1 refers to the F1-scores of all the discriminative tasks and HR refers to hallucination rate in the image descriptions.

(a) Ablation study on reference data.

Ref. Data	CHAIR		AMBER		MME-Hall
	$C_i \downarrow$	$C_s \downarrow$	F1 $\uparrow$	HR $\downarrow$	Score $\uparrow$
Unseen data	12.7	47.4	81.7	34.7	<b>668.3</b>
Seen data	<b>11.7</b>	<b>41.4</b>	<b>83.4</b>	<b>32.2</b>	665.0

(b) Ablation study on reference model.

Ref. Model	CHAIR		AMBER		MME-Hall
	$C_i \downarrow$	$C_s \downarrow$	F1 $\uparrow$	HR $\downarrow$	Score $\uparrow$
<b>7B</b>	<b>11.7</b>	<b>41.4</b>	<b>83.4</b>	<b>32.2</b>	<b>665.0</b>
13B	12.4	45.2	80.1	34.7	640.0

### C.3 ABLATION ON DIVERGENCE MEASURE

**Reference data.** We experiment with the reference data that has been used to measure KL divergence with respect to the reference model. We briefly experiment in two setups:

- Unseen data: we directly use the vision-language instructions and *correct* responses as the reference samples.
- Seen data: we take a fraction of the instruction tuning dataset the base model is originally trained on, and use them as reference samples.

We perform this experiment on HALVA<sub>7B</sub> and the results are presented in Table S3 (a). The results demonstrate that using seen samples to measure divergence gives a better estimate of model state during training, and accordingly the tuned model overall performs better, across various benchmarks.

**Reference model.** By default, we initialize the reference model (the model kept frozen) and the online model (the model being trained) from the same checkpoint. Additionally, we experiment with initializing the reference model different than the model being trained. In particular, we experiment with training LLaVA<sub>7B</sub> while using LLaVA<sub>13B</sub> as the reference model. We find this interesting to explore as both LLaVA<sub>7B</sub> and LLaVA<sub>13B</sub> are originally trained in a similar setup, and LLaVA<sub>13B</sub> performs relatively better compared to the LLaVA<sub>7B</sub>, on most of the benchmarks (Liu et al., 2023c). The results presented in Table S3 (b) show that initializing the reference model and the online model from the same checkpoint, achieve optimal performance. We believe this is likely since the reference model initialized from an identical state of the model being trained, gives a true estimate of divergence and accordingly optimized model performs better across a variety of benchmarks.

### C.4 DETAILED RESULTS OF MME-HALL

In Table S4, we present the detailed results of the MME-Hall (Fu et al., 2023) benchmark across its four sub-categories: existence, count, position, and color. Our results indicate that DPA mitigates (or retains the same performance as the base model) object hallucination across different aspects, unlike prior finetuning methods such as HA-DPO (Zhao et al., 2023b) and EOS (Yue et al., 2024), or inference-based methods such as VCD (Leng et al., 2023) and Woodpecker (Yin et al., 2023), which either degrade overall performance or show improvement in one category but suffer in others.

Table S4: Detailed results on **MME-Hall**.

Method	Object ( $\uparrow$ )		Attribute ( $\uparrow$ )		Total ( $\uparrow$ )
	Existence	Count	Position	Color	
LLaVA-v1.5 <sub>7B</sub>	190.0	155.0	133.3	170.0	648.3
HA-DPO <sub>7B</sub>	190.0	133.3	136.7	158.3	618.3
EOS <sub>7B</sub>	190.0	138.3	118.3	160.0	606.7
VCD <sub>7B</sub>	184.7	138.3	128.7	153.0	604.7
Woodpecker <sub>7B</sub>	165.0	98.3	56.7	46.7	366.7
<b>HALVA<sub>7B</sub> (Ours)</b>	190.0	165.0	135.0	175.0	<b>665.0</b>
LLaVA-v1.5 <sub>13B</sub>	185.0	155.0	133.3	170.0	643.3
<b>HALVA<sub>13B</sub> (Ours)</b>	190.0	163.3	141.7	180.0	<b>675.0</b>
VILA-v1.5 <sub>13B/384</sub>	185.0	170.0	148.3	185.0	688.3
<b>HALVA<sub>13B/384</sub> (Ours)</b>	185.0	173.3	148.3	185.0	<b>691.7</b>

## C.5 DETAILED RESULTS OF MMHAL-BENCH

In Table S5, we present the detailed results of MMHal-Bench (Sun et al., 2023) across its eight sub-categories. Our proposed DPA demonstrates consistent effectiveness in mitigating object hallucinations in the following types: adversarial, comparison, relation, and holistic on both HALVA<sub>7B</sub> and HALVA<sub>13B</sub>. Additionally, DPA improves performance in 6 out of 8 subcategories for both the 13B variants. Moreover, recent hallucination mitigation methods such as HA-DPO and EOS prove ineffective in addressing such broad categories of hallucinations, even resulting in worsened baseline performance.

Table S5: Detailed results on MMHal-Bench.

Method	Overall Score ( $\uparrow$ )	Hall. Rate ( $\downarrow$ )	Score in Each Question Type ( $\uparrow$ )							
			Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
LLaVA-v1.5 <sub>7B</sub>	2.11 $\pm$ 0.06	0.56 $\pm$ 0.01	3.06 $\pm$ 0.27	1.00 $\pm$ 0.00	1.61 $\pm$ 0.05	1.97 $\pm$ 0.09	2.36 $\pm$ 0.05	3.20 $\pm$ 0.05	2.14 $\pm$ 0.30	1.53 $\pm$ 0.25
HA-DPO <sub>7B</sub>	1.97 $\pm$ 0.04	0.59 $\pm$ 0.01	3.56 $\pm$ 0.17	1.08 $\pm$ 0.09	1.14 $\pm$ 0.13	1.89 $\pm$ 0.21	2.22 $\pm$ 0.33	3.31 $\pm$ 0.10	1.42 $\pm$ 0.14	1.17 $\pm$ 0.00
EOS <sub>7B</sub>	2.03 $\pm$ 0.02	0.59 $\pm$ 0.02	2.69 $\pm$ 0.13	1.78 $\pm$ 0.09	1.89 $\pm$ 0.13	1.53 $\pm$ 0.18	2.09 $\pm$ 0.14	3.08 $\pm$ 0.30	1.67 $\pm$ 0.29	1.53 $\pm$ 0.09
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>2.25<math>\pm</math>0.10</b>	<b>0.54<math>\pm</math>0.01</b>	<b>2.78<math>\pm</math>0.09</b>	<b>1.47<math>\pm</math>0.18</b>	<b>1.97<math>\pm</math>0.13</b>	<b>1.89<math>\pm</math>0.05</b>	<b>3.03<math>\pm</math>0.21</b>	<b>3.20<math>\pm</math>0.05</b>	<b>2.42<math>\pm</math>0.43</b>	<b>1.22<math>\pm</math>0.27</b>
LLaVA-v1.5 <sub>13B</sub>	2.38 $\pm$ 0.02	0.50 $\pm$ 0.01	3.20 $\pm$ 0.05	2.53 $\pm$ 0.18	2.55 $\pm$ 0.05	2.20 $\pm$ 0.05	1.97 $\pm$ 0.05	3.33 $\pm$ 0.14	1.50 $\pm$ 0.22	1.72 $\pm$ 0.13
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>2.58<math>\pm</math>0.08</b>	<b>0.46<math>\pm</math>0.02</b>	<b>3.03<math>\pm</math>0.09</b>	<b>2.58<math>\pm</math>0.09</b>	<b>2.66<math>\pm</math>0.14</b>	<b>2.08<math>\pm</math>0.14</b>	<b>2.45<math>\pm</math>0.05</b>	<b>3.36<math>\pm</math>0.17</b>	<b>2.44<math>\pm</math>0.39</b>	<b>2.00<math>\pm</math>0.08</b>
VILA-v1.5 <sub>13B/384</sub>	2.58 $\pm$ 0.02	0.46 $\pm$ 0.01	3.36 $\pm$ 0.13	1.08 $\pm$ 0.09	3.39 $\pm$ 0.13	2.05 $\pm$ 0.05	2.97 $\pm$ 0.21	3.11 $\pm$ 0.05	2.19 $\pm$ 0.13	2.47 $\pm$ 0.05
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>2.58<math>\pm</math>0.06</b>	<b>0.45<math>\pm</math>0.01</b>	<b>3.11<math>\pm</math>0.05</b>	<b>1.47<math>\pm</math>0.05</b>	<b>3.47<math>\pm</math>0.05</b>	<b>2.08<math>\pm</math>0.00</b>	<b>3.11<math>\pm</math>0.13</b>	<b>3.19<math>\pm</math>0.13</b>	<b>1.64<math>\pm</math>0.24</b>	<b>2.58<math>\pm</math>0.09</b>

## C.6 DETAILED RESULTS OF HALLUSIONBENCH

In Table S6, we present the detailed results of HallusionBench (Liu et al., 2023a), which evaluates MLLMs beyond object hallucination, including those may cause by visual illusions and quantitative analysis form charts or graphs, among others. In addition to improving the overall performance, the results demonstrate the effectiveness of DPA on all the sub-categories (i.e., easy set, hard set) of HallusionBench as well. For example, we find that HALVA<sub>7B</sub> and HALVA<sub>13B</sub> substantially improve performance (4.34%-6.90%) on the *Hard Set* of HallusionBench, which consists of human-edited image-question pairs specially crafted to elicit hallucinations in MLLMs. We note that, in addition to hallucination mitigation, DPA helps MLLMs in reducing Yes/No bias. As discussed earlier, LLaVA-v1.5 is prone to answering ‘Yes’, in most cases. Our proposed DPA effectively reduces Yes/No bias from 0.31 to 0.17 and from 0.38 to 0.20 on HALVA<sub>7B</sub> and HALVA<sub>13B</sub>, respectively. Moreover, in the case of HALVA<sub>13B/384</sub>, the Yes/No bias is reduced from 0.19 to 0.02, with 0 being ideal.

Table S6: Detailed results on HallusionBench.

Method	Yes/No Bias		Question Pair Acc.	Fig. Acc.	Easy Acc.	Hard Acc.	All Acc.
	Pct. Diff ( $\sim$ 0)	FP Ratio ( $\sim$ 0.5)	( <i>qAcc</i> ) $\uparrow$	( <i>fAcc</i> ) $\uparrow$	( <i>Easy aAcc</i> ) $\uparrow$	( <i>Hard aAcc</i> ) $\uparrow$	( <i>aAcc</i> ) $\uparrow$
LLaVA-v1.5 <sub>7B</sub>	0.31 $\pm$ 0.00	0.79 $\pm$ 0.00	10.70 $\pm$ 0.13	19.65 $\pm$ 0.00	42.34 $\pm$ 0.13	41.47 $\pm$ 0.13	47.09 $\pm$ 0.14
HA-DPO <sub>7B</sub>	0.26	0.76	11.21	19.08	42.86	44.19	48.36
EOS <sub>7B</sub>	0.29	0.78	11.21	18.50	43.96	42.09	48.72
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>0.17<math>\pm</math>0.00</b>	<b>0.67<math>\pm</math>0.00</b>	<b>13.85<math>\pm</math>0.00</b>	<b>21.48<math>\pm</math>0.17</b>	<b>42.71<math>\pm</math>0.13</b>	<b>45.81<math>\pm</math>0.00</b>	<b>48.95<math>\pm</math>0.14</b>
LLaVA-v1.5 <sub>13B</sub>	0.38 $\pm$ 0.00	0.85 $\pm$ 0.00	8.79 $\pm$ 0.22	15.22 $\pm$ 0.17	44.25 $\pm$ 0.13	35.97 $\pm$ 0.13	46.50 $\pm$ 0.09
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>0.20<math>\pm</math>0.00</b>	<b>0.70<math>\pm</math>0.00</b>	<b>13.85<math>\pm</math>0.22</b>	<b>20.13<math>\pm</math>0.17</b>	<b>44.47<math>\pm</math>0.13</b>	<b>42.87<math>\pm</math>0.13</b>	<b>49.10<math>\pm</math>0.05</b>
VILA-v1.5 <sub>13B/384</sub>	0.19 $\pm$ 0.00	0.71 $\pm$ 0.00	18.90 $\pm$ 0.00	24.86 $\pm$ 0.29	52.38 $\pm$ 0.13	46.20 $\pm$ 0.27	55.39 $\pm$ 0.05
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>0.02<math>\pm</math>0.00</b>	<b>0.53<math>\pm</math>0.00</b>	<b>22.71<math>\pm</math>0.46</b>	<b>27.65<math>\pm</math>0.17</b>	<b>52.89<math>\pm</math>0.34</b>	<b>46.96<math>\pm</math>0.23</b>	<b>56.60<math>\pm</math>0.18</b>

## C.7 A CRITICAL ANALYSIS OF OUR PROPOSED DPA

Here, we critically assess whether the performance enhancement observed in our proposed DPA is attributable to generative data augmentation, the proposed training objective, or their combination. To investigate this, we apply our generative data augmentation directly to another finetuning-based hallucination mitigation approach, HA-DPO (Zhao et al., 2023b). In HA-DPO, correct and hallucinated pairs are employed to finetune MLLMs, aiming to maximize the reward margin between the correct responses and the hallucinated ones. Accordingly, we train HA-DPO by replacing their data with the output of our generative data augmentation module. We utilize the official code released by (Zhao et al., 2023b) and conduct hyper-parameter tuning (mainly with varying  $\beta$  and learning rate) ensure effective training. Subsequently, we evaluate the performance of the newly trained HA-DPO on both hallucination (CHAIR, AMBER, MME-Hall) and non-hallucination (MME) benchmarks. The results presented in Table S7 indicate that applying our proposed generative data augmentation to HA-DPO does not yield the same level of performance boost as HALVA. This confirms that the performance boost of our proposed method stems from a combination of the KL-regularized phrase-level alignment objective and the data



augmentation setup. Note that since our proposed method necessitates a pair of aligned correct and hallucinated phrases, and the descriptive responses utilized in HA-DPO do not meet this requirement, we are unable to apply DPA directly to their data.

Table S7: Effect of generative data augmentation on HA-DPO. Here, CHAIR, AMBER, and MME-Hall are hallucination benchmarks, and MME is a general vision-language benchmark.

	CHAIR ( $C_i$ ) ↓	AMBER F1 ↑	MME-Hall ↑	MME ↑
HA-DPO <sub>7B</sub>	<b>11.0</b>	78.1	618.3	1502.6
HA-DPO <sub>7B</sub> w/ Generative Data Aug.	14.6	77.7	631.7	1508.9
HALVA <sub>7B</sub>	11.7	<b>83.4</b>	<b>665.0</b>	<b>1527.0</b>

### C.8 RESULTS ON POPE

In addition to the hallucination benchmarks in the main paper, we also evaluate HALVA using POPE (Li et al., 2023d). While POPE is used in prior works, we note a few key limitations and find it to be a not well suited benchmark for evaluating MLLMs, as listed below. Please note that the similar concerns are also echoed in recent works (Wang et al., 2023a; Bai et al., 2024).

First, POPE employs a Yes-or-No protocol to check for existence of an object, but lacks coverage of other types of object hallucinations, such as object attributes (e.g., color, count) and object relations (e.g., position, environment). Second, the questions are formulated based on only 500 images and include a total of 79 unique objects, which fails to capture object hallucinations across diverse visual concepts. Third, POPE does not evaluate hallucinations in descriptive tasks (e.g., image description), where MLLMs tend to hallucinate more. These limitations led to introduction of more comprehensive benchmarks such as AMBER and MME among others, which we are used as the primary evaluation benchmarks in this work.

As shown in Table S8, we observe that while models such as GPT-4o and InternVL2 perform considerably better than others on MME and HallusionBench, they are not well-represented by POPE. Despite these shortcomings, we were able to obtain 87.1 and 87.9 for HALVA<sub>7B</sub> and HALVA<sub>13B</sub> using a different  $\alpha = 0.005$ .

Table S8: The results on POPE are presented. \* Results are obtained using a different  $\alpha$  than our default. † Added here for reference only, and should not be directly compared with 7B and 13 models, due to the large discrepancy in their model sizes.

Method	POPE (F1 ↑)	AMBER (F1 ↑)	HallusionBench (Acc. ↑)	MME-Hall (Score ↑)	MME (Score ↑)
LLaVA-v1.5 <sub>7B</sub>	85.9	74.7	47.1	684.3	1510.7
LLaVA-RLHF <sub>7B</sub>	81.5	76.3	43.0	493.3	1190.0
HA-DPO <sub>7B</sub>	86.9	78.1	48.4	618.3	1502.6
EOS <sub>7B</sub>	86.0	75.6	48.7	606.7	1424.4
<b>HALVA<sub>7B</sub> (Ours)</b>	<b>84.8/87.1*</b>	<b>83.4</b>	<b>49.0</b>	<b>665.0</b>	<b>1527.0</b>
LLaVA-v1.5 <sub>13B</sub>	85.9	73.1	46.5	643.3	1530.1
LLaVA-RLHF <sub>13B</sub>	81.9	83.7	46.4	585.0	1367.7
<b>HALVA<sub>13B</sub> (Ours)</b>	<b>84.9/87.9*</b>	<b>86.5</b>	<b>49.1</b>	<b>675.0</b>	<b>1544.0</b>
VILA-v1.5 <sub>13B</sub>	86.3	82.2	55.4	688.3	1569.6
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>86.1</b>	<b>87.9</b>	<b>56.6</b>	<b>691.7</b>	<b>1575.7</b>
GPT-4o <sup>†</sup> (v.0513, detail-high)	85.6	-	55.0	-	2310.3
InternVL2 <sub>40B</sub> <sup>†</sup> (Chen et al., 2024)	81.9	-	56.5	-	2293.1

### C.9 RESULTS ON LINGUISTIC QUALITY

To analyse whether DPA training have an adverse affect on the linguistic quality of the responses generated by MLLMs, we evaluate the responses on four aspects: grammatical correctness, fluency, detailedness, and choice of words. Since there is no standard or commonly used benchmark for these tasks, we use randomly selected 100 detailed image descriptions (a subset from the AMBER (Wang et al., 2023a) image description task) generated

by LLaVA 1.5<sub>7B</sub> and HALVA<sub>7B</sub>, with GPT-4o-mini as the judge to rate them on a scale of 0 to 10. The template used in evaluation is presented in Figure S6. As shown in Table S9, HALVA<sub>7B</sub> exhibits the same performance as LLaVA 1.5<sub>7B</sub>.

Model	Grammatical Correctness	Fluency	Detailedness	Choice of Words
LLaVA 1.5 <sub>7B</sub>	9.90 ± 0.30	9.64 ± 0.52	8.37 ± 0.48	8.93 ± 0.26
<b>HALVA<sub>7B</sub> (Ours)</b>	9.99 ± 0.10	9.51 ± 0.50	8.35 ± 0.48	8.99 ± 0.23

Table S9: Results on linguistic qualities of the responses.

Following is a detailed image description.  
 Your task is to assess the response on the following criteria:  
 1. Grammatical Correctness: Analyze the response for grammar, punctuation, and syntax accuracy.  
 2. Fluency: Evaluate whether the response flows smoothly, reads naturally, and maintains coherence throughout.  
 3. Detailedness: Check if the response provides sufficient and relevant detail to address the topic comprehensively, without redundancy or unnecessary information.  
 4. Choice of Words: Assess if the words used are appropriate, varied, and effectively convey the intended message.

Rate each criterion on a scale from 0 to 10, where 0 indicates poor quality and 10 signifies an excellent response.

Here is the image description to evaluate:

{description}

Your response should be in this format:

Grammatical Correctness: SCORE  
 Fluency: SCORE  
 Detailedness: SCORE  
 Choice of Words: SCORE

Figure S6: The template for evaluating the linguistic quality of the responses.

## D IMPLEMENTATION DETAILS

### D.1 TRAINING HYPERPARAMETERS

The details of training hyperparameters used in DPA training is presented in Table S10.

Table S10: Details of training hyperparameters used in DPA training.

	HALVA <sub>7B</sub>	HALVA <sub>13B</sub>	HALVA <sub>13B/384</sub>
Base model	LLaVA-v1.5 <sub>7B</sub>	LLaVA-v1.5 <sub>13B</sub>	VILA-v1.5 <sub>13B</sub>
LLM	Vicuna-v1.5 <sub>7B</sub>	Vicuna-v1.5 <sub>13B</sub>	
Vision encoder	CLIP ViT-L <sub>336/14</sub>		SigLIP-L-400M
Trainable module	LoRA in LLM and everything else is kept frozen		
LoRA setup (Hu et al., 2021)	rank=128, alpha=256		
Learning rate	5e-6		2.5e-5
Learning rate scheduler		Cosine	
Optimizer	AdamW (Loshchilov & Hutter, 2017)		
Weight decay		0.	
Warmup ratio		0.03	
Epoch		1 (342 steps)	
Batch size per GPU		16	
Batch size (total)		64	
$\alpha$ (loss coefficient)	0.4	0.5	0.2
Memory optimization	Zero stage 3 (Ren et al., 2021; Rajbhandari et al., 2021)		
Training time	1.5 hrs	3 hrs.	3 hrs.

### D.2 LICENSES OF EXISTING ASSETS USED

For images, we use publicly-available Visual Genome dataset (Krishna et al., 2017). This dataset can be downloaded from <https://homes.cs.washington.edu/~ranjay/visualgenome/api.html> and is licensed under a Creative Commons Attribution 4.0 International License.

For the base MLLM, we use LLaVA-v1.5 (Liu et al., 2023c) and VILA-v1.5 (Lin et al., 2024). LLaVA-v1.5 is publicly available and its Apache license 2.0 can be found at <https://github.com/haotian-liu/LLaVA/blob/main/LICENSE>. VILA-v1.5 is publicly available and its Apache license 2.0 can be found at <https://github.com/NVlabs/VILA/blob/main/LICENSE>. The weights used in this work are available as follows:

- LLaVA-v1.5<sub>7B</sub>: <https://huggingface.co/liuhaotian/llava-v1.5-7b>
- LLaVA-v1.5<sub>13B</sub>: <https://huggingface.co/liuhaotian/llava-v1.5-13b>
- VILA-v1.5<sub>13B</sub>: <https://huggingface.co/Efficient-Large-Model/VILA1.5-13b>

### D.3 GENERATIVE DATA AUGMENTATION SETUP

We present the prompt templates that are used to prepare correct and hallucinated descriptions in Figures S7, S8, and S9. The full list of instructions used in generating image descriptions is presented in Figure S10. We leverage Gemini Vision Pro (gemini-1.0-pro-vision) in preparing the responses. Complete examples depicting the pipeline of generating correct descriptions, closed-set hallucinated descriptions, and open-set descriptions are presented in Figures S11, S12, and S13. We present additional examples of training samples for one sentence image caption, short image description, detailed image description, and Yes-or-No questions in Figures S14, S15, S16, and S17, respectively.

1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

```
# Input  
## Image  
<Image>  
## Text  
Here are the region descriptions of the given image.  
  
<Region description 1>  
<Region description 2>  
<Region description 3>  
...  
  
The descriptions are the ground truth information for the image.  
  
Based on the given region descriptions,  
write a response for the following question.  
  
Question:  
  
<Instruction>  
  
The response must be correct and has strong readability.  
  
Do NOT add any new information or additional details.  
  
# Output  
  
<Correct description>
```

Figure S7: The **template** for generating the **correct** image descriptions.

1512  
1513  
1514  
1515  
1516  
1517  
1518  
1519  
1520  
1521  
1522  
1523  
1524  
1525  
1526  
1527  
1528  
1529  
1530  
1531  
1532  
1533  
1534  
1535  
1536  
1537  
1538  
1539  
1540  
1541  
1542  
1543  
1544  
1545  
1546  
1547  
1548  
1549  
1550  
1551  
1552  
1553  
1554  
1555  
1556  
1557  
1558  
1559  
1560  
1561  
1562  
1563  
1564  
1565

**# Input**

**## Text**

The given text is a description of an image.  
<Correct description>

Please rewrite the given text by replacing the mentioned words  
with those from the given options.

Please choose the replacement that sounds the most appropriate.

Replace the word: <ground-truth object 1> - with a word from the  
given options: <list of hallucinated objects 1>

Replace the word: <ground-truth object 2> - with a word from the  
given options: <list of hallucinated objects 2>

...

The description should logically make sense, the style of the  
new text should be the same as the original text, and  
has strong readability.

Please make sure to NOT include the following words in the  
description: <list of ground-truth objects>.

Your response should only include the new description  
and nothing else.

**# Output**

<Hallucinated description>

Figure S8: The **template** for generating the **closed-set** hallucinated descriptions.

1566  
1567  
1568  
1569  
1570  
1571  
1572  
1573  
1574  
1575  
1576  
1577  
1578  
1579  
1580  
1581  
1582  
1583  
1584  
1585  
1586  
1587  
1588  
1589  
1590  
1591  
1592  
1593  
1594  
1595  
1596  
1597  
1598  
1599  
1600  
1601  
1602  
1603  
1604  
1605  
1606  
1607  
1608  
1609  
1610  
1611  
1612  
1613  
1614  
1615  
1616  
1617  
1618  
1619

```
# Input  
  
## Text  
  
The given text is a description of an image.  
<Correct description>  
  
Please rewrite the given text by replacing the mentioned object  
with another object of similar types or categories.  
For example, an animal can be replaced with another  
animal or one type of vehicle can be replaced by another type  
of vehicle and so on.  
  
The description should logically makes sense, the style of  
the new text should be the same as the original text,  
and has strong readability.  
  
Your response should only include the new description and  
nothing else.  
  
The following objects need to be replaced:  
<list of ground-truth objects>.  
  
# Output  
  
<Hallucinated description>
```

Figure S9: The **template** for generating the **open-set** hallucinated descriptions.

1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631  
1632  
1633  
1634  
1635  
1636  
1637  
1638  
1639  
1640  
1641  
1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653  
1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666  
1667  
1668  
1669  
1670  
1671  
1672  
1673

**# Instructions for one sentence caption:**

Provide a one-sentence caption for the provided image.

**# Instructions for short description:**

Describe the image concisely.  
Provide a brief description of the given image.  
Offer a succinct explanation of the picture presented.  
Summarize the visual content of the image.  
Give a short and clear explanation of the subsequent image.  
Share a concise interpretation of the image provided.  
Present a compact description of the photo's key features.  
Relay a brief, clear account of the picture shown.  
Render a clear and concise summary of the photo.  
Write a terse but informative summary of the picture.  
Create a compact narrative representing the image presented.  
Please provide a short description of this image.

**# Instructions for detailed description:**


Provide a detailed description of the given image.  
Give an elaborate explanation of the image you see.  
Share a comprehensive rundown of the presented image.  
Offer a thorough analysis of the image.  
Explain the various aspects of the image before you.  
Clarify the contents of the displayed image with great detail.  
Characterize the image using a well-detailed description.  
Break down the elements of the image in a detailed manner.  
Walk through the important details of the image.  
Portray the image with a rich, descriptive narrative.  
Narrate the contents of the image with precision.  
Analyze the image in a comprehensive and detailed manner.  
Illustrate the image through a descriptive explanation.  
Examine the image closely and share its details.  
Write an exhaustive depiction of the given image.  
Write a detailed description of the given image.

Figure S10: **Instructions** for different types of image descriptions. These instructions are directly taken from (Liu et al., 2024), we list them here for the sake of completeness.

1674  
1675  
1676  
1677  
1678  
1679  
1680  
1681  
1682  
1683  
1684  
1685  
1686  
1687  
1688  
1689  
1690  
1691  
1692  
1693  
1694  
1695  
1696  
1697  
1698  
1699  
1700  
1701  
1702  
1703  
1704  
1705  
1706  
1707  
1708  
1709  
1710  
1711  
1712  
1713  
1714  
1715  
1716  
1717  
1718  
1719  
1720  
1721  
1722  
1723  
1724  
1725  
1726  
1727

**# Input**

**## Image**



**## Text**

Here are the region descriptions of the given image.

a man jumping on a skateboard  
a skate park with colorful graffiti  
white clouds over mountains  
man in white shirt and blue jeans  
skating ramp with graffiti markings  
Skateboarder making a jump  
... 45 more

The descriptions are the ground truth information for the image.

Based on the given region descriptions,  
write a response for the following question.

Question: Walk through the important details of the image.

The response must be correct and has strong readability.

Do NOT add any new information or additional details.

**# Output**

A young man in a white shirt and blue jeans is skateboarding in a skate park. The skate park has colorful graffiti on the walls and ramps. There are stairs and a fence behind the skate park. There are also some trees and buildings in the background. The sky is blue with some clouds.

Figure S11: A **complete example** of generating the **correct** image descriptions.



1728  
1729  
1730  
1731  
1732  
1733  
1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781

**# Input**

**## Image**



**## Text**

The given text is a description of an image.

Correct description:

The image shows a variety of fruits on wooden tables at a farmer's market. There are pears, apples, oranges, and pineapples. The fruits are arranged in boxes and crates. There is a price sign on some of the boxes.

Please rewrite the given text by replacing the mentioned words with those from the given options.

Please choose the replacement that sounds the most appropriate.

Replace the word: fruit - with a word from the given options:  
plate, leaf, food, basket, vegetable

Replace the word: apple - with a word from the given options:  
table, banana, root, bowl, shirt

Replace the word: pear - with a word from the given options:  
tomato, gummed label, container, writing, hair

The description should logically make sense, the style of the new text should be the same as the original text, and has strong readability.

Please make sure to NOT include the following words in the description: apple, box, crate, fruit, ground, letter, line, pear, tree trunk, wall, word.

Your response should only include the new description and nothing else.

**# Output**

The image displays an assortment of vegetables on wooden tables at a farmer's market. There are tomatoes, bananas, oranges, and pineapples. The vegetables are organized in containers and signs. There is a price tag on some of the containers.

Figure S12: A complete example of generating closed set hallucinated descriptions is provided. The image is not fed to Gemini in generating hallucinated response, shown here for reference.

1782  
1783  
1784  
1785  
1786  
1787  
1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830  
1831  
1832  
1833  
1834  
1835

**# Input**

**## Image**



**## Text**

The given text is a description of an image.

Correct description:

A young man in a white shirt and blue jeans is skateboarding in a skate park. The skate park has colorful graffiti on the walls and ramps. There are stairs and a fence behind the skate park. There are also some trees and buildings in the background. The sky is blue with some clouds.

Please rewrite the given text by replacing the mentioned object with another object of similar types or categories. For example, an animal can be replaced with another animal or one type of vehicle can be replaced by another type of vehicle and so on.

The description should logically makes sense, the style of the new text should be the same as the original text, and has strong readability.

Your response should only include the new description and nothing else.

The following objects need to be replaced:  
advertisement, arm, baseball cap, building, cloud, design, distance, fence, head, jean, leaf, man, mountain, park, pole, ramp, shadow, sign, skate park, skateboard, skating, sky, stair, t-shirt, tree, trick, wheel.

**# Output**

A young woman in a black dress and red sneakers is rollerblading in a roller rink. The roller rink has vibrant murals on the walls and floors. There are benches and a railing behind the roller rink. There are also some plants and shops in the background. The ceiling is white with some lights.

Figure S13: A complete example of generating open-set hallucinated descriptions is provided. The image is not fed to Gemini in generating hallucinated response, shown here for reference.

1836  
1837  
1838  
1839  
1840  
1841  
1842  
1843  
1844  
1845  
1846  
1847  
1848  
1849  
1850  
1851  
1852  
1853  
1854  
1855  
1856  
1857  
1858  
1859  
1860  
1861  
1862  
1863  
1864  
1865  
1866  
1867  
1868  
1869  
1870  
1871  
1872  
1873  
1874  
1875  
1876  
1877  
1878  
1879  
1880  
1881  
1882  
1883  
1884  
1885  
1886  
1887  
1888  
1889



**Question:** Provide a one-sentence caption for the provided image.

**Correct:** There are three people holding and using their black smartphones.

**Hallucinated:** There are three people holding and using their black tablets.



**Question:** Provide a one-sentence caption for the provided image.

**Correct:** The image shows a variety of donuts on metal shelves in a donut shop.

**Hallucinated:** The image depicts an assortment of cupcakes on wooden shelves in a cupcake shop.



**Question:** Please provide a short description of this image.

**Correct:** A man is snowboarding down a snowy slope at night.

**Hallucinated:** A person is snowboarding down a snowy hill at night.



**Question:** Provide a one-sentence caption for the provided image.

**Correct:** The image shows a blonde woman wearing a pink dress with a red bow in her hair.

**Hallucinated:** The image displays a blonde woman wearing a pink gown with a red hat on her head.

Figure S14: Examples of one sentence image captions used in DPA training.

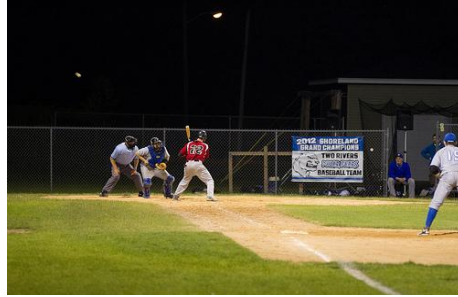
1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897  
1898  
1899  
1900  
1901  
1902  
1903  
1904  
1905  
1906  
1907  
1908  
1909  
1910  
1911  
1912  
1913  
1914  
1915  
1916  
1917  
1918  
1919  
1920  
1921  
1922  
1923  
1924  
1925  
1926  
1927  
1928  
1929  
1930  
1931  
1932  
1933  
1934  
1935  
1936  
1937  
1938  
1939  
1940  
1941  
1942  
1943



**Question:** Provide a brief description of the given image.

**Correct:** A black **cat** is lying on the ground next to a small brown **bird**. The **cat** has its eyes open and is looking at the **bird**.

**Hallucinated:** A black **dog** is lying on the **grass** next to a small brown **leaf**. The **dog** has its eyes open and is looking at the **leaf**.



**Question:** Write a terse but informative summary of the picture.

**Correct:** The image is a night view of a **baseball** game. There are two **baseball** players, one is the **batter** and the other is the **catcher**.

**Hallucinated:** The image is a night view of a **hockey** game. There are two hockey players, one is the **shooter** and the other is the **goalie**.



**Question:** Provide a brief description of the given image.

**Correct:** Two boys are standing in a **baseball** field. They are wearing **baseball** uniforms and holding **baseball** mitts. The boy on the left is wearing a **green** and white uniform and the boy on the right is wearing a **black** and **white** uniform.

**Hallucinated:** Two children are standing in a **soccer** field. They are wearing **soccer** uniforms and holding **soccer** balls. The child on the left is wearing a **blue** and white uniform and the child on the right is wearing a **red** and **black** uniform.



**Question:** Summarize the visual content of the image.

**Correct:** A **brown** and **white** **bulldog** is standing on a **blue** **surfboard** in the **ocean**. The **bulldog** is looking at the camera with an **overbite**. There is a big **splash** of **water** in front of the **surfboard**.

**Hallucinated:** A **gray** and **white** **cat** is standing on a **yellow** **skateboard** in the **snow**. The **cat** is looking at the camera with a **snaggletooth**. There is a big pile of **snow** in front of the **skateboard**.

Figure S15: Examples of short image descriptions used in DPA training.

1944  
1945  
1946  
1947  
1948  
1949  
1950  
1951  
1952  
1953  
1954  
1955  
1956  
1957  
1958  
1959  
1960  
1961  
1962  
1963  
1964  
1965  
1966  
1967  
1968  
1969  
1970  
1971  
1972  
1973  
1974  
1975  
1976  
1977  
1978  
1979  
1980  
1981  
1982  
1983  
1984  
1985  
1986  
1987  
1988  
1989  
1990  
1991  
1992  
1993  
1994  
1995  
1996  
1997



**Question:** Illustrate the image through a descriptive explanation.

**Correct:** There are a few motorcycles parked in a parking lot. There is a man standing behind one of the motorcycles. He is looking at the motorcycle. The motorcycle is orange and black. It has a chrome exhaust pipe. There are some trees and buildings in the background.

**Hallucinated:** There are a few trucks parked in a parking lot. There is a person standing behind one of the trucks. He is looking at the truck. The truck is orange and black. It has a chrome license plate. There are some plants and houses in the background.



**Question:** Clarify the contents of the displayed image with great detail.

**Correct:** A yellow container house is placed on the sidewalk. The house has a red and yellow sign on the front. There are some buckets in front of the house. A man is squatting on the sidewalk next to the house. There are green bushes and a brick sidewalk.

**Hallucinated:** A yellow trailer home is placed on the grass. The home has a blue and yellow flag on the front. There are some barrels in front of the home. A woman is kneeling on the grass next to the home. There are red flowers and a stone path.

Figure S16: Examples of detailed image descriptions used in DPA training.

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051



**Question:** Describe the image in detail.

**Correct:** A woman is sitting behind a table in a park. There is a sign on the table that says Just see the results and its free. The woman is wearing a blue jean jacket and a beaded necklace. There is a stack of pamphlets on the table. The table is covered with a red tablecloth. The ground is covered with brown leaves. There is a large tree in the background.

**Hallucinated:** A man is sitting behind a chair in a garden. There is a poster on the chair that says Just see the outcome and its free. The man is wearing a black leather coat and a golden chain. There is a pile of leaflets on the chair. The chair is covered with a blue sheet. The floor is covered with green grass. There is a tall building in the background.



**Question:** Explain the various aspects of the image before you.

**Correct:** This image shows a brown horse standing in a stall. The horse has a white blaze on its forehead and white socks on its back legs. The stall is made of cinder blocks and has a metal gate. There is a pile of manure in the stall. The horse is standing on dirt. There are green hills in the background.

**Hallucinated:** This image depicts a black cow standing in a pen. The cow has a black spot on its forehead and black socks on its front legs. The pen is made of wooden planks and has a wooden gate. There is a pile of hay in the pen. The cow is standing on straw. There are brown hills in the background.

Figure S16 (Continued): Examples of detailed image descriptions used in DPA training.

2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097  
2098  
2099  
2100  
2101  
2102  
2103  
2104  
2105



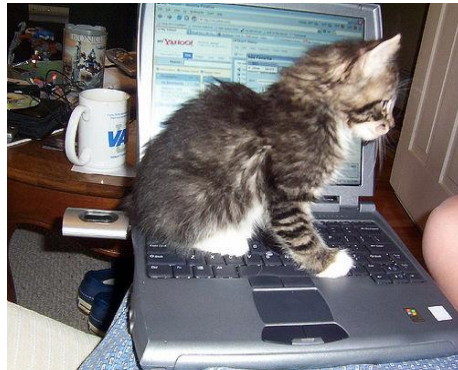
**Question:** Are there a total of nine benches visible in the scene? Please answer in one word Yes or No.  
**Correct:** No  
**Hallucinated:** Yes



**Question:** Is there a nightstand on the left side of the bed? Please answer in one word Yes or No.  
**Correct:** Yes  
**Hallucinated:** No



**Question:** Is there a person located on the left side of the image? Please answer in one word Yes or No.  
**Correct:** No  
**Hallucinated:** Yes



**Question:** Is the cup located on the left side of the table? Please answer in one word Yes or No.  
**Correct:** Yes  
**Hallucinated:** No


Figure S17: Examples of **Yes-or-No** questions used in DPA training.

2106  
2107  
2108  
2109  
2110  
2111  
2112  
2113  
2114  
2115  
2116  
2117  
2118  
2119  
2120  
2121  
2122  
2123  
2124  
2125  
2126  
2127  
2128  
2129  
2130  
2131  
2132  
2133  
2134  
2135  
2136  
2137  
2138  
2139  
2140  
2141  
2142  
2143  
2144  
2145  
2146  
2147  
2148  
2149  
2150  
2151  
2152  
2153  
2154  
2155  
2156  
2157  
2158  
2159

## E QUALITATIVE RESULTS

### E.1 QUALITATIVE COMPARISON BETWEEN HALVA AND LLaVA-v1.5

Describe the image in detail.



User

The image features a dirty, old toilet sitting in a dirt-covered area, possibly an outhouse or an abandoned building. The toilet is positioned in the center of the scene, surrounded by dirt and debris. There are several **carrots** scattered around the area, with some close to the toilet and others further away. The overall scene appears to be quite unclean and neglected.

The image features a dirty, old toilet sitting in a dirt-covered room. The toilet is positioned in the middle of the room, surrounded by a pile of dirt and debris. The walls of the room are made of cinder blocks, giving the space a somewhat unfinished appearance.

Figure S18: Comparing detailed image descriptions between HALVA [🤖] and its base model LLaVA-v1.5 [🤖]. Unlike HALVA, LLaVA-v1.5 hallucinates 'carrot' in the image description. Hallucinations are highlighted in red.



2160  
2161  
2162  
2163  
2164  
2165  
2166  
2167  
2168  
2169  
2170  
2171  
2172  
2173  
2174  
2175  
2176  
2177  
2178  
2179  
2180  
2181  
2182  
2183  
2184  
2185  
2186  
2187  
2188  
2189  
2190  
2191  
2192  
2193  
2194  
2195  
2196  
2197  
2198  
2199  
2200  
2201  
2202  
2203  
2204  
2205  
2206  
2207  
2208  
2209  
2210  
2211  
2212  
2213



Describe the image in detail.

User



The image captures a busy street at night with a car driving down the road. The car is positioned in the middle of the scene, and its headlights are on, illuminating the street. There are several traffic lights visible in the image, with one located near the center and another towards the right side of the scene.

In addition to the car, there are two other people in the scene, one standing closer to the left side and the other near the center. A handbag can be seen on the ground, possibly belonging to one of the pedestrians. The overall atmosphere of the image is bustling and lively, with the car and people navigating the busy street at night.




The image captures a busy street at night with a car driving down the road. The street is filled with various signs, including a traffic light and a street sign, which are both lit up. The traffic light is positioned near the center of the scene, while the street sign is located on the right side.

There are also two other cars visible in the scene, one behind the main car and another further down the road. The street appears to be in a foreign country, as indicated by the presence of foreign signs.


Figure S19: Comparing detailed image descriptions between HALVA [🤖] and its base model LLaVA-v1.5 [🤖]. LLaVA-v1.5 wrongfully describes the position of the ‘cars’ and ‘traffic lights’. Moreover, it hallucinates the presence of ‘people’, ‘pedestrian’, and ‘handbag’. In contrast, HALVA correctly identifies the relative positions of the ‘traffic lights’ and ‘cars’, while accurately describing the image. Hallucinations are highlighted in red.

2214  
2215  
2216  
2217  
2218  
2219  
2220  
2221  
2222  
2223  
2224  
2225  
2226  
2227  
2228  
2229  
2230  
2231  
2232  
2233  
2234  
2235  
2236  
2237  
2238  
2239  
2240  
2241  
2242  
2243  
2244  
2245  
2246  
2247  
2248  
2249  
2250  
2251  
2252  
2253  
2254  
2255  
2256  
2257  
2258  
2259  
2260  
2261  
2262  
2263  
2264  
2265  
2266  
2267

Describe the image in detail.




User



The image captures a baseball game in progress, with a pitcher standing on the pitcher's mound, preparing to throw the ball. The pitcher is wearing a red shirt and is in the process of winding up for the pitch.

In the scene, there are several other people present, including a batter, a catcher, and an umpire. The batter is holding a baseball bat, ready to swing at the incoming pitch. The catcher is positioned behind the batter, wearing a baseball glove, and the umpire is standing nearby, observing the game closely.

The baseball is visible in the air, close to the pitcher, as the pitcher prepares to release it towards the batter



The image captures a baseball pitcher in the middle of a pitch on a baseball field. The pitcher is wearing a red shirt and is in the process of throwing the ball. He is standing on the pitcher's mound, which is a raised area in the center of the field.

The baseball can be seen in the air, close to the pitcher's hand, as he releases it. The pitcher's baseball glove is also visible, positioned near his other hand. The scene is set against a backdrop of a lush green field, showcasing the baseball diamond and the pitcher's skillful performance.

Figure S20: Comparing detailed image descriptions between HALVA [🤖] and its base model LLaVA-v1.5 [🤖]. LLaVA-v1.5 exhibits object hallucination that may happen due to object co-occurrence in the training data, such as the occurrence of 'pitcher' with 'batter', 'catcher', and 'umpire'. HALVA effectively mitigates such hallucination and accurately describes the image with necessary details. Hallucinations are highlighted in red.

2268  
2269  
2270  
2271  
2272  
2273  
2274  
2275  
2276  
2277  
2278  
2279  
2280  
2281  
2282  
2283  
2284  
2285  
2286  
2287  
2288  
2289  
2290  
2291  
2292  
2293  
2294  
2295  
2296  
2297  
2298  
2299  
2300  
2301  
2302  
2303  
2304  
2305  
2306  
2307  
2308  
2309  
2310  
2311  
2312  
2313  
2314  
2315  
2316  
2317  
2318  
2319  
2320  
2321



User: Are there three people in this image?

LLaVA-v1.5: Yes, there are three people in this image.

HALVA: Yes, there are three people in this image.

The same image given in a separate trial.

User: Are there four people in this image?

LLaVA-v1.5: Yes, there are four people in this image.

HALVA: No, there are three people in this image.



User: Is there direct contact between the dog and beach?

LLaVA-v1.5: Yes, the dog is running directly on the beach, with its paws touching the sand.

HALVA: Yes, the dog is running directly on the beach, with its paws touching the sand.

The same image given in a separate trial.

User: Is there direct contact between the dog and sea?

LLaVA-v1.5: Yes, the dog is running on the beach near the water, which suggests that it is in direct contact with the sea.

HALVA: No, the dog is not directly touching the sea. It is running on the beach near the water, but there is a gap between the dog and the water.

Figure S21: HALVA [🤖] does not exhibit a tendency to answer ‘Yes’ for any questions asked, like LLaVA-v1.5 [🗃️]. Moreover, HALVA can accurately lay out the details of the image to further support the answer. As presented in the right side example, HALVA answers with “the dog is not directly touching the sea. It is running on the beach near the water, but there is a gap between the dog and the water.”. Hallucinations are highlighted in red.

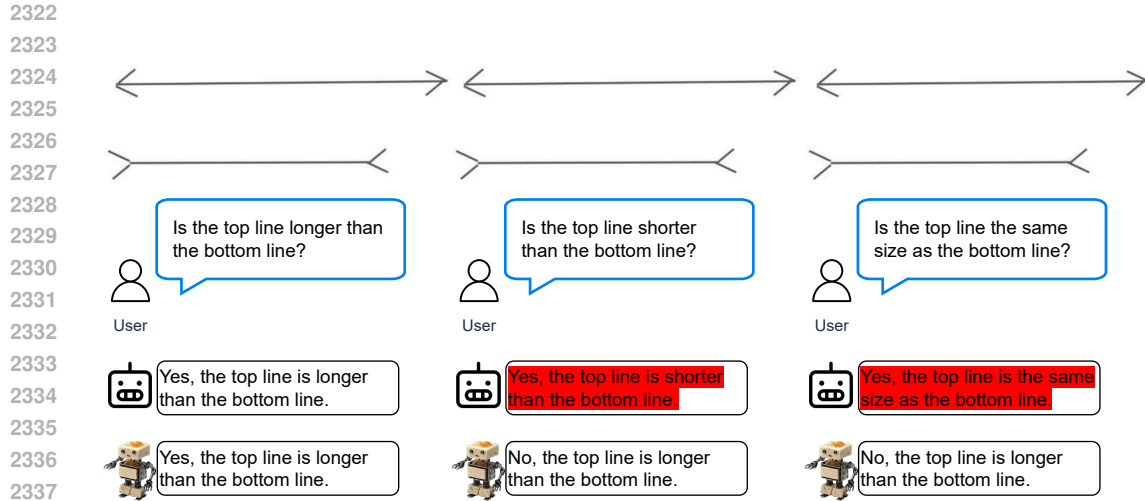


Figure S22: Examples illustrating the effectiveness of our proposed DPA in mitigating hallucinations that arise due to visual illusions. HALVA [🤖] does not exhibit a tendency to answer ‘Yes’ to any questions asked, like LLaVA-v1.5 [👤]. Moreover, HALVA exhibits consistency in its responses unlike LLaVA-v1.5. Hallucinations are highlighted in red.

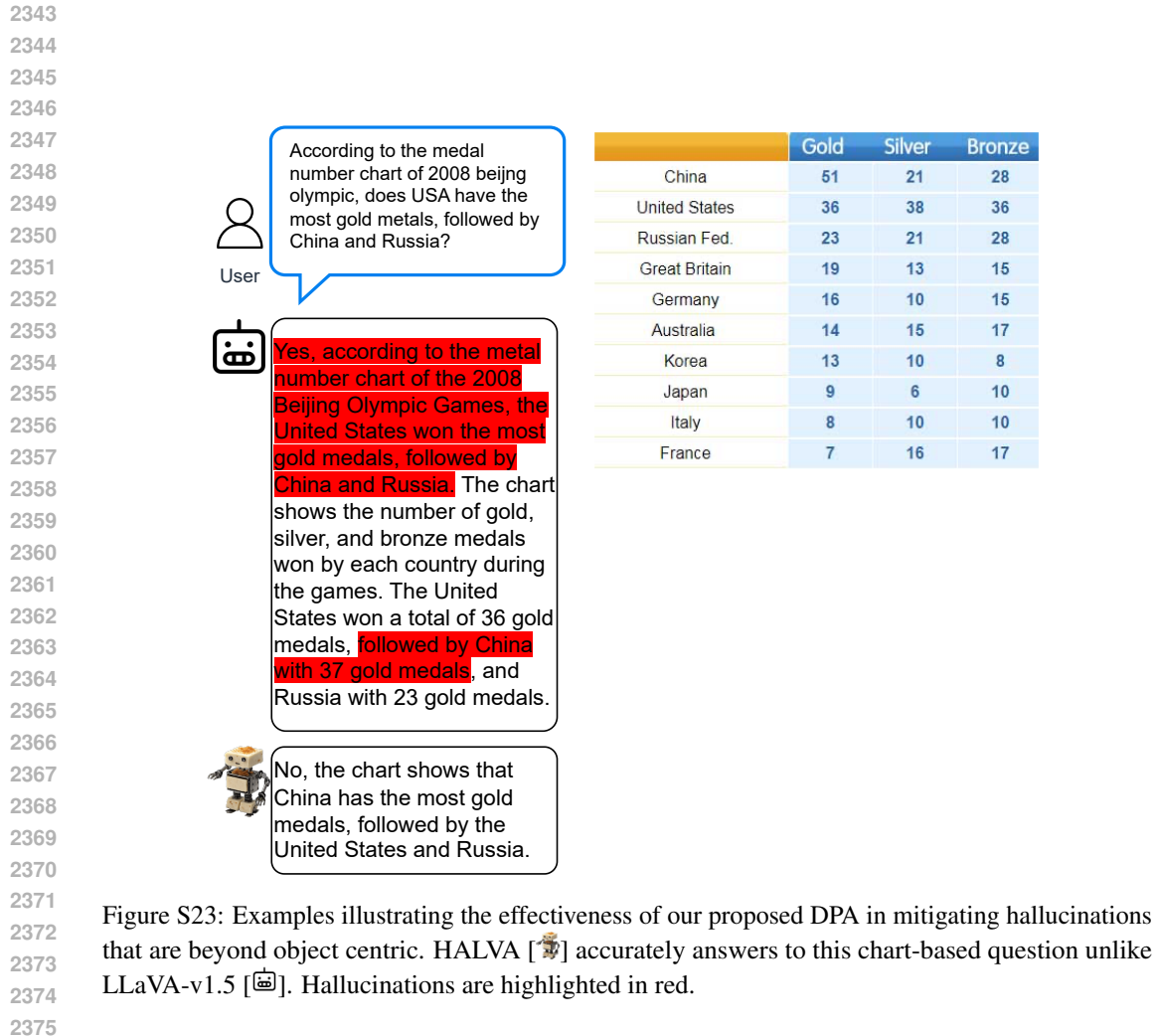


Figure S23: Examples illustrating the effectiveness of our proposed DPA in mitigating hallucinations that are beyond object centric. HALVA [🤖] accurately answers to this chart-based question unlike LLaVA-v1.5 [👤]. Hallucinations are highlighted in red.

2376  
2377  
2378  
2379  
2380  
2381  
2382  
2383  
2384  
2385  
2386  
2387  
2388  
2389  
2390  
2391  
2392  
2393  
2394  
2395  
2396  
2397  
2398  
2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2417  
2418  
2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429

E.2 QUALITATIVE COMPARISON BETWEEN HALVA AND VILA-v1.5



Is there one knife in this image?

User

Yes, there is one knife in the image, which the man is holding.

Yes, there is one knife in the image.

Is there direct contact between the dog and skate?

User

Yes, the dog is standing on top of the skateboard, with its paws on the skateboard.

Yes, the dog is standing on the skateboard, with its paws on the skateboard.

The same image given in a separate trial.

The same image given in a separate trial.

Are there two knives in this image?

User

Yes, there are two knives in the image. One knife is being held by the man, and the other knife is sticking out of the coconut.

No, there is only one knife in the image, which the man is holding.

Is there direct contact between the dog and skate?

User

Yes, the dog is standing on the skateboard, which is in direct contact with the ground.

No, the dog is not directly touching the ground. It is standing on a skateboard, which is placed on the ground.

Figure S24: HALVA [🤖] does not exhibit a tendency to answer ‘Yes’ for any questions asked, like VILA-v1.5 [🤖]. Moreover, HALVA can accurately lay out the details of the image to further support the answer. As presented in the right side example, HALVA answers with “the dog is not directly touching the ground. It is standing on a skateboard, which is placed on the ground.”. Hallucinations are highlighted in red.

2430  
2431  
2432  
2433  
2434  
2435  
2436  
2437  
2438  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2457  
2458  
2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2476  
2477  
2478  
2479  
2480  
2481  
2482  
2483



Are there two horses in this image?

User

Yes, there are two horses in this image.

Yes, there are two horses in this image.

The same image given in a separate trial.

Is there direct contact between the giraffe and grass?

User

Yes, the giraffe is walking through the grass, which means it is in direct contact with the grass.

Yes, the giraffe is walking through the grass, which means it is in direct contact with the grass.

The same image given in a separate trial.

Are there three horses in this image?

User

Yes, there are three horses in this image.

No, there are two horses in the image.

Is there direct contact between the tree and giraffe?

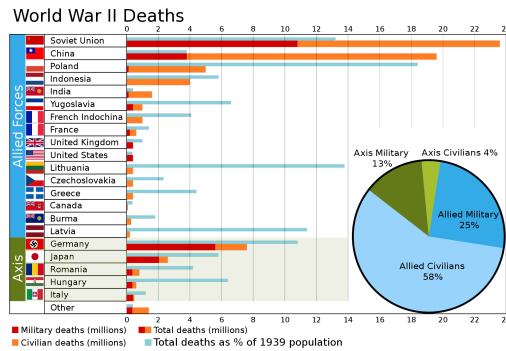
User

Yes, there is direct contact between the tree and the giraffe, as the giraffe is standing under the tree.

No, there is no direct contact between the tree and the giraffe. The giraffe is standing in the grassy field, and the tree is located in the background.

Figure S25: HALVA [🤖] does not exhibit a tendency to answer ‘Yes’ for any questions asked, like base model VILA-v1.5 [🤖]. Moreover, HALVA exhibit consistency in its response unlike VILA-v1.5, as shown in the left example, HALVA confirms the presence of two horses in both the time. Hallucinations are highlighted in red.

2484  
2485  
2486  
2487  
2488  
2489  
2490  
2491  
2492  
2493  
2494  
2495  
2496  
2497  
2498  
2499  
2500  
2501  
2502  
2503  
2504  
2505  
2506  
2507  
2508  
2509  
2510  
2511  
2512  
2513  
2514  
2515  
2516  
2517  
2518  
2519  
2520  
2521  
2522  
2523  
2524  
2525  
2526  
2527  
2528  
2529  
2530  
2531  
2532  
2533  
2534  
2535  
2536  
2537



User: According to the image, did the United States suffer the greatest number of total deaths in World War II, with China having the second-highest?

AI: Yes, according to the image, the United States suffered the greatest number of total deaths in World War II, with China having the second-highest.

AI: No, according to the image, the Soviet Union suffered the greatest number of total deaths in World War II, with China having the second-highest.

Figure S26: Examples illustrating the effectiveness of our proposed DPA in mitigating hallucinations that are beyond object centric. HALVA [17] accurately answers to this chart-based question unlike VILA-v1.5 [16]. Hallucinations are highlighted in red.