

# Data-augmented phrase-level alignment for mitigating object hallucination

Pritam Sarkar<sup>\*1,2</sup> Sayna Ebrahimi<sup>3</sup> Ali Etemad<sup>\*1</sup>

Ahmad Beirami<sup>4</sup> Sercan Ö. Arik<sup>3</sup> Tomas Pfister<sup>3</sup>

<sup>1</sup>Queen’s University <sup>2</sup>Vector Institute <sup>3</sup>Google Cloud AI Research <sup>4</sup>Google DeepMind

## 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 Hal-lucination 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) [1, 2, 3, 4, 5, 6, 7] have laid the foundation for the development of highly capable multimodal LLMs (MLLMs) [6, 8, 9, 10, 11, 12]. 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 [13, 14, 15, 16]. 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

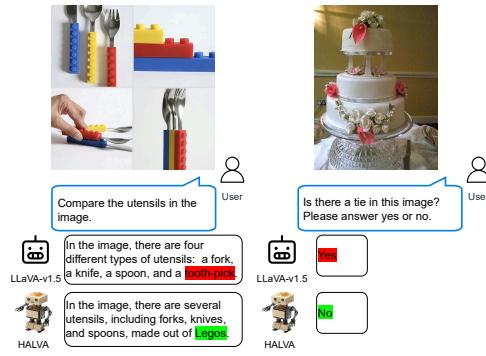


Figure 1: Examples of object hallucinations.

\*This work was partially done when Pritam Sarkar was an intern at Google Cloud AI Research and Ali Etemad was a visiting faculty researcher at Google Research. Corresponding author: pritam.sarkar@queensu.ca

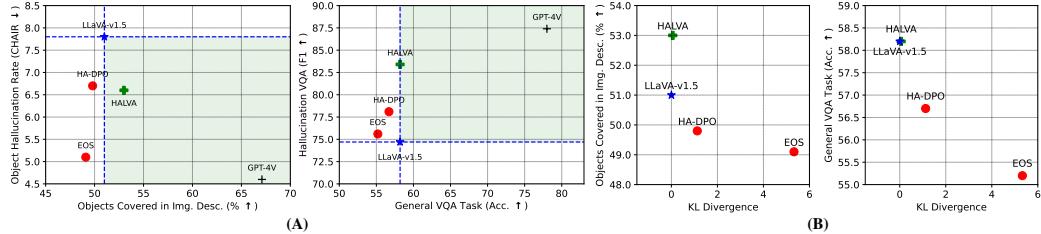


Figure 2: (A): A high-level overview comparing the performance of HALVA (the finetuned model with DPA) with existing finetuning methods in mitigating object hallucination, and their ability on general vision-language tasks. (B): Unlike HALVA, the existing finetuning approaches (e.g., HA-DPO and EOS) substantially diverge from their base model (LLaVA-v1.5<sub>13B</sub>).

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 [17, 16].

Prior work have attempted to address object hallucination in one of three key stages: inference [18, 19, 20, 21, 22, 23], pretraining [24, 25, 17], and finetuning [26, 27]. Inference-based methods aim to mitigate hallucinations during text generation, either through specialized decoding [20, 18, 28] or through iterative corrections [21, 29, 22], 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 [21, 16]. Pretraining techniques, such as negative instruction tuning or contrastive learning, have also been used to mitigate object hallucination [17, 25]. 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 [26] or human feedback [24, 30], among others [31, 27]. Nonetheless, these methods may lead to decreased performance on general vision-language tasks as illustrated in Figure 2 (A), which could be attributed to substantial divergence from their initial states (see Figure 2 (B)). We believe that the use of ‘sequence’-level losses between correct and hallucinated responses contributes to this phenomenon, as penalties are applied to all tokens in a hallucinated response, causing the model to diverge too far from its initial state.

Our goal is to achieve a method that can mitigate object hallucination in MLLMs without adding to inference time or requiring substantial re-training, while retaining the out-of-the-box performance on general vision-language tasks. To this end, we first use generative data augmentation [32, 33] 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 novel 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

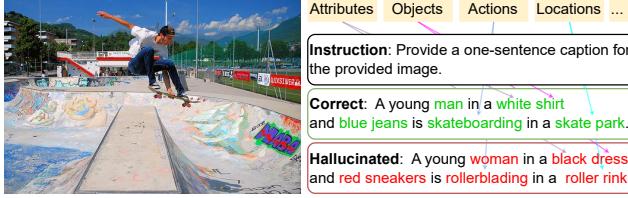


Figure 3: An example of correct and hallucinated response pairs constructed through our generative data-augmentation. The hallucinated responses are generated by selectively altering the true concepts in the correct response. For instance, we alter ‘objects’: shirt → dress, & jeans → sneakers; ‘attributes’: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

methods, 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<sup>1</sup>.

## 2 Method: Data-augmented phrase-level alignment

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 [22, 16, 15, 34]. 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. An example is presented 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.

<sup>1</sup>A GitHub link will be added here.

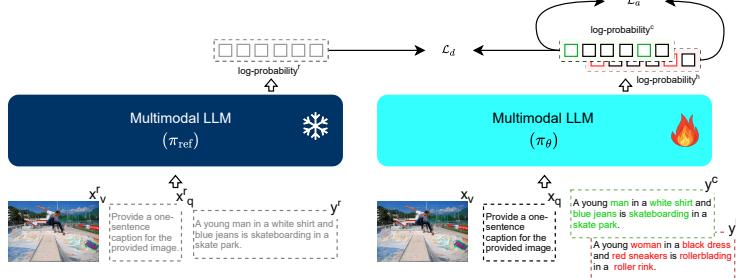


Figure 4: **Overview of our method:** Given a vision-language instruction and its correct and hallucinated response pair, the alignment objective ( $\mathcal{L}_a$ ) reduces the log-likelihood of hallucinated tokens compared to the correct ones. Also, a token-wise KL divergence regularizer ( $\mathcal{L}_d$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

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 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 [35, 36].

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 [35]. 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 defined as:

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

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) [37], which is an object-centric image dataset consisting of a total of 108K images and their annotations. Accordingly, we prepare the correct responses with both descriptive (e.g., *Describe the image in detail.*) and non-descriptive (e.g., <Question>, *Please answer in one word, yes or no*) instructions. Descriptive instructions include one-sentence captions, short descriptions, and detailed descriptions of images. Moreover, the non-descriptive question-answers are directly taken from [26]. We prepare the correct responses using Gemini Vision Pro [6] and based on the original images and ground-truth annotations. Subsequently, we perform generative data augmentation to obtain hallucinated responses, as described in Section 2. Our final training set consists of a total of 21.5K vision-language instructions and their corresponding correct and hallucinated responses.

**Implementation details.** We use LLaVA-v1.5 [9] and VILA-v1.5 [38] as our base models considering their superior performance in general vision-language tasks and the availability of their code and models. LLaVA-v1.5 uses Vicuna-v1.5 [39, 5] as the language encoder and CLIP ViT-L<sub>14</sub> [40] as the vision encoder. VILA-v1.5 uses Vicuna-v1.5 [39, 5] as the language encoder and SigLip-L-400M [41] as the vision encoder. Note that while LLaVA-v1.5 uses images of resolution 336 pixels, VILA-v1.5 is trained with images of resolution 384 pixels. During training, we freeze the vision encoder and projection layers, and only train the LLM using LoRA [42]. We refer to the resulting DPA trained checkpoints as HALVA, i.e., HALVA<sub>7B</sub> based on 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>. All experiments are conducted on 4 A100-80GB GPUs. We utilize an effective batch size of 64 and train for 1 epoch or 342 steps. The training time ranges from 1.5 to 3 hours for 7B and 13B variants. The additional implementation details are presented in Appendix D.

**Evaluation setup.** First, we evaluate HALVA on four object hallucination benchmarks encompassing both generative and discriminative tasks, including **CHAIR** [15], **MME-Hall** [43], **AMBER** [44], and **MMHal-Bench** [24]. Additionally, we perform a curiosity driven experiment to critically test the impact of our proposed DPA beyond object hallucination, using **HallusionBench** [45]. Furthermore, to ensure that DPA does not adversely affect the general language generation capabilities of MLLMs, we evaluate HALVA on four popular vision-language benchmarks: **VQA-v2** [46], **MM-Vet** [47], **TextVQA** [48], and **MME** [43]. All evaluations are conducted three times, and we report average scores. In the case of GPT-4-based evaluation, the performance slightly varies due to the randomness of GPT-4 generations. Therefore we also report the standard deviations.

### 4 Results

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

**Image description task.** In Figure 2 (A) Left, we compare MLLMs on image description tasks in terms of both hallucination rate (AMBER CHAIR) and their detailedness, captured through the number of ground-truth objects covered (AMBER Cover). Our goal is to mitigate hallucinations while retaining or improving the richness of image descriptions compared to the base model. As shown, HALVA captures more ground-truth objects while hallucinating less than HA-DPO. Moreover, while EOS achieves a slightly lower hallucination rate, it degrades the detailedness of image descriptions, performing worse than the base model. This is an undesired artifact in MLLMs, particularly for tasks that require detailedness such as medical imaging analysis [13, 14].

**Question answering task.** In Figure 2 (A) Right, we compare the performance of MLLMs on visual question-answering tasks using both object hallucination (AMBER) and general vision-language (TextVQA) benchmarks. As shown, both HA-DPO and EOS underperform HALVA in mitigating

object hallucination and even deteriorate general vision-language abilities compared to the base model. These results show the shortcomings of existing approaches, which we address in this work.

To further understand the limitations of existing methods in greater detail, we measure divergence 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.

#### 4.1 Evaluation on object hallucination

**CHAIR.** MLLMs can be prone to hallucinations when generating detailed image descriptions [16, 15, 44]. To assess the impact of DPA in such scenarios, we evaluate HALVA on CHAIR, which stands for Caption Hallucination Assessment with Image Relevance [15]. 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 [49], EOS [27], and HA-DPO [26]. It should be noted that our proposed DPA does not negatively impact the language generation ability or expressiveness of MLLMs, unlike EOS [27], 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 [15] 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 [44], a more recent object hallucination benchmark, which we discuss later.

**MME-Hall.** We evaluate HALVA on discriminative tasks using MME [43]. 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.

**AMBER.** To evaluate performance on both generative and discriminative tasks, we use AMBER [44], which measures hallucination using several metrics. For generative tasks, AMBER assesses the frequency of hallucinated objects in image descriptions, similar to [15]. 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 [44].

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 [19], 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

Table 1: Results on **CHAIR**.  $\ddagger$  and  $\dagger$  indicate that the reported values are from [50] and [27].  $*$ Results are computed by us, using their official checkpoints.  $C_i$  and  $C_s$  refer to CHAIR at instance and sentence levels. **Red:** worsen base model.

Method	$C_i(\downarrow)$	$C_s(\downarrow)$	Len.
mPLUG-Owl $^{\ddagger}$ 7B [51]	30.2	76.8	98.5
MultiModal-GPT $^{\ddagger}$ 7B [52]	18.2	36.2	45.7
MiniGPT-v2 $^{\ddagger}$ 7B [50]	8.7	25.3	56.5
InstructBlip $_{7B}$ [10]	17.5	62.9	102.9
LLaVA-v1.5 $^{\dagger}$ $_{7B}$ [9]	15.4	50.0	100.6
EOS $_{7B}$ [27]	12.3	40.2	79.7
OPERA $_{7B}$ [49]	12.8	44.6	-
DoLA $_{7B}$ [53]	13.8	47.8	-
HA-DPO $^*$ $_{7B}$ [26]	<b>11.0</b>	<b>38.2</b>	91.0
<b>HALVA<math>_{7B}</math> (Ours)</b>	<b>11.7</b> $\downarrow$ <b>3.7</b>	<b>41.4</b> $\downarrow$ <b>8.6</b>	92.2
MiniGPT-4 $^{\dagger}$ 13B [54]	9.2	31.5	116.2
InstructBlip $_{13B}$ [10]	16.0	51.2	95.6
LLaVA $^{\ddagger}$ $_{13B}$ [8]	18.8	62.7	90.7
LLaVA-v1.5 $^{\dagger}$ $_{13B}$ [9]	13.0	47.2	100.9
EOS $_{13B}$ [27]	11.4	36.8	85.1
<b>HALVA<math>_{13B}</math> (Ours)</b>	<b>12.8</b> $\downarrow$ <b>0.2</b>	<b>45.4</b> $\downarrow$ <b>1.8</b>	98.0
VILA-v1.5 $_{13B/384}$ [38]	9.2	33.0	183.4
<b>HALVA<math>_{13B/384}</math> (Ours)</b>	<b>8.4</b> $\downarrow$ <b>0.8</b>	<b>30.0</b> $\downarrow$ <b>3.0</b>	182.6

Table 2: Results on **MME-Hall**.  $\ddagger$  indicating reported values from [16].  $*$ Results are computed by us, using official checkpoints. **Red:** worsen base model.

Method	MME-Hall ( $\uparrow$ )
Cheetor $_{7B}$ $\ddagger$ [55]	473.4
LRV-Instruction $_{7B}$ $\ddagger$ [17]	528.4
Otter $_{7B}$ $\ddagger$ [56]	483.3
mPLUG-Owl $_{27B}$ $\ddagger$ [57]	578.3
Lynx $_{7B}$ [58]	606.7
Qwen-VL-Chat $_{7B}$ $\ddagger$ [59]	606.6
LLaMA-Adapter V2 $_{7B}$ $\ddagger$ [60]	493.3
LLaVA-v1.5 $_{7B}$ [9]	648.3
HA-DPO $^*$ $_{7B}$ [26]	<b>618.3</b>
EOS $^*$ $_{7B}$ [27]	<b>606.7</b>
VCD $_{7B}$ [20]	<b>604.7</b>
Woodpecker $^*$ $_{7B}$ [19]	<b>366.7</b>
<b>HALVA<math>_{7B}</math> (Ours)</b>	<b>665.0</b> $\uparrow$ <b>16.7</b>
BLIVA $_{11B}$ $\ddagger$ [61]	580.0
MMICL $_{12B}$ $\ddagger$ [62]	568.4
InstructBLIP $_{13B}$ $\ddagger$ [10]	548.3
SPHINX $_{13B}$ $\ddagger$ [63]	668.3
Muffin $_{13B}$ $\ddagger$ [64]	590.0
RLHF-V $_{13B}$ [30]	585.0
LLaVA-v1.5 $_{13B}$ [9]	643.3
<b>HALVA<math>_{13B}</math> (Ours)</b>	<b>675.0</b> $\uparrow$ <b>31.7</b>
VILA-v1.5 $_{13B/384}$ [38]	688.3
<b>HALVA<math>_{13B/384}</math> (Ours)</b>	<b>691.7</b> $\uparrow$ <b>3.4</b>

enhances performance on discriminative tasks, for both 7B and 13B variants. For instance, HALVA $_{7B}$  improves the F1-score on the existence category from 64.6% to 93.3%. Additionally, HALVA $_{13B}$  improves the F1 score on relation-based tasks from 45.0% to 73.5%. Overall, HALVA $_{7B}$  outperforms both HA-DPO and EOS on discriminative tasks by a large margin, achieving a 5.3 and 7.8 point higher F1 score respectively. Furthermore, HALVA $_{13B}$  and HALVA $_{13B/384}$  perform better or on par with GPT-4V on discriminative tasks, i.e., F1-score of 86.5 by HALVA $_{13B}$ , 87.9 by HALVA $_{13B/384}$ , 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 [15]. We utilize MMHal-Bench [24], which evaluates hallucination across 12 object-topics, including object attributes, presence of adversarial objects, and spatial relations, among others. Following [24], we use GPT-4 [12] 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 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 $_{7B}$  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 $_{13B}$  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

Table 3: Results on **AMBER**. Cover.: coverage of ground-truth objects; Hall.: Hallucination Rate; Cog.: Cognition;  $F1_E$ ,  $F1_A$ , and  $F1_R$  refer to F1 scores of Existence, Attribute, and Relation subsets. The final F1 is calculated across all sub-tasks.  $\dagger$  indicates that the reported values are from [44]. \*Results are computed by us, using their official checkpoint. **Red**: worsens base model.

Method	Generative Task				Discriminative Task			
	CHAIR ( $\downarrow$ )	Cover. ( $\uparrow$ )	Hall. ( $\downarrow$ )	Cog. ( $\downarrow$ )	$F1_E(\uparrow)$	$F1_A(\uparrow)$	$F1_R(\uparrow)$	$F1(\uparrow)$
mPLUG-Owl $\dagger$ $_{7B}$ [51]	21.6	50.1	76.1	11.5	17.2	22.9	6.2	18.9
LLaVA $\dagger$ $_{7B}$ [8]	11.5	51.0	48.8	5.5	8.4	48.6	58.1	32.7
MiniGPT-4 $\dagger$ $_{7B}$ [54]	13.6	63.0	65.3	11.3	80.0	43.7	52.7	64.7
mPLUG-Owl2 $\dagger$ $_{7B}$ [57]	10.6	52.0	39.9	4.5	89.1	72.4	54.3	78.5
InstructBLIP $\dagger$ $_{7B}$	8.8	52.2	38.2	4.4	89.0	76.3	67.6	81.7
LLaVA-v1.5 $\dagger$ $_{7B}$	7.8	51.0	36.4	4.2	64.6	65.6	62.4	74.7
HA-DPO $^*$ $_{7B}$ [26]	6.7	49.8	30.9	3.3	88.1	66.1	68.8	78.1
EOS $^*$ $_{7B}$ [27]	5.1	49.1	22.7	2.0	82.8	67.4	69.2	75.6
Woodpecker $^*$ $_{7B}$ [19]	6.9	48.9	30.4	3.6	81.7	53.5	41.5	67.0
<b>HALVA<math>_{7B}</math> (Ours)</b>	<b>6.6<math>\downarrow</math>1.2</b>	<b>53.0<math>\uparrow</math>2.0</b>	<b>32.2<math>\downarrow</math>4.2</b>	<b>3.4<math>\downarrow</math>0.8</b>	<b>93.3<math>\uparrow</math>28.7</b>	<b>77.1<math>\uparrow</math>11.5</b>	<b>63.1<math>\uparrow</math>0.7</b>	<b>83.4<math>\uparrow</math>8.7</b>
RLHF-V $_{13B/448}$ [30]	6.8	46.1	27.4	2.5	95.7	80.0	71.8	87.1
LLaVA-v1.5 $_{13B}$ [9]	6.6	51.9	30.5	3.3	78.5	70.2	45.0	73.1
<b>HALVA<math>_{13B}</math> (Ours)</b>	<b>6.4<math>\downarrow</math>0.2</b>	<b>52.6<math>\uparrow</math>0.7</b>	<b>30.4<math>\downarrow</math>0.1</b>	<b>3.2<math>\downarrow</math>0.1</b>	<b>92.6<math>\uparrow</math>14.1</b>	<b>81.4<math>\uparrow</math>11.2</b>	<b>73.5<math>\uparrow</math>28.5</b>	<b>86.5<math>\uparrow</math>13.4</b>
VILA-v1.5 $_{13B/384}$ [38]	9.9	63.3	56.1	4.8	87.5	77.8	66.7	82.2
<b>HALVA<math>_{13B/384}</math> (Ours)</b>	<b>9.1<math>\downarrow</math>0.8</b>	<b>63.9<math>\uparrow</math>0.6</b>	<b>54.2<math>\downarrow</math>1.9</b>	<b>4.0<math>\downarrow</math>0.8</b>	<b>93.9<math>\uparrow</math>16.4</b>	<b>82.6<math>\uparrow</math>4.8</b>	<b>75.9<math>\uparrow</math>9.2</b>	<b>87.9<math>\uparrow</math>5.7</b>
GPT-4V $\dagger$ [12]	4.6	67.1	30.7	2.6	94.5	82.2	83.2	87.4

[45]. The results presented in Table 5 demonstrate that our proposed method directly benefits other forms of vision-language hallucinations as well. HALVA $_{7B}$ , HALVA $_{13B}$ , and HALVA $_{13B/384}$  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 $_{13B/384}$  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 [46], MM-Vet [47], TextVQA [48], and MME [43]. We follow the evaluation protocol mentioned in LLaVA-v1.5 [9]. The results presented in Table 6 show that HALVA maintains or improves performance with respect to the base models. For example, HALVA $_{7B}$  improves on MME and MM-Vet by 16.3 and 1% respectively, while retaining the same performance on TextVQA and VQA-v2. A similar trend is noticed in the case of HALVA $_{13B}$ . Unlike HALVA $_{7B}$ , existing finetuning methods such as HA-DPO $_{7B}$  and EOS $_{7B}$ , based on LLaVA-v1.5 $_{7B}$ , exhibit deterioration in general tasks when tuned for hallucination mitigation.

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 $\uparrow$	MM-Vet $\uparrow$	TextVQA $\uparrow$	MME $\uparrow$
LLaVA-v1.5 $_{7B}$	<b>78.5</b>	31.1	<b>58.2</b>	1510.7
HA-DPO $_{7B}$	77.6 $^{*}$ $\downarrow$ 0.9	30.7 $^{*}$ $\uparrow$ 0.4	56.7 $^{*}$ $\downarrow$ 1.5	1502.6 $\downarrow$ 8.1
EOS $_{7B}$	77.6 $^{*}$ $\downarrow$ 0.9	31.4 $^{*}$ $\uparrow$ 0.3	55.2 $^{*}$ $\downarrow$ 3.0	1424.4 $^{*}$ $\downarrow$ 102.6
<b>HALVA<math>_{7B}</math></b>	<b>78.5</b> $\downarrow$ 0.0	<b>32.1<math>\uparrow</math>1.0</b>	<b>58.2</b> $\downarrow$ 0.0	<b>1527.0<math>\uparrow</math>16.3</b>
LLaVA-v1.5 $_{13B}$	<b>80.0</b>	36.1	<b>61.2</b>	1530.1
<b>HALVA<math>_{13B}</math></b>	<b>80.0</b> $\downarrow$ 0.0	<b>37.8<math>\uparrow</math>1.0</b>	<b>61.2</b> $\downarrow$ 0.0	<b>1544.0<math>\uparrow</math>13.9</b>

### 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

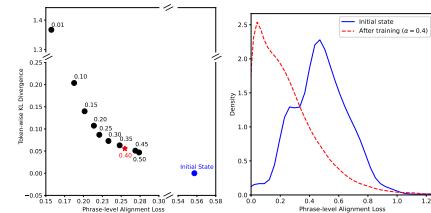


Figure 5: **Left:** Changes in the model state due to DPA training with varying  $\alpha$ . **Right:** Changes in alignment loss before and after training across all training samples. Default  $\alpha$  is 0.4 for HALVA $_{7B}$ .

Table 4: Results on **MMHal-Bench**.  $\dagger$  and  $\ddagger$  indicate that the reported values are from [24] and [25].  $*$ Results are computed by us, using their official checkpoint. **Red:** worsen base model.

Method	Overall Score ( $\uparrow$ )	Hall. Rate ( $\downarrow$ )
Kosmos-2 $\ddagger$ [65]	1.69	0.68
IDEFIC $\ddagger$ 9B [66]	1.89	0.64
InstructBLIP $\ddagger$ 7B [10]	2.10	0.58
LLaVA $\ddagger$ 7B [8]	1.55	0.76
VCD $_7$ B [20]	2.12	0.54
OPERA $_7$ B [49]	2.33	0.50
LURE $_7$ B [22]	1.64	0.60
LLaVA-SFT $_7$ B [24]	1.76	0.67
LLaVA-RLHF $_7$ B [24]	2.05	0.68
LLaVA-v1.5 $_7$ B [9]	$2.11 \pm 0.05$	$0.54 \pm 0.01$
HAACL $_7$ B [25]	2.13	<b>0.50</b>
HA-DPO $^*$ 7B [26]	<b>1.97</b>	<b>0.60</b>
EOS $^*$ 7B [27]	<b>2.03</b>	<b>0.59</b>
<b>HALVA<math>_7</math>B (Ours)</b>	<b>2.25</b> $\pm 0.09$ <b>↑0.14</b>	$0.54 \pm 0.01$ <b>↓0.00</b>
LLaVA $^\dagger$ 13B [8]	1.11	0.84
InstructBLIP $^\dagger$ 13B [10]	2.14	0.58
RLHF-V $_{13B/448}$ [30]	-	0.52
LLaVA-SFT $_{13B}$ [24]	2.43	0.55
LLaVA-RLHF $_{13B}$ [24]	2.53	0.57
LLaVA-v1.5 $_{13B}$ [9]	$2.37 \pm 0.02$	$0.50 \pm 0.00$
<b>HALVA<math>_{13B}</math> (Ours)</b>	<b>2.58</b> $\pm 0.07$ <b>↑0.21</b>	<b>0.45</b> $\pm 0.02$ <b>↓0.05</b>
VILA-v1.5 $_{13B/384}$ [38]	<b>2.58</b> $\pm 0.02$	$0.46 \pm 0.01$
<b>HALVA<math>_{13B/384}</math> (Ours)</b>	<b>2.58</b> $\pm 0.06$	<b>0.45</b> $\pm 0.01$ <b>↓0.01</b>
GPT4V [12]	3.49	0.28

Table 5: Results on **HallusionBench**.  $\dagger$  indicates that the reported values are from [45].  $*$ Results are computed by us, using their official checkpoint.

Method	Yes/No Bias (~0)	Overall Acc. ( $\uparrow$ )
mPLUG_Owl-v1 $^\dagger$ 7.2B [51]	0.32	43.93
MiniGPT5 $^\dagger$ 7B [67]	0.25	40.30
MiniGPT4 $^\dagger$ 7B [54]	0.19	35.78
InstructBLIP $^\dagger$ 7B [10]	-0.13	45.26
BLIP2 $^\dagger$ 7B [11]	0.18	40.48
mPLUG_Owl-v2 $^\dagger$ 7B [57]	0.25	47.30
LRV-Instruction $^\dagger$ 7B [17]	0.26	42.78
LLaVA-1.5 $^*$ 7B [9]	0.31	$47.09 \pm 0.14$
LLaVA-RLHF $^*$ 7B [24]	0.24	43.0
HA-DPO $^*$ 7B [26]	0.26	48.4
EOS $^*$ 7B [27]	0.29	48.7
<b>HALVA<math>_7</math>B (Ours)</b>	<b>0.17</b>	<b>48.95</b> $\pm 0.13$ <b>↑1.86</b>
Qwen-VL $^{9.6B}$ [59]	0.12	39.15
Open-Flamingo $^\dagger$ 9B [68]	0.33	38.44
BLIP2-T5 $^\dagger$ 12B [11]	0.08	48.09
RLHF-V $^{13B/448}$ [30]	0.13	47.47
LLaVA-1.5 $^\dagger$ 13B [9]	0.26	46.94
LLaVA-1.5 $^*$ 13B [9]	0.38	$46.50 \pm 0.09$
LLaVA-RLHF $^*$ 13B [24]	<b>0.17</b>	46.41
<b>HALVA<math>_{13B}</math> (Ours)</b>	0.20	<b>49.10</b> $\pm 0.05$ <b>↑2.16</b>
VILA-v1.5 $^*$ 13B/384 [38]	0.19	$55.39 \pm 0.05$
<b>HALVA<math>_{13B/384}</math> (Ours)</b>	<b>0.02</b>	<b>56.60</b> $\pm 0.18$ <b>↑1.21</b>
GPT4V $^\dagger$ [12]	0.06	65.28
Gemini Pro Vision $^\dagger$ [6]	-0.02	36.85

Figure 5. 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 $_7$ B. 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.

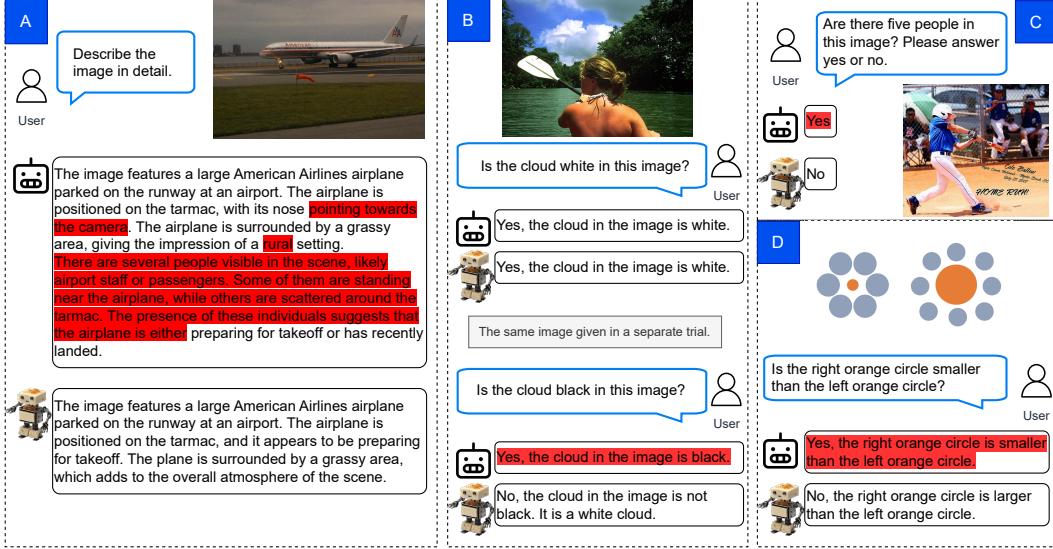


Figure 6: Qualitative comparisons between HALVA [2] and LLaVA-v1.5 [6]. Our proposed DPA effectively mitigates hallucination under different setups: (A) detail image description, (B) visual question-answering, (C) Yes-or-No answer, (D) visual illusion. Hallucinations are highlighted in red. More examples, comparing with LLaVA-v1.5 and VILA-v1.5, are in Appendix E.

## 5 Related work

**Multimodal LLMs.** Vision-language models (VLMs) often align image and text features in a shared embedding space, as pioneered by CLIP [40] and ALIGN [69], followed by others [70, 71, 72, 73]. 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 [8, 54, 6, 12, 10, 11, 65, 61, 10, 59, 74] further enhance visual perception, understanding, and reasoning. While some MLLMs are open-source, others are only accessible through APIs [12, 6, 59]. Among the publicly available MLLMs, LLaVA [8, 9] and VILA [38] 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 the context of object hallucination [15, 16, 22, 24, 23]. 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 [17] or generating and correcting image-instruction pairs on-the-fly [75]; inference, with methods involving specialized decoding strategies [20, 18, 28] or iterative corrections using offline models to detect and correct hallucinations at inference time [22, 19]; and finetuning, with approaches relying on human feedback [24, 30] to train reward models or employing preference optimization techniques [26, 30, 76, 77, 78, 79]. 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 deteriorate the performance of the base model on general vision-language tasks (see 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 vision-language ability of the base model.

## 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

- [1] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *JMLR*, 24(240):1–113, 2023. [1](#)
- [2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. [1](#)
- [3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 21(140):1–67, 2020. [1](#)
- [4] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. [1](#)
- [5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [1, 5](#)
- [6] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. [1, 5, 9, 10](#)
- [7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. [1](#)
- [8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. [1, 7, 8, 9, 10, 29](#)
- [9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. [1, 5, 7, 8, 9, 10, 22, 25](#)
- [10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. [1, 7, 9, 10](#)
- [11] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [1, 9, 10](#)
- [12] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. [1, 7, 8, 9, 10](#)
- [13] Sheng Wang, Zihao Zhao, Xi Ouyang, Qian Wang, and Dinggang Shen. Chatcad: Interactive computer-aided diagnosis on medical image using large language models. *arXiv preprint arXiv:2302.07257*, 2023. [1, 5](#)

- [14] Mingzhe Hu, Shaoyan Pan, Yuheng Li, and Xiaofeng Yang. Advancing medical imaging with language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*, 2023. 1, 5
- [15] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018. 1, 3, 5, 6, 7, 10
- [16] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1, 2, 3, 6, 7, 10, 23
- [17] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*, 2023. 2, 7, 9, 10
- [18] Ailin Deng, Zhirui Chen, and Bryan Hooi. Seeing is believing: Mitigating hallucination in large vision-language models via clip-guided decoding. *arXiv preprint arXiv:2402.15300*, 2024. 2, 10
- [19] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. 2, 6, 7, 8, 10, 22
- [20] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. 2, 7, 9, 10, 22
- [21] Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv preprint arXiv:2311.07362*, 2023. 2
- [22] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023. 2, 3, 9, 10
- [23] Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *WACV*, pages 1381–1390, 2022. 2, 10
- [24] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*, 2023. 2, 5, 7, 9, 10, 22
- [25] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. *arXiv preprint arXiv:2312.06968*, 2023. 2, 9
- [26] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lylms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 2, 5, 6, 7, 8, 9, 10, 17, 22, 23
- [27] Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*, 2024. 2, 6, 7, 8, 9, 22
- [28] Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. *arXiv preprint arXiv:2402.18476*, 2024. 2, 10
- [29] Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. *arXiv preprint arXiv:2402.11622*, 2024. 2
- [30] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023. 2, 7, 8, 9, 10, 17
- [31] Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mocha: Multi-objective reinforcement mitigating caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2023. 2
- [32] Yao Qin, Chiyuan Zhang, Ting Chen, Balaji Lakshminarayanan, Alex Beutel, and Xuezhi Wang. Understanding and improving robustness of vision transformers through patch-based negative augmentation. *NeurIPS*, 35:16276–16289, 2022. 2

- [33] Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. *NeurIPS*, 36, 2024. 2
- [34] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 3, 23
- [35] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017. 4, 17
- [36] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 36, 2024. 4, 17
- [37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5, 25
- [38] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *CVPR*, pages 26689–26699, 2024. 5, 7, 8, 9, 10, 25
- [39] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality, March 2023. 5
- [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 5, 10
- [41] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *ICCV*, pages 11975–11986, 2023. 5
- [42] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5, 25
- [43] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 5, 6, 8, 22
- [44] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. 5, 6, 8, 23
- [45] Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*, 2023. 5, 8, 9, 22
- [46] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, pages 6904–6913, 2017. 5, 8
- [47] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 5, 8
- [48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 5, 8
- [49] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:2311.17911*, 2023. 6, 7, 9
- [50] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large

- language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 7
- [51] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 7, 8, 9
- [52] Tao Gong, Chengqi Lyu, Shilong Zhang, Yudong Wang, Miao Zheng, Qian Zhao, Kuikun Liu, Wenwei Zhang, Ping Luo, and Kai Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023. 7
- [53] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023. 7
- [54] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 7, 8, 9, 10
- [55] Juncheng Li, Kaihang Pan, Zhiqi Ge, Minghe Gao, Hanwang Zhang, Wei Ji, Wenqiao Zhang, Tat-Seng Chua, Siliang Tang, and Yueting Zhuang. Empowering vision-language models to follow interleaved vision-language instructions. *arXiv preprint arXiv:2308.04152*, 2023. 7
- [56] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*, 2023. 7
- [57] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv preprint arXiv:2311.04257*, 2023. 7, 8, 9
- [58] Yan Zeng, Hanbo Zhang, Jiani Zheng, Jiangnan Xia, Guoqiang Wei, Yang Wei, Yuchen Zhang, and Tao Kong. What matters in training a gpt4-style language model with multimodal inputs? *arXiv preprint arXiv:2307.02469*, 2023. 7
- [59] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 7, 9, 10
- [60] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 7
- [61] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu. Bliva: A simple multimodal llm for better handling of text-rich visual questions. In *AAAI*, volume 38, pages 2256–2264, 2024. 7, 10
- [62] Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*, 2023. 7
- [63] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, et al. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv preprint arXiv:2311.07575*, 2023. 7
- [64] Renze Lou, Kai Zhang, Jian Xie, Yuxuan Sun, Janice Ahn, Hanzi Xu, Yu Su, and Wenpeng Yin. Muffin: Curating multi-faceted instructions for improving instruction following. In *ICLR*, 2023. 7
- [65] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 9, 10
- [66] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *NeurIPS*, 36, 2024. 9
- [67] Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*, 2023. 9

- [68] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 9
- [69] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 10
- [70] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022. 10
- [71] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 10
- [72] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, pages 12888–12900. PMLR, 2022. 10
- [73] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022. 10
- [74] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm’s referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 10
- [75] Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, et al. Vigc: Visual instruction generation and correction. In *AAAI*, volume 38, pages 5309–5317, 2024. 10
- [76] Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*, 2024. 10, 17
- [77] Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. Strengthening multimodal large language model with bootstrapped preference optimization. *arXiv preprint arXiv:2403.08730*, 2024. 10
- [78] Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024. 10
- [79] Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. *arXiv preprint arXiv:2405.19716*, 2024. 10
- [80] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 17
- [81] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024. 24
- [82] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 25
- [83] Jie Ren, Samyam Rajbhandari, Reza Yazdani Aminabadi, Olatunji Ruwase, Shuangyan Yang, Minjia Zhang, Dong Li, and Yuxiong He. Zero-offload: Democratizing billion-scale model training. In *USENIX ATC*, pages 551–564, 2021. 25
- [84] Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *SC*, pages 1–14, 2021. 25

# 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 eesults
- 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.

```
import torch
import torch.nn.functional as F

def forward(self, **inputs):
    """x: vision-language input
    y_pos: correct response of x
    y_neg: hallucinated response of x constructed through gen. data aug.
    x_ref, y_ref: reference input-output pair to calculate divergence
    """

    batch_size = x.shape[0]

    # forward pass with correct and hallucinated responses
    pos_logits = self.model(x, y_pos)
    neg_logits = self.model(x, y_neg)

    # calculate log-probabilities
    pos_logps, pos_labels = self.log_softmax(pos_logits, y_pos)
    neg_logps, neg_labels = self.log_softmax(neg_logits, y_neg)

    # accumulate log-probabilities of
    # correct and hallucinated tokens at phrase level
    pos_logps = self.accumulate_logps(pos_logps)
    neg_logps = self.accumulate_logps(neg_logps)

    # phrase-level alignment loss
    alignment_loss = torch.log(1 + torch.exp(neg_logps - pos_logps))
    alignment_loss = alignment_loss.mean()

    # forward pass with the reference samples
    logits = self.model(x_ref, y_ref)
    with torch.no_grad():
        reference_logits = self.reference_model(x_ref, y_ref)

    # calculate probability
    proba = F.softmax(logits, dim=-1)
    reference_proba = F.softmax(reference_logits, dim=-1)

    # token-wise KL divergence
    divergence = (reference_proba*(reference_proba.log()-proba.log()))
    divergence = divergence.sum()/batch_size

    # final loss
    loss = alignment_loss + self.alpha*divergence

    return loss
```

## B Distinction between ours DPA and DPO-based hallucination mitigation methods

Several existing and concurrent works, such as HA-DPO [26], RLHF-V [30], and RLAIF [76], have introduced hallucination mitigation techniques for MLLMs, that are derived from DPO [36]. Following, we discuss the differences between our proposed DPA and DPO.

We write both DPA (ours) and the DPO [36] 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:

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

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 [36] and other existing alignment techniques [35, 80]. 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 [26] 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.



**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>.

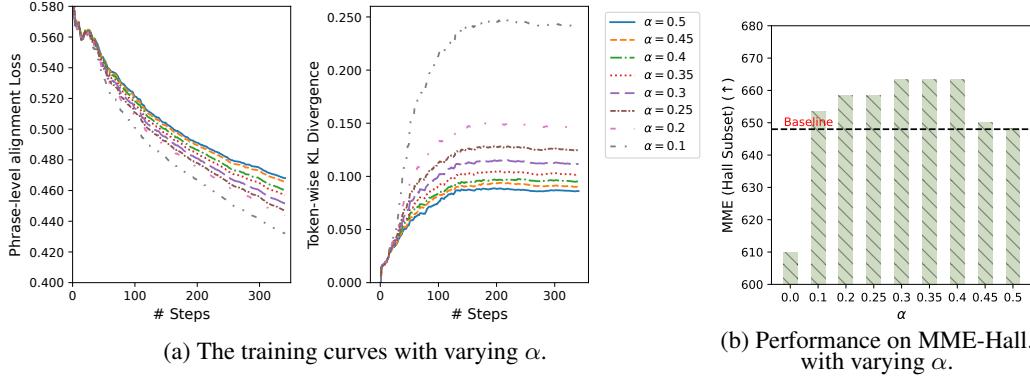


Figure S2: The training curves with varying  $\alpha$  (a) and their performance on object hallucination (b) are presented.  $\alpha$  in the range of 0.3 to 0.4 achieves optimal performance on the 7B variant.

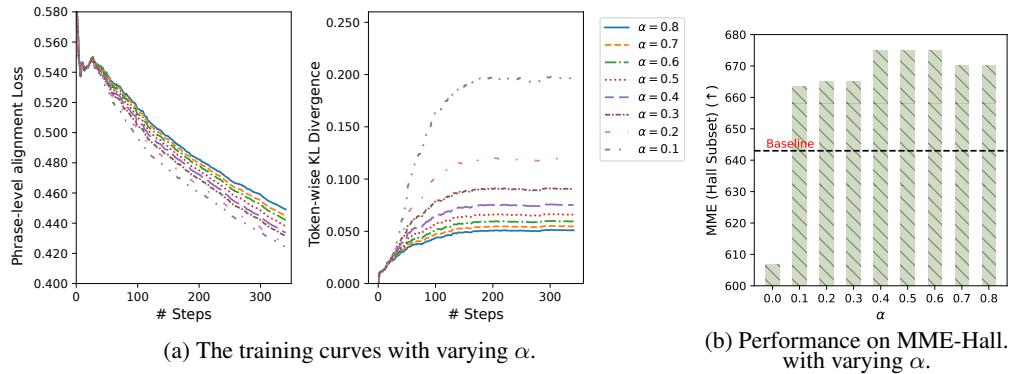


Figure S3: The training curves with varying  $\alpha$  (a) and their performance on object hallucination (b) are presented.  $\alpha$  in the range of 0.4 to 0.6 achieves optimal performance on the 13B variant.



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

$\alpha = 0.4$  (default): A painting of a house with a white porch and pink flowers.

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

$\alpha = 0.01$ : The image is a black and white photo of a man and a little girl sitting together on a couch. The man is holding the little girl, and they are both smiling.

$\alpha = 0.4$  (default): The image is a black and white photo of a man and a young girl sitting together on a couch. The man is wearing a tie, and the little girl is wearing a dress. They are both smiling and appear to be enjoying their time together.

Figure S4: Qualitative examples demonstrating the impact of DPA training with a very low  $\alpha$ . As presented here, training with a very low  $\alpha$  of 0.01 may occasionally hurt the language generation ability of an MLLM. The example on the left side shows an instance of degeneration, while the example on the right side shows a lack of descriptive power, failing to mention key details in the description, such as ‘the man is wearing a tie’ or ‘the girl is wearing a dress’. The 7B variant is used in this study.

## 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 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.

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
	$C_i \downarrow$	$C_s \downarrow$	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>	641.0
<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

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↑	HR↓	Score↑
Unseen data	12.7	47.4	81.7	34.7	<b>668.3</b>
<b>Seen data</b>	<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↑	HR↓	Score↑
7B	<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 [9]. 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 [43] 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 [26] and EOS [27], or inference-based methods such as VCD [20] and Woodpecker [19], 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>	<b>190.0</b>	155.0	133.3	170.0	648.3
HA-DPO <sub>7B</sub>	<b>190.0</b>	133.3	<b>136.7</b>	158.3	618.3
EOS <sub>7B</sub>	<b>190.0</b>	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>	<b>190.0</b>	<b>165.0</b>	135.0	<b>175.0</b>	<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>	<b>190.0</b>	<b>163.3</b>	<b>141.7</b>	<b>180.0</b>	<b>675.0</b>
VILA-v1.5 <sub>13B/384</sub>	<b>185.0</b>	170.0	<b>148.3</b>	<b>185.0</b>	688.3
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>185.0</b>	<b>173.3</b>	<b>148.3</b>	<b>185.0</b>	<b>691.7</b>

#### C.5 Detailed results of MMHal-Bench

In Table S5, we present the detailed results of MMHal-Bench [24] 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	<b>1.97</b> $\pm$ 0.09	2.36 $\pm$ 0.05	3.20 $\pm$ 0.05	2.14 $\pm$ 0.30	<b>1.53</b> $\pm$ 0.25
HA-DPO <sub>7B</sub>	1.97 $\pm$ 0.04	0.59 $\pm$ 0.01	<b>3.56</b> $\pm$ 0.17	1.08 $\pm$ 0.09	1.14 $\pm$ 0.13	1.89 $\pm$ 0.21	2.22 $\pm$ 0.33	<b>3.31</b> $\pm$ 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	<b>1.78</b> $\pm$ 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</b> $\pm$ 0.10	<b>0.54</b> $\pm$ 0.01	2.78 $\pm$ 0.09	1.47 $\pm$ 0.18	<b>1.97</b> $\pm$ 0.13	1.89 $\pm$ 0.05	<b>3.03</b> $\pm$ 0.21	3.20 $\pm$ 0.05	<b>2.42</b> $\pm$ 0.43	1.22 $\pm$ 0.27
LLaVA-v1.5 <sub>13B</sub>	2.38 $\pm$ 0.02	0.50 $\pm$ 0.01	<b>3.20</b> $\pm$ 0.05	2.53 $\pm$ 0.18	2.55 $\pm$ 0.05	<b>2.20</b> $\pm$ 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</b> $\pm$ 0.08	<b>0.46</b> $\pm$ 0.02	3.03 $\pm$ 0.09	<b>2.58</b> $\pm$ 0.09	<b>2.66</b> $\pm$ 0.14	2.08 $\pm$ 0.14	<b>2.45</b> $\pm$ 0.05	<b>3.36</b> $\pm$ 0.17	<b>2.44</b> $\pm$ 0.39	<b>2.00</b> $\pm$ 0.08
VILA-v1.5 <sub>13B/384</sub>	<b>2.58</b> $\pm$ 0.02	0.46 $\pm$ 0.01	<b>3.36</b> $\pm$ 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	<b>2.19</b> $\pm$ 0.13	2.47 $\pm$ 0.05
<b>HALVA<sub>13B/384</sub> (Ours)</b>	<b>2.58</b> $\pm$ 0.06	<b>0.45</b> $\pm$ 0.01	3.11 $\pm$ 0.05	<b>1.47</b> $\pm$ 0.05	<b>3.47</b> $\pm$ 0.05	<b>2.08</b> $\pm$ 0.00	<b>3.11</b> $\pm$ 0.13	<b>3.19</b> $\pm$ 0.13	1.64 $\pm$ 0.24	<b>2.58</b> $\pm$ 0.09

#### C.6 Detailed results of HallusionBench

In Table S6, we present the detailed results of HallusionBench [45], 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 (~ 0)	FP Ratio (~ 0.5)	(qAcc) $\uparrow$	(fAcc) $\uparrow$	(Easy aAcc) $\uparrow$	(Hard aAcc) $\uparrow$	(aAcc) $\uparrow$
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	<b>43.96</b>	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>	42.71 $\pm$ 0.13	<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 [26]. 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 [26] 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 <sub>i</sub> ) $\downarrow$	AMBER F1 $\uparrow$	MME-Hall $\uparrow$	MME $\uparrow$
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
<b>HALVA<sub>7B</sub></b>	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 [34]. 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 [44, 16].

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

Table S8: The results on POPE are presented. \* Results are obtained using a different  $\alpha$  than our default.  $\dagger$  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.

<b>Method</b>	<b>POPE</b> (F1 $\uparrow$ )	<b>AMBER</b> (F1 $\uparrow$ )	<b>HallusionBench</b> (Acc. $\uparrow$ )	<b>MME-Hall</b> (Score $\uparrow$ )	<b>MME</b> (Score $\uparrow$ )
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>	84.8/87.1*	<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>	84.9/87.9*	<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>	86.1	<b>87.9</b>	<b>56.6</b>	<b>691.7</b>	<b>1575.7</b>
GPT-4o $^\dagger$ (v.0513, detail-high)	85.6	-	55.0	-	2310.3
InternVL2 $^\dagger$ <sub>40B</sub> [81]	81.9	-	56.5	-	2293.1

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  $\alpha = 0.005$ .

## D Implementation details

### D.1 Training hyperparameters

The details of training hyperparameters used in DPA training is presented in Table S9.

Table S9: Details of training hyperparameters used in DPA training.

	<b>HALVA<sub>7B</sub></b>	<b>HALVA<sub>13B</sub></b>	<b>HALVA<sub>13B/384</sub></b>
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 [42]		rank=128, alpha=256	
Learning rate	5e-6		2.5e-5
Learning rate scheduler		Cosine	
Optimizer		AdamW [82]	
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 [83, 84]	
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 [37]. 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 [9] and VILA-v1.5 [38]. 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 S5, S6, and S7. The full list of instructions used in generating image descriptions is presented in Figure S8. 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 S9, S10, and S11. 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 S12, S13, S14, and S15, respectively.

```
# 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 S5: The **template** for generating the **correct** image descriptions.

```
# 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 S6: The **template** for generating the **closed-set** hallucinated descriptions.

**# 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 S7: The **template** for generating the **open-set** hallucinated descriptions.

**# 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 S8: **Instructions** for different types of image descriptions. These instructions are directly taken from [8], we list them here for the sake of completeness.

## # 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 S9: A **complete example** of generating the **correct** image descriptions.

# 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 S10: 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.

# 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 S11: 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.



**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 S12: Examples of **one sentence image captions** used in DPA training.



**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.

**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 S13: Examples of **short image descriptions** used in DPA training.



**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 S14: Examples of **detailed image descriptions** used in DPA training.



**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 S14 (Continued): Examples of **detailed image descriptions** used in DPA training.



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

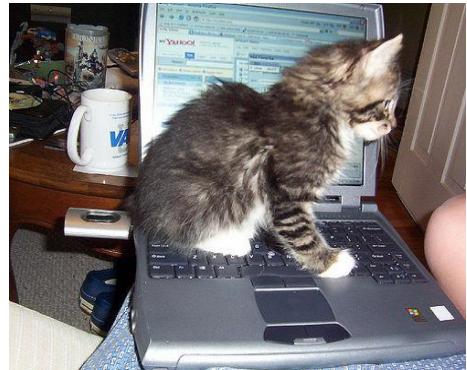
**Correct:** No

**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 S15: Examples of **Yes-or-No** questions used in DPA training.

## E Qualitative results

### E.1 Qualitative comparison between HALVA and LLaVA-v1.5

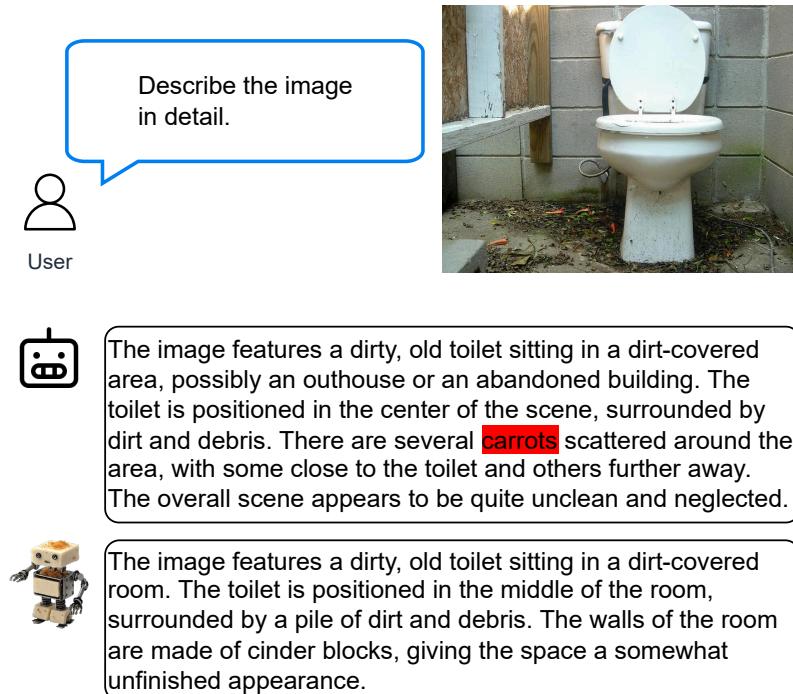


Figure S16: Comparing detailed image descriptions between HALVA [?] and its base model LLaVA-v1.5 [?]. Unlike HALVA, LLaVA-v1.5 hallucinates ‘carrot’ in the image description. Hallucinations are highlighted in red.



User

Describe the image in detail.



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 S17: Comparing detailed image descriptions between HALVA [HALVA] and its base model LLaVA-v1.5 [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.

User

Describe the image in detail.



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

HALVA

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 S18: Comparing detailed image descriptions between HALVA [HALVA] and its base model LLaVA-v1.5 [LLaVA]. 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.

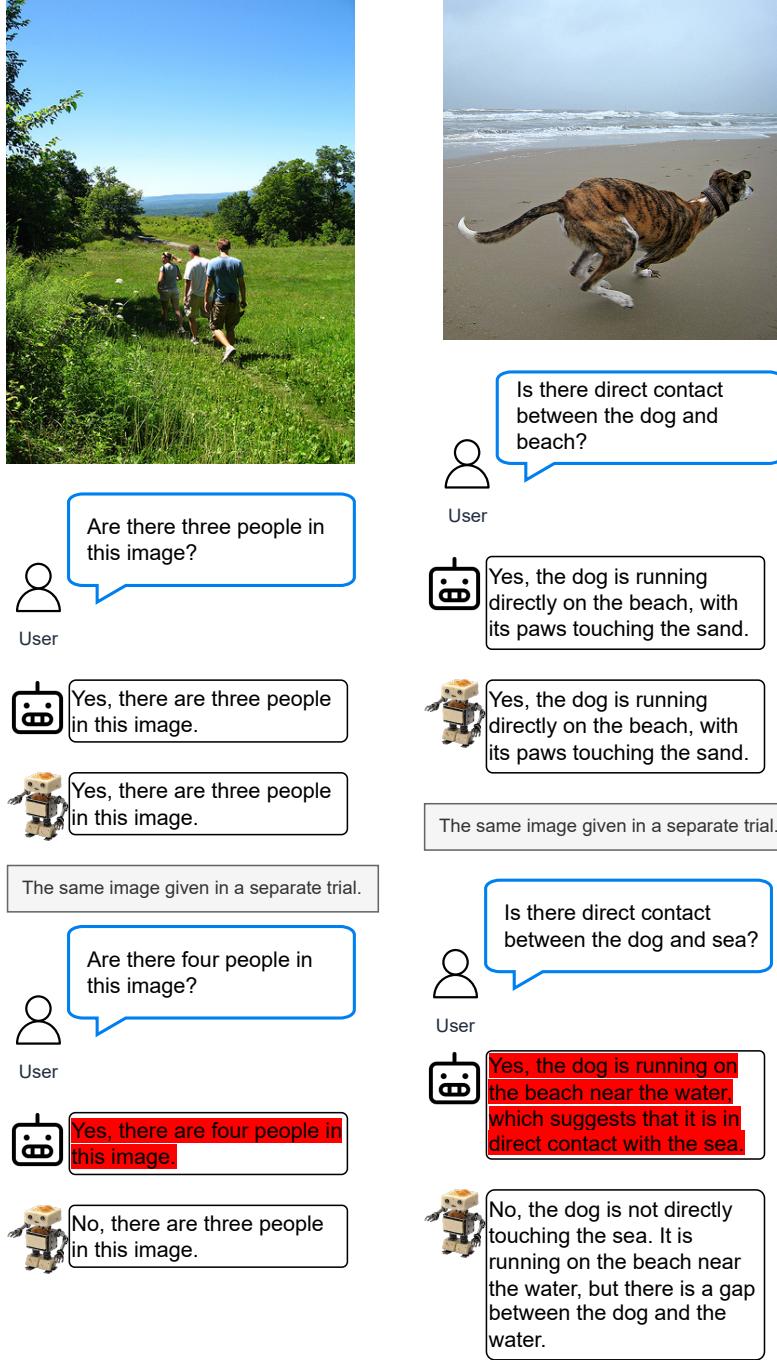


Figure S19: HALVA [1] does not exhibit a tendency to answer ‘Yes’ for any questions asked, like LLaVA-v1.5 [2]. Moreover, HALVA can accurately lay out the details of the image to further support the answer. As presented in the right side example, HALVA answers with “the dog is not directly touching the 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.

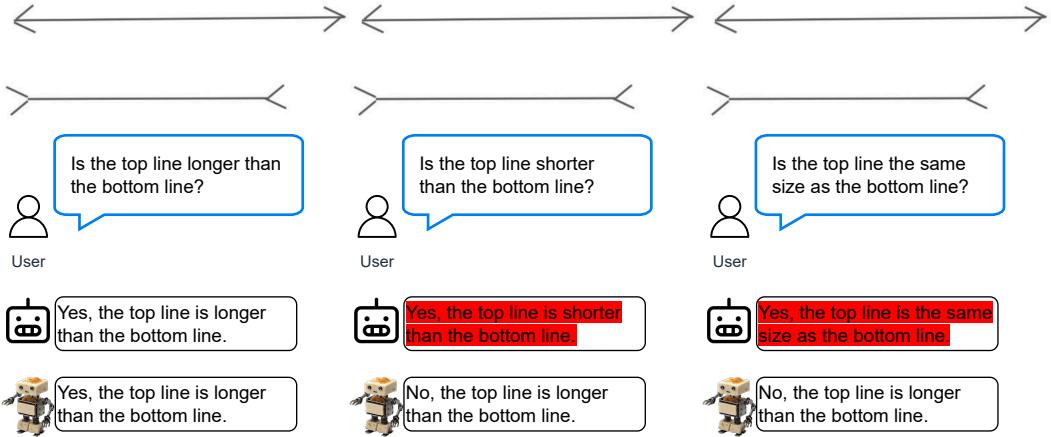


Figure S20: Examples illustrating the effectiveness of our proposed DPA in mitigating hallucinations that arise due to visual illusions. HALVA [?] does not exhibit a tendency to answer ‘Yes’ to any questions asked, like LLaVA-v1.5 [?]. Moreover, HALVA exhibits consistency in its responses unlike LLaVA-v1.5. Hallucinations are highlighted in red.



Figure S21: 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.

## E.2 Qualitative comparison between HALVA and VILA-v1.5

 <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">User</span> <p>Is there one knife in this image?</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">HALVA</span> <p>Yes, there is one knife in the image, which the man is holding.</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">VILA-v1.5</span> <p>Yes, there is one knife in the image.</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <p>The same image given in a separate trial.</p> </div>	 <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">User</span> <p>Is there direct contact between the dog and skate?</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">HALVA</span> <p>Yes, the dog is standing on top of the skateboard, with its paws on the skateboard.</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">VILA-v1.5</span> <p>Yes, the dog is standing on the skateboard, with its paws on the skateboard.</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <p>The same image given in a separate trial.</p> </div>
<div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">User</span> <p>Are there two knives in this image?</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">HALVA</span> <p>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.</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">VILA-v1.5</span> <p>No, there is only one knife in the image, which the man is holding.</p> </div>	<div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">User</span> <p>Is there direct contact between the dog and skate?</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">HALVA</span> <p>Yes, the dog is standing on the skateboard, which is in direct contact with the ground.</p> </div> <div style="border: 1px solid #ccc; padding: 5px; border-radius: 10px; margin-top: 10px;"> <span style="border-bottom: 2px solid #007bff; padding-bottom: 5px;">VILA-v1.5</span> <p>No, the dog is not directly touching the ground. It is standing on a skateboard, which is placed on the ground.</p> </div>

Figure S22: 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.



User

Are there two horses in this image?



Yes, there are two horses in this image.



Yes, there are two horses in this image.

The same image given in a separate trial.



User

Are there three horses in this image?



Yes, there are three horses in this image.



No, there are two horses in the image.



User

Is there direct contact between the giraffe and grass?



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.



User

Is there direct contact between the tree and giraffe?

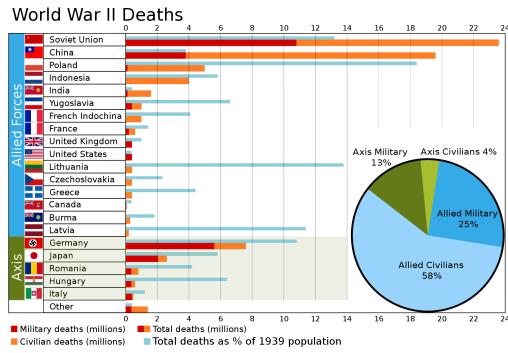


Yes, there is direct contact between the tree and the giraffe, as the giraffe is standing under the tree.



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 S23: HALVA [HALVA] does not exhibit a tendency to answer ‘Yes’ for any questions asked, like base model VILA-v1.5 [VILA]. 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.



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?

User

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.



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 S24: Examples illustrating the effectiveness of our proposed DPA in mitigating hallucinations that are beyond object centric. HALVA [HALVA] accurately answers to this chart-based question unlike VILA-v1.5 [VILA]. Hallucinations are highlighted in red.