Direct Preference Optimization for Perceptual Alignment via Vision-Language Consistency

Seungyeon Lee[1] and Dong-Gyu Lee[1] *

Department of Artificial Intelligence, Kyungpook National University,
Daegu, Republic of Korea
{statai3237,dglee}@knu.ac.kr

Abstract. Recent advances in Large Vision Language Models (LVLMs) have led to interest in alignment methods such as Direct Preference Optimization (DPO), which improves multi-modal instruction following without costly supervision or explicit reward modeling. However, previous attempts to improve DPO through data augmentation and preference ranking fall short of bridging vision-language perceptual gaps, limiting consistent multi-modal alignment. In this paper, we propose a novel **Vision-Language Consistency based Direct Preference Optimization (VLC-DPO)**, which automatically constructs preference data and leverages consistent information from visual inputs and corresponding textual descriptions as alignment signals. The VLC-DPO consists of three key components: (1) an initial response generation that produces image- and description-based responses to bridge the modality gap in LVLMs by capturing complementary aspects of visual understanding, (2) a preference data construction via vision-language consistency score that automatically selects high-quality preference pairs based on response consistency, and (3) a VLC-DPO training that uses consistency-based implicit rewards to optimize preferences, assigning dynamic weights to less image-related or distant samples to improve DPO learning. Extensive experiments on hallucination mitigation and zero-shot VQA demonstrate the effectiveness of our method, which outperforms existing approaches.

Keywords: Vision-Language Consistency · Direct Preference Optimization · Vision-Language Model

1 Introduction

Large Vision Language Models (LVLMs) have significantly improved visual comprehension with language instruction following ability across a wide range of multi-modal tasks such as visual question answering (VQA), hallucination mitigation [47,2,19]. Despite these advances, LVLMs are prone to generating plausible responses that include unnecessary or contextually irrelevant content, resulting in outputs misaligned with user preferences [51,46]. Addressing these issues

* Corresponding author

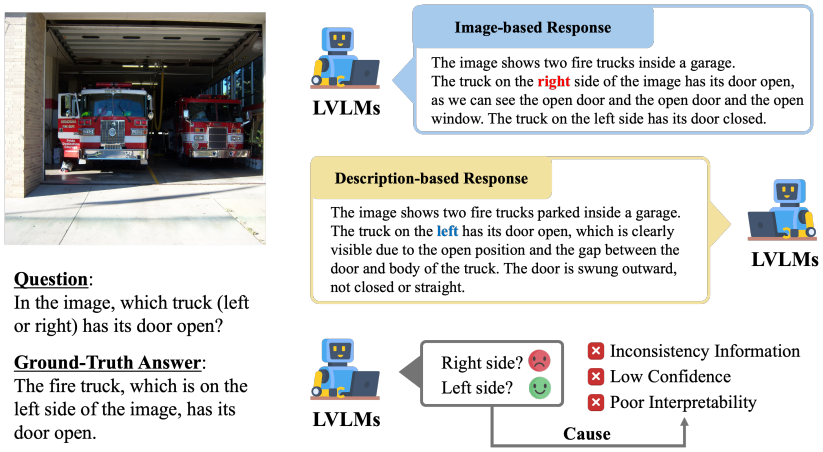

Fig. 1. Comparison between image-based and description-based responses from the same LVLMs. Although the image shows that the left door of the truck is open, the image-based response incorrectly states that the right door is open. In contrast, the description-based response directly identifies the left door. The example shows that it occurs due to the discrepancy in information contained in the visual-language model.

involves improving instruction-following capabilities, which in turn requires high-quality data collection and training, both of which demand substantial time and resources [45,40]. Recent research has increasingly investigated alignment techniques aimed at reducing reliance on high-quality labeled data and costly reward models, including those employed in reinforcement learning with human feedback (RLHF) [37,42].

A representative approach is Direct Preference Optimization (DPO) [26], which aligns models directly to preference data without the need for explicit reward modeling or reinforcement learning. In an effort to improve the performance and preference alignment quality of DPO, researchers have investigated methods for augmenting training data, including the use of synthetic variations and the integration of AI-generated preferences and rankings [24,42,20].

While prior studies have shown promising results, a key challenge remains: perceptual discrepancies between vision and language modalities, referring to the distinct methods each modality uses to represent and interpret the same information. The discrepancies cause inconsistencies and distortions in preference alignment, undermining the multi-modal understanding and reliability of the model [49,50,4]. As shown in Figure 1, existing LVLMs generate inconsistent outputs when responding to an image input versus a text description of the same content. These inconsistent results demonstrate that modality-specific differences in interpretation reduce the reliability and interpretability of model outputs. In light of these observations, the presence of unaddressed modality-specific discrepancies in the training data distorts preference signals, causing suboptimal alignment and degraded model performance.

Motivated by these limitations, we propose a novel **Vision-Language Con-**

sistency based Direct Preference Optimization (VLC-DPO) method that optimizes preferences using open-source LVLMs while ensuring consistency between the vision and language modalities. Specifically, VLC-DPO integrates visual and language-based representations by converting visual inputs into explicit textual descriptions and leveraging modality-consistency to guide preference optimization. The proposed VLC-DPO method consists of three main components: (1) **Initial Response Generation** obtains information from two modalities: it generates an image-based response using the input image and instruction, and a description-based response by producing a detailed caption from the image, then generating a response based on that caption. (2) **Preference data construction via vision-language consistency score (VLC score)** involves selecting automatically consistent response pairs using a new proposed VLC score, which measures the consistency between image-based and description-based outputs and their relevance with the input image. (3) **VLC-DPO training** aims to optimize preference using a self-generated preference pair dataset. The VLC-DPO objective assigns stronger penalties to preference pair when the two candidates exhibit substantial differences in vision-language consistency or image-text alignment. To steer the model effectively away from generating incoherent or irrelevant outputs.

We conduct extensive experiments on four hallucination mitigation benchmarks: MMHal-Bench [31], Object HalBench [27], AMBER [35], and POPE [16], and two VQA benchmarks: A-OKVQA [29], OK-VQA [22]. In the experiments, the VLC-DPO achieves state-of-the-art performance on the hallucination benchmarks AMBER and POPE, surpassing existing methods. Moreover, our approach outperforms prior methods in the zero-shot VQA setting, achieving a 15.1% relative improvement on A-OKVQA and a 5.3% gain on OK-VQA.
The main contributions are summarized as follows:

- We propose a novel VLC-DPO that leverages multi-modal consistency between vision and language to optimize the learning of response preference without any human-written data, ground-truth labels, or using a proprietary model.

- We introduce a new VLC score that jointly evaluates inter-response consistency and image-grounded relevance, enabling the automatic and scalable collection of high-quality preference pairs.

- We conduct comprehensive experiments to demonstrate the effectiveness of our proposed method across both hallucination mitigation, as well as zero-shot VQA tasks.

2 Related Work

2.1 Direct preference optimization in LVLMs

Direct preference optimization in LVLMs aims to align model outputs with human-like preferences of multi-modal responses. Wang et al. [34] introduces

mDPO, which prevents the decrease in response probability by introducing a reward anchor that forces the reward for the selected response to be positive. SeVa [53] proposes an unsupervised preference-ranking framework that constructs chosen and rejected responses from original-augmented image pairs, enabling robust answer learning via DPO training. Deng et al. [5] conduct self-training to generate preference data from unlabeled images and enhance reasoning with self-generated descriptions appended to instruction-tuning prompts. RLAIF-V [45] aligns LVLMs using iterative feedback of open-source models generated via a deconfounded divide-and-conquer approach. OPA-DPO [42] proposes as an on-policy alignment framework to correct hallucination responses by leveraging expert feedback and to align original responses with expert-corrected responses in an on-policy manner. HSA-DPO [37] presents a detect-and-rewrite pipeline using a proprietary model and introduce a hallucination severity-aware DPO that incorporates the severity of hallucinations into training. Compared to prior studies, we leverage vision-language consistency to construct high-reliability data and use consistency and image-text matching as adaptive reward signals.

2.2 Inconsistency in Vision-Language Modalities

Ensuring consistent responses from LVLMs is increasingly recognized as a crucial factor for building reliable and trustworthy systems [49,13,50]. Zhang et al. [50] point out that inconsistent responses from LVLMs to diverse prompts reduce user trust, and construct discriminative prompts based on low-confidence words. Wang et al. [33] suggest a caption bridge-based cross-modality alignment and contrastive learning model that leverages auxiliary captions closely related to visual information to reduce the semantic gap between vision and language modalities. Li et al. [15] use a visual description module to generate object-level and image-level text, enriching semantic information and reducing the modality gap by providing detailed descriptions. OmniCaptioner [21] bridges the vision-language modalities gap by generating fine-grained textual descriptions across diverse visual domains. KnowAda [43] addresses the modality gap through a data-centric fine-tuning method that adapts captions to an existing knowledge of model, effective reducing hallucinations while preserving descriptiveness. In this work, we construct training data for DPO using a newly proposed VLC score, which comprehensively evaluates both vision-language consistency and image-text alignment to select reliable preference pairs.

3 Method

3.1 Problem Definition

We consider the preference optimization task in vision-language response generation to mitigate hallucination. Given an initial set of responses $D = \{R_I, R_T\}$ generated from two modalities (image, text), the proposed method is able to automatically construct preference pairs by leveraging a VLC score. We assess

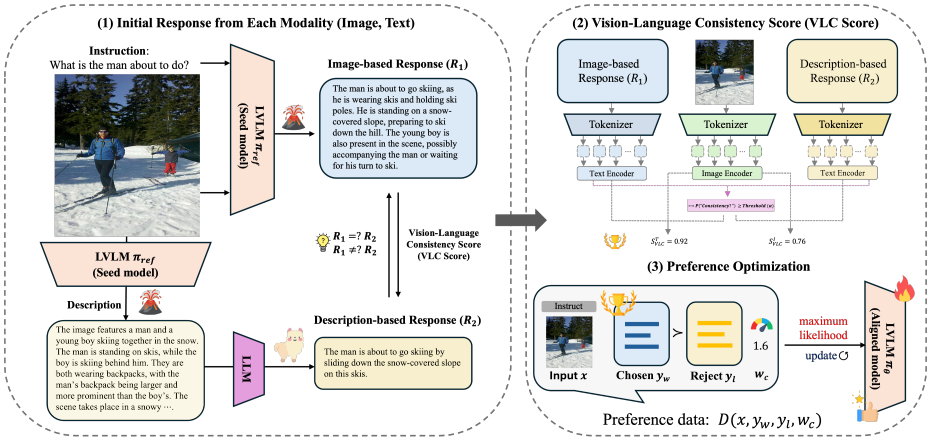

Fig. 2. Our proposed VLC-DPO consists of three main components: (1) Initial response generation from vision-language modalities, (2) Preference Data Construction via VLC Score, and (3) Preference optimization for VLC-DPO training using self-generated preference data pairs.

the degree of consistency and semantic alignment between the responses and their corresponding visual-textual inputs. The VLC-DPO training objective is to apply the magnitude of discrepancy in consistency and alignment between response pairs to enforce adaptive penalties. The overall flow of our proposed VLC-DPO is shown in Figure 2.

3.2 Preliminaries

Direct Preference Optimization. Direct Preference Optimization (DPO) [26] is a recently utilized alternative to traditional RLHF that simplifies the training pipeline without explicit reward modeling and reinforcement learning. DPO directly fine-tunes a policy model using preference data instead of optimizing a reward function. DPO optimizes a likelihood-based loss function that encourages the policy to prefer responses judged favorable by humans. Given a dataset of pairwise preference $D = \{x, y_w, y_l\}$, where y_w is the preferred response and y_l the dispreferred one for input x. DPO seeks to optimize policy π by maximizing the relative log-likelihood of the preferred responses relative to the dispreferred ones.
The DPO objective function is as follows:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\,\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\Big[\log\sigma\Big(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\Big)\Big], \tag{1}$$

$\log\pi(y|x)$ is formulated as:

$$\log \pi(y|x) = \sum_{y_i \in y} \log p(y_i|x, y_{<i}), \tag{2}$$

Here, π_θ is the current policy model parameterized by θ, and π_{ref} is a fixed reference model. The The β is a hyperparameter that controls how strongly the new policy deviates from the reference policy. The σ is a sigmoid function that maps the scaled log-likelihood difference to a probability, enabling preference learning as a binary classification task.

3.3 Initial Response Generation

In this paper, we introduce a method to build preference data for VLC-DPO training using open-source LVLMs, rather than using human-written or proprietary models (e.g., GPT-4V). We use a portion of the RLAIF-V [45] dataset to generate image and description-based responses. First, given an input image I_i, we generate the corresponding response R_{I_i} by using the open-source model LLaVA-1.5 [19], which is designed to elicit an instruction-aware response. Then, for each image I_i, we generate a detailed description D_i using LVLMs. The D_i and instructions are then used as input to a language-only model, such as Llama 2 [32], to produce the final text-based response R_{T_i}. Description-based responses serve as intermediate representations to connect visual and textual modalities, performing the role of bridging the gap between the multi-modal, following the approach of prior works [33,7]. We build an initial set of responses $D = \{R_{I_i}, R_{T_i}\}_i^N$ using image and text-based responses, where N is the total number of data.

3.4 Preference Data Construction via Vision-Language Consistency

In this phase, we introduce a vision-language consistency based data selection method for constructing preference data pairs from responses obtained for images and text. We present a new VLC score (S_{VLC}) by synthesizing vision-language semantic consistency (S_{SC}) and image-text matching (S_{ITM}) scores. The proposed VLC score is employed during data construction to evaluate the semantic consistency and image-text matching scores, effectively filtering out inaccurate captions and reducing the risk of misleading training feedback.

Semantic Consistency of Each Response. We evaluate sentence-level semantic consistency to filter out inconsistent image- and text-based responses from the initial response set D. Specifically, we employ a model pre-trained on natural language inference datasets to estimate the semantic relationship between sentence pairs[1]. Based on prior observations that low semantic consistency is associated with instability or hallucinated outputs [1,9,25], the resulting score is used to remove semantically inconsistent response pairs. Given two responses r_1 and r_2, we denote their respective token embedding matrices as $H_1 \in \mathbb{R}^{L_1 \times d}$

[1] https://huggingface.co/sentence-transformers/nli-mpnet-base-v2

and $H_2 \in \mathbb{R}^{L_2 \times d}$, where L_i is the token length of response and d is the embedding dimension. The corresponding attention masks are given as $m_1, m_2 \in \{0,1\}^{L_i}$. We compute sentence embeddings e_1 and e_2 as the attention-weighted mean of token embeddings:

$$e_i = \frac{\sum_{j=1}^{L_i} m_i^j \cdot h_i^j}{\sum_{j=1}^{L_i} m_i^j}, \qquad \text{for} \quad i \in \{1,2\} \tag{3}$$

where h_i^j is the embedding of the j-th token in response r_i, and m_i^j is the corresponding attention mask.
Finally, the semantic consistency score S_{SC} between two responses is computed as the cosine similarity of their embeddings:

$$S_{\text{SC}} = \frac{e_1 \cdot e_2}{||e_1|| \cdot ||e_2||} \tag{4}$$

Image-Text Matching Score. The image-text matching score is calculated using the weights of a pre-trained model that predicts the likelihood of an answer given an image and a question, which combines the image encoder CLIP and Flan-T5 with an encoder-decoder structure [18]. To evaluate the overall alignment between an image and a response, we decompose the response into individual sentences and compute the alignment score for each sentence. Specifically, for a given image I and a set of sentences $\{s_1, s_2, \cdots, s_n\}$ extracted from the response, we compute the generative likelihood of answering "yes" to a question derived from each sentence. The image-text matching score S_{ITM} is obtained by summing the individual sentence-level scores and dividing by the number of sentences:

$$S_{\text{ITM}} = \frac{1}{n} \sum_{i=1}^{n} P(\text{“yes”} | I, Q(s_i)) \tag{5}$$

Here, $Q(s_i)$ is a natural language question derived from the sentence s_i, using the prompt template following [18], where question prompt is "Does this figure show s_i? Please answer yes or no". The S_{ITM} allows for assessing whether the image semantically supports each sentence from the response, enabling fine-grained alignment scoring.

Vision-Language Consistency Score. We present a new VLC score that combines inter-response semantic consistency with image-text matching scores to generate high-quality preference data pairs. We begin by considering response pairs whose mutual consistency exceeds a pre-defined τ and assign a final score based on how closely each response aligns with the given image.

$$S_{\text{VLC}} = \begin{cases} \alpha \cdot S_{\text{SC}} + (1-\alpha) \cdot S_{\text{ITM}}, & \text{if} \quad S_{\text{SC}} \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where α adjusts the balance between S_{SC} and S_{ITM}, and τ is the threshold of the consistency score. The S_{VLC} reflects the internal agreement between responses and their individual relevance to the image. We ensure that selected responses are coherent and grounded in visual content by enforcing thresholds on both consistency and the composite score. We compute S_{VLC} for each R_I and R_T pair to obtain a final response score, as referred to S^I_{VLC} and S^T_{VLC}. We construct a preference pair where the response with a higher score is R_{chosen}, and the response with a lower score is R_{rejected}. To effectively optimize DPO training, it is crucial to construct high-confidence response pairs in which the distinction in rewards between the chosen and rejected responses is unambiguous [52,20,30]. That is, ensuring a sufficiently meaningful preference margin during DPO data construction helps the model better distinguish between chosen and rejected responses and contributes to more stable optimization [28,5]. To this end, we adopt a strategy of intentionally injecting noise into responses with relatively low scores when the scores of two responses are excessively similar. As part of our noise injection strategy, we perturb the response by randomly replacing words, masking words, or shuffling word order within the sentence. These perturbations are designed to slightly degrade the response quality while preserving grammatically and overall plausibility, rather than producing broken or nonsensical sentences. This noise injection strategy ensures that rejected responses have lower scores than the chosen responses, resulting in a clearer reward margin between the two responses.

3.5 VLC-DPO Training

We now conduct preference training using the automatically generated preference pair data based on VLC scores. To train VLC-DPO, we simply modify the standard DPO objective function in Eq. 1 to adjust the penalty using VLC score. The VLC-DPO adaptively optimizes preference by increasing the log-likelihood of chosen response y_w and decreasing the log-likelihood of rejected response y_l. In detail, our proposed method for VLC-DPO utilizes the difference in VLC scores as a consistency signal between S^I_{VLC} and S^T_{VLC} to assign a penalty with a large score difference. We adaptively incorporate the difference in VLC scores into the implicit reward model of DPO, so that responses with larger discrepancies in VLC scores receive stronger penalties. In this way, VLC-DPO encourages the model to prefer chosen responses, particularly those that higher consistency or semantic alignment. The proposed VLC-DPO method adjusts the contribution of each preference pair during optimization based on the difference score of VLC score as $|w_{\text{c}}|$. The continuous $|w_{\text{c}}|$ is the absolute value of the difference between S^I_{VLC} and S^T_{VLC}.

$$\mathcal{L}_{\text{VLC-DPO}}(\pi_\theta; \pi_{\text{ref}}) = \\ - \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - |w_{\text{c}}| \cdot \beta \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)} \right) \right], \quad (7)$$

where $|w_c|$ is the absolute value of the difference between S^I_{VLC} and S^T_{VLC}. The $|w_c|$ is scaled by 10 to amplify its role as a penalty during training, while being capped at a maximum of 3 to preserve optimization stability. The penalty value serves to suppress responses that lack consistency or relevance. We report the effectiveness of the VLC score in detail in Section 4.5 and the Appendix.

4 Experiments

4.1 Datasets

In the training, we construct a self-generated preference dataset by extracting 16k image-instruction pairs from the RLAIF-V [45] dataset, to accommodate a diverse range of knowledge sources[2]. After filtering, we use 11.5k (approximately 72%) preference pairs for training. We evaluate the performance of our proposed VLC-DPO against other baseline models on four hallucination benchmarks: MMHal-Bench [31], Object HalBench [27], AMBER [35], and POPE [16]. We also use two VQA benchmarks to compare the performance of zero-shot VQA: A-OKVQA [29] and OK-VQA [22].

Hallucination benchmarks: (1) **MMHal-Bench** is a benchmark designed to evaluate the hallucination problem in LVLMs under real-world scenarios, consisting of 96 samples across 12 object topics. The evaluation metric uses GPT-4 to compare model responses against ground-truth, measuring both response quality and hallucination rate. (2) **Object HalBench** is a dataset used for evaluating object-level hallucination. Following [44], we use 300 instances evaluated using 8 diverse prompts, with two metrics: response-level and mention-level hallucination rate. (3) **AMBER** is a multi-dimensional benchmark for hallucination analysis, consisting of over 15,000 examples. We use the discriminative task of the dataset and present results in terms of accuracy and F_1-score. (4) **POPE** is a benchmark that evaluates the hallucination discrimination ability of LVLMs by designing binary (yes or no) questions about the presence of objects in images. We use adversarial part of the dataset and evaluate the accuracy.

VQA benchmarks: (1) **A-OKVQA** is a crowdsourced dataset that requires general knowledge and knowledge to answer 25k questions. The train, validation, and test sets comprise 17.1k, 1.1k, and 6.7k and are composed into image/question/answer/rationale triplets. (2) **OK-VQA** is a knowledge-based visual question-answering dataset that requires external knowledge, containing more than 14k open-ended questions. It uses random images from the COCO [17] dataset, and consists of about 9k train sets and 5k test sets. We use the validation set of A-OKVQA and the test set of OK-VQA.

In this study, these datasets are utilized to investigate the effectiveness of DPO training in enhancing multi-modal responses. Specifically, hallucination datasets measure the model's ability to generate text faithful to visual content, directly

[2] In our preference learning, we use a subset of LCS-558K, COCO, VQAv2, OCR-VQA, ShareGPT4V, MovieNet, ART500K, and Google-Landmark as knowledge sources.

Table 1. Performance comparison with state-of-the-art methods on MMHal-Bench and Object HalBench benchmarks. Value in **bold** indicates the best performance, and the second best results are underlined. G.T indicates whether ground-truth is used during training. The 4 iter. indicates that RLAIF-V [45] was applied with 4 feedback iteration.

Model	Size	G.T	Feedback	MMHal-Bench		Object HalBench	
				Score (↑)	Hall. (↓)	Rsp. (↓)	Men. (↓)
AMP-MEG [48]	13B	✗	Rule	**3.08**	0.37	31.7	20.6
V-DPO [38]	7B	✗	LLaVA-1.5	2.16	0.56	-	-
mDPO [34]	7B	✓	GPT-4V	2.39	0.54	35.7	9.8
POVID [52]	7B	✓	Rule	1.98	0.60	39.9	19.9
RLHF-V [44]	13B	✗	Human	2.45	0.51	12.2	7.5
RLAIF-V [45]	7B	✗	LLaVA-NeXT (4 iter.)	2.95	**0.32**	10.5	5.2
oDPO [8]	7B	✓	LLaVA-1.5	2.50	0.49	34.3	9.5
OPA-DPO [42]	7B	✓	GPT-4V	2.83	0.45	13.0	4.25
HSA-DPO [37]	13B	✗	GPT-4, GPT-4V	2.61	0.48	**5.3**	**3.2**
DAMA [20]	7B	✓	LLaVA-1.5	2.76	0.41	9.1	4.7
LLaVA-1.5 [19]	7B	✗	✗	1.86	0.64	54.5	27.8
+ VLC-DPO (ours)	7B	✗	LLaVA-1.5	2.95	0.33	9.6	5.0
LLaVA-1.5 [19]	13B	✗	✗	2.36	0.56	50.0	23.6
+ VLC-DPO (ours)	13B	✗	LLaVA-1.5	2.98	**0.32**	10.1	5.6
GPT-4V [23]	-	-	Unknown	3.49	0.28	13.6	7.3

reflecting its capability to maintain factual consistency. VQA datasets assess a model's ability to integrate and reason over visual and textual information, providing a measure of multi-modal understanding. These two tasks provide a comprehensive evaluation of both factual grounding and reasoning skills, which are central to the objectives of VLC-DPO training.

4.2 Baselines

For hallucination mitigation task, we compare the proposed VLC-DPO with multiple vision-language alignment models for LVLMs, including AMP-MEG [48], V-DPO [38], mDPO [34], POVID [52], RLHF-V [44], RLAIF-V [45], oDPO [8], OPA-DPO [42], HSA-DPO [37], LLaVA-1.5 [19], and GPT-4V [23]. The results are taken from [45,42,20,37].

For zero-shot VQA task, we compare our proposed method with previous state-of-the-art methods. We use PICa [41], Img2LLM [7], LAMOC [6], L2A [39], VCTP [3], DIETCOKE [14], Brote-IM-XXL [36], and LLaVA-1.5 [19] as baseline models. The results are taken from [7,6,39,3,14,36]. We compare our method with several baselines using standard VQA scores: $\min(\frac{\#\text{human that provided that answer}}{3}, 1)$.

4.3 Implementation Details

We conduct experiments using LLaVA-1.5 7B and 13B, combining CLIP ViT-L/14@336px as a visual encoder and Vicuna 1.5 as a language decoder. Our backbone is chosen due to its strong reasoning performance on multi-modal instruction-tuning benchmarks and its widespread adoption in prior DPO studies [19,52,45]. Following prior work, we adopt the same backbone to enable fair

Table 2. Performance comparison with state-of-the-art methods on AMBER (Discriminative part) and POPE (Adversarial setting) benchmarks.

Model	Size	G.T	Feedback	AMBER (Disc.)		POPE (Adv.)
				Acc.	F_1	Acc.
AMP-MEG [48]	13B	✗	Rule	79.5	84.6	-
V-DPO [38]	7B	✗	LLaVA-1.5	-	81.6	-
POVID [52]	7B	✓	Rule	75.2	79.9	84.7
RLHF-V [44]	13B	✗	Human	72.6	75.0	-
RLAIF-V [45]	7B	✗	LLaVA-NeXT (4 iter.)	76.8	84.5	81.6
oDPO [8]	7B	✓	LLaVA-1.5	80.2	84.1	-
OPA-DPO [42]	7B	✓	GPT-4V	-	-	82.6
HSA-DPO [37]	13B	✗	GPT-4, GPT-4V	-	-	84.0
DAMA [20]	7B	✓	LLaVA-1.5	<u>83.3</u>	<u>87.0</u>	-
LLaVA-1.5 [19]	7B	✗	✗	73.5	77.7	80.8
+ VLC-DPO (ours)	7B	✗	LLaVA-1.5	83.1	**87.2**	<u>85.2</u>
LLaVA-1.5 [19]	13B	✗	✗	71.2	73.0	83.5
+ VLC-DPO (ours)	13B	✗	LLaVA-1.5	**83.5**	86.8	**85.6**
GPT-4V [23]	-	-	Unknown	83.4	87.4	-

comparison and ensure reproducibility. Our method is model-agnostic, and we further validate its applicability on additional LVLM backbones in the Appendix. We perform the experiments using 8 NVIDIA RTX A6000 GPUs. We train the VLC-DPO model for five epochs with a learning rate of 2e-7, a batch size of 16, and hyperparameters $\alpha = 0.5$ and $\beta = 0.1$. We fine-tune the LLaVA-1.5 model efficiently using LoRA (Low-Rank Adaptation) [10], with a rank of 128 and alpha set to 256. We also use Flash Attention to accelerate the training process by improving computational efficiency.

4.4 Main Results

Performance on Hallucination Mitigation Tables 1 and 2 report the quantitative results of our VLC-DPO compared to existing baselines. Our method achieves state-of-the-art performance, particularly on the AMBER (F_1-score) and POPE benchmarks, demonstrating the effectiveness of our hallucination mitigation. In this work, we utilize the open-source LLaVA-1.5 7B model without incorporating any ground-truth answers or feedback from proprietary models during training or inference. Despite this setup, our method outperforms prior approaches in reducing hallucinations, achieving superior results without reliance on external annotations. This demonstrates the effectiveness of leveraging vision-language consistency for data construction and adaptively assigning more substantial penalties to rewards plays a critical role in improving hallucination mitigation performance.

Performance on zero-shot VQA As shown in Table 3, VLC-DPO outperforms the baseline models on the A-OKVQA and OK-VQA datasets. Following [19], we append a prompt "Answer the question using a single word or phrase." to the end of the question to handle short answers. The VLC-DPO demonstrates strong performance in zero-shot VQA settings. These results show that ensuring

Table 3. Main experimental results on A-OKVQA and OK-VQA for zero-shot VQA performance comparison. Value in **bold** indicate the best performance, and the second best results are underlined.

Method	A-OKVQA	OK-VQA
PICa [41]	42.4	42.9
Img2LLM [7]	42.9	45.6
LAMOC [6]	37.9	40.3
L2A [39]	48.5	46.2
VCTP [3]	53.2	56.2
DIETCOKE [14]	47.5	49.2
Brote-IM-XXL [36]	57.8	-
LLaVA-1.5 7B [19]	46.2	46.4
+ VLC-DPO (ours)	<u>60.7</u>	<u>57.3</u>
LLaVA-1.5 13B [19]	52.3	51.8
+ VLC-DPO (ours)	**66.5**	**59.2**

Table 4. Ablation results of vision-language consistency usage strategies on the AMBER and POPE benchmarks.

Methods	AMBER		POPE
	Accuracy	F_1	Accuracy
VLC-DPO (7B)	**83.1**	**87.2**	**85.2**
w/o description	80.4	82.0	82.6
w/o consistency	78.5	81.7	81.8
w/o penalty (w_c)	75.0	77.9	81.0

vision-language consistency plays a significant role in the overall performance improvements. On the A-OKVQA dataset, VLC-DPO achieves a VQA score of 66.5, representing an improvement of approximately 15.1% over the prior state-of-the-art score of 57.8. Similarly, on the OK-VQA dataset, our approach achieved 59.2, surpassing previous methods and setting a new state-of-the-art performance. We attribute this improvement to our use of vision-language consistency as a guiding signal for alignment, which effectively filters and guides the model toward more faithful and contextually relevant responses, thereby enhancing zero-shot VQA performance.

4.5 Ablation study

Effect of vision-language consistency strategy. We conduct an ablation study comparing model input selection strategies to prove the effect of vision-language consistency. Specifically, we evaluate model performance when generating responses based on (1) image-only responses, (2) consistency disabled, and (3) vision-language responses without an added consistency constraint. Table 4 shows that incorporating vision-language consistency enables more accurate response data filtering, which facilitates the construction of high-quality

Table 5. Ablation results of the VLC scores threshold (τ) on the AMBER and POPE benchmarks.

Threshold (τ)	AMBER		POPE
	Accuracy	F_1	Accuracy
0.5	81.0	82.9	83.2
0.6	82.4	86.0	84.1
0.7	**83.1**	**87.2**	**85.2**
0.8	82.6	86.1	84.7

Table 6. Impact of different noise types (noise-free, replacement, masking, order shuffling, mixture) on rejected responses.

Type of Noise	AMBER		POPE
	Accuracy	F_1	Accuracy
Noise-free	80.1	81.4	81.6
Replacement	81.9	86.1	83.9
Masking	80.8	82.5	83.0
Order Shuffling	76.2	79.0	81.3
Mixture	**83.1**	**87.2**	**85.2**

training data and training. This improved data quality contributes to better performance in hallucination mitigation. We observe that leveraging consistency across modalities effectively guides the model toward faithful and semantically aligned responses, while also serving as a training signal.

Effect of VLC score. In Table 5, we assess the impact of the VLC score by comparing performance under different threshold values τ, which controls how strongly the score influences preference response selection. We use a VLC score that reflects both image-text alignment and internal response consistency to guide data selection. We find that a τ value of 0.7 yields the best overall performance, automatically selecting high-quality preference training data samples. This suggests that the threshold of 0.7 results in stable and reliable performance, retaining visually relevant and coherent data. We further validate the effectiveness of our proposed VLC score based training by constructing preference data pairs through an LVLM-as-a-Judge based on LLaVA-1.5 13B. The results are presented in Figure 3. The experimental results demonstrate that our VLC score-based method achieves superior performance compared to the simple LVLM-as-a-Judge approach. Interestingly, we have observed that constructing chosen-rejected pairs using an LVLM-as-a-Judge through pairwise comparisons is susceptible to position bias, exhibiting order-dependent judgments similar to those observed in previous works [11,12]. These observations further underscore the importance of the proposed VLC score-based data construction strategy.

Impact of noise injection on model performance. We compare different noise strategies applied to rejected responses to assess their role in increasing

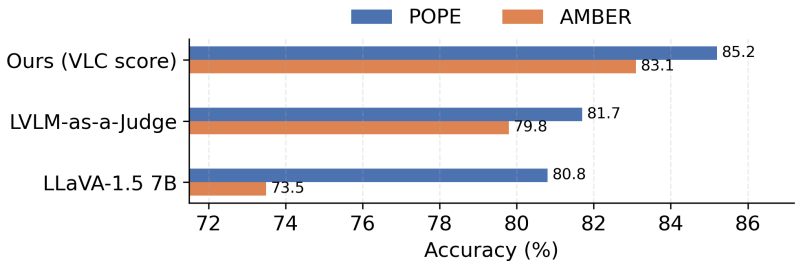

Fig. 3. Accuracy comparison of the base LLaVA-1.5 7B, LVLM-as-a-Judge, and the proposed VLC score based method on the POPE and AMBER benchmarks.

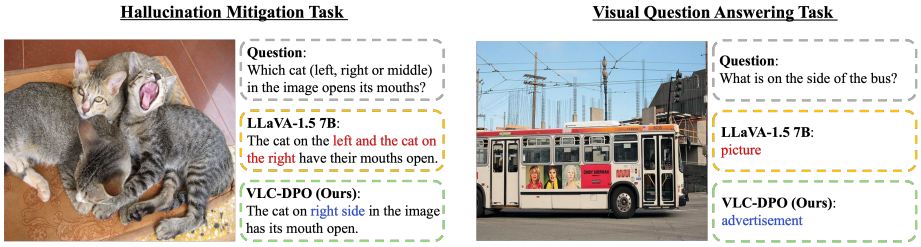

Fig. 4. Qualitative results of VLC-DPO. Comparison of model outputs: (Top) MMHal-Bench for hallucination mitigation, and (Bottom) A-OKVQA for VQA task.

distinction from chosen responses under the VLC score filtering. To analyzing the impact of different noise types on overall performance, we conduct a comparative study using perturbation strategies: word replacement, masking, order shuffling, and mixture setting. The mixture setting is composed of 40% replacement, 40% masking, and 20% order shuffling, which yields the best performance. Table 6 shows that incorporating diverse noise types during data generation helps the model generalize better and learn more robust patterns, compared to using any single noise type or a noise-free setting.

4.6 Qualitative Analysis

Figure 4 shows qualitative results that the VLC-DPO model generates more accurate and reliable responses than the LLaVA-1.5 on both the hallucination and VQA tasks. In the hallucination task, the VLC-DPO model substantially reduces the generation of content that is unsupported by the image. This result indicates improved vision-language alignment, resulting in responses that are faithful to the visual evidence and semantically consistent with the input images. For the VQA task, our model generates answers that are closely aligned with the

visual content than those generated by LLaVA-1.5. These results highlight that our approach enhances the model's ability to align visual cues with linguistic context, resulting in more accurate and contextually relevant responses.

5 Discussion

In this study, we propose a new preference data construction and optimization strategy to ensure consistency between the vision and language modalities. In contrast to existing methods, our VLC-DPO method incorporates visual-language consistency to guide preference data construction and alignment, reducing inconsistencies and distortions and enhancing multi-modal understanding. This leads to improved alignment between visual and textual modalities, yielding more consistent and reliable model outputs. The proposed VLC score is designed to improve robustness against false positive in data construction phase and ensure compatibility with various LVLMs architectures. Despite these considerations, our DPO training is conducted at the response level without incorporating local feedback on specific tokens, phrases, or objects, which limits the model's ability to detect fine-grained errors and effectively reduce hallucinations. Although our data construction effectively isolates alignment and consistency differences, some pairs happen to share the same hallucination across both responses, an issue worth addressing in subsequent work. In future work, we plan to extend reward modeling to the object or token level to enable finer-grained learning, which will also help address the shared hallucination issue and further improve multi-modal consistency and reliability.

6 Conclusion

In this work, we present a novel VLC-DPO that ensures consistency between the vision and language modalities. The proposed method increases the trustworthiness and correctness of open-source LVLMs without any human intervention, ground-truth labels, or proprietary model. To bridge the gap between vision and language information, we propose a new VLC score, which incorporates both response consistency and alignment with the image. Our method is able to utilize the proposed VLC score to automatically generate high-quality responses and use them as preference training data. Moreover, we use the difference in VLC scores between image and description-based responses as an adaptive reward signal, encouraging the model to maximize the reward margins. Extensive experiments have demonstrated the effectiveness of our VLC-DPO in hallucination mitigation and VQA tasks.

Acknowledgements

Please insert your acknowledgments here. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean Government (MSIT) (No. RS-2025-23324166) and (No. RS-2025-02214941).

References

1. Abdaljalil, S., Kurban, H., Sharma, P., Serpedin, E., Atat, R.: Sindex: Semantic inconsistency index for hallucination detection in llms. arXiv preprint arXiv:2503.05980 (2025)
2. Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A.A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.C., Chen, Y.L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A.D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R.J., Hu, W., Huynh, J., Iter, D., Jacobs, S.A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y.J., Kurilenko, L., Lee, J.R., Lee, Y.T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C.C.T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L.L., Zhang, Y., Zhang, Y., Zhang, Y., Zhou, X.: Phi-3 technical report: A highly capable language model locally on your phone (2024), https://arxiv.org/abs/2404.14219
3. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. Proceedings of the AAAI Conference on Artificial Intelligence **38**(2), 1254–1262 (Mar 2024). https://doi.org/10.1609/aaai.v38i2.27888, https://ojs.aaai.org/index.php/AAAI/article/view/27888
4. Dagan, G., Loginova, O., Batra, A.: CAST: Cross-modal alignment similarity test for vision language models. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B.D., Schockaert, S. (eds.) Proceedings of the 31st International Conference on Computational Linguistics. pp. 1387–1402. Association for Computational Linguistics, Abu Dhabi, UAE (Jan 2025), https://aclanthology.org/2025.coling-main.93/
5. Deng, Y., Lu, P., Yin, F., Hu, Z., Shen, S., Gu, Q., Zou, J., Chang, K.W., Wang, W.: Enhancing large vision language models with self-training on image comprehension. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=FZW7Ctyjm3
6. Du, Y., Li, J., Tang, T., Zhao, W.X., Wen, J.R.: Zero-shot visual question answering with language model feedback. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 9268–9281. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.590, https://aclanthology.org/2023.findings-acl.590/
7. Guo, J., Li, J., Li, D., Tiong, A.M.H., Li, B., Tao, D., Hoi, S.: From images to textual prompts: Zero-shot visual question answering with frozen large language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10867–10877 (2023)

8. He, Y., Sun, H., Ren, P., Wang, J., Wang, H., Qi, Q., Zhuang, Z., Wang, J.: Evaluating and mitigating object hallucination in large vision-language models: Can they still see removed objects? In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 6841–6858. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.349, https://aclanthology.org/2025.naacl-long.349/
9. Hossain, A.I., Choudhury, S.R., Alhoori, H.: SciHallu: A multi-granularity hallucination detection dataset for scientific writing. In: Inui, K., Sakti, S., Wang, H., Wong, D.F., Bhattacharyya, P., Banerjee, B., Ekbal, A., Chakraborty, T., Singh, D.P. (eds.) Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. pp. 1277–1304. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, Mumbai, India (Dec 2025), https://aclanthology.org/2025.ijcnlp-long.70/
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
11. Hwang, Y., Lee, D., Min, K., Kang, T., Kim, Y., Jung, K.: Fooling the LVLM judges: Visual biases in LVLM-based evaluation. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 23186–23205. Association for Computational Linguistics, Suzhou, China (Nov 2025). https://doi.org/10.18653/v1/2025.emnlp-main.1182, https://aclanthology.org/2025.emnlp-main.1182/
12. Laskar, M.T.R., Islam, M.S., Mahbub, R., Masry, A., Rahman, M., Bhuiyan, A., Nayeem, M.T., Joty, S., Hoque, E., Huang, J.X.: Judging the judges: Can large vision-language models fairly evaluate chart comprehension and reasoning? In: Rehm, G., Li, Y. (eds.) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track). pp. 1203–1216. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.acl-industry.83, https://aclanthology.org/2025.acl-industry.83/
13. Lee, J., Son, J., Seok, J., Jang, W., Kwon, Y.D.: Preference consistency matters: Enhancing preference learning in language models with automated self-curation of training corpora. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 12150–12169. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.606, https://aclanthology.org/2025.naacl-long.606/
14. Li, M., Li, H., Du, Z., Li, B.: Diversify, rationalize, and combine: Ensembling multiple QA strategies for zero-shot knowledge-based VQA. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 1552–1566. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.84, https://aclanthology.org/2024.findings-emnlp.84/
15. Li, P., Si, Q., Fu, P., Lin, Z., Wang, Y.: Object attribute matters in visual question answering. In: Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial

Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence. AAAI'24/IAAI'24/EAAI'24, AAAI Press (2024). https://doi.org/10.1609/aaai.v38i17.29816, https://doi.org/10.1609/aaai.v38i17.29816
16. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.20, https://aclanthology.org/2023.emnlp-main.20/
17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
18. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. In: European Conference on Computer Vision. pp. 366–384. Springer (2024)
19. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 26296–26306 (June 2024)
20. Lu, J., Wu, J., Li, J., Jia, X., Wang, S., Zhang, Y., Fang, J., Wang, X., He, X.: DAMA: Data- and model-aware alignment of multi-modal LLMs. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=zvZTIXkzDe
21. Lu, Y., Yuan, J., Li, Z., Zhao, S., Qin, Q., Li, X., Zhuo, L., Wen, L., Liu, D., Cao, Y., Yan, X., Li, X., Peng, T., Zhang, S., Shi, B., Chen, T., Chen, Z., Bai, L., Gao, P., Zhang, B.: Omnicaptioner: One captioner to rule them all (2025), https://arxiv.org/abs/2504.07089
22. Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: Proceedings of the IEEE/cvf conference on computer vision and pattern recognition. pp. 3195–3204 (2019)
23. OpenAI: Gpt-4v(ision) system card (2023), https://api.semanticscholar.org/CorpusID:263218031
24. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In: European Conference on Computer Vision. pp. 395–413. Springer (2024)
25. Park, S., Du, X., Yeh, M.H., Wang, H., Li, Y.: Steer LLM latents for hallucination detection. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=UMqNQEPNT3
26. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: your language model is secretly a reward model. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
27. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Riloff, E., Chiang, D., Hockenmaier, J., Tsujii, J. (eds.) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 4035–4045. Association for Computational Linguistics, Brussels, Belgium (Oct-Nov 2018). https://doi.org/10.18653/v1/D18-1437, https://aclanthology.org/D18-1437/
28. Sarkar, P., Ebrahimi, S., Etemad, A., Beirami, A., Arik, S.O., Pfister, T.: Mitigating object hallucination via data augmented contrastive tuning. CoRR **abs/2405.18654** (2024), https://doi.org/10.48550/arXiv.2405.18654

29. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
30. Sun, J., Wu, J., Wu, J., Zhu, Z., Lu, X., Zhou, J., Ma, L., Wang, X.: Robust preference optimization via dynamic target margins (2025), https://arxiv.org/abs/2506.03690
31. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L., Wang, Y.X., Yang, Y., Keutzer, K., Darrell, T.: Aligning large multimodal models with factually augmented RLHF. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Findings of the Association for Computational Linguistics: ACL 2024. pp. 13088–13110. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.findings-acl.775, https://aclanthology.org/2024.findings-acl.775/
32. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
33. Wang, B., Ma, Y., Li, X., Gao, J., Hu, Y., Yin, B.: Bridging the cross-modality semantic gap in visual question answering. IEEE Transactions on Neural Networks and Learning Systems **36**(3), 4519–4531 (2025). https://doi.org/10.1109/TNNLS.2024.3370925
34. Wang, F., Zhou, W., Huang, J.Y., Xu, N., Zhang, S., Poon, H., Chen, M.: mDPO: Conditional preference optimization for multimodal large language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 8078–8088. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.emnlp-main.460, https://aclanthology.org/2024.emnlp-main.460/
35. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., Sang, J.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation (2024), https://arxiv.org/abs/2311.07397
36. Wang, Z., Chen, C., Zhu, Y., Luo, F., Li, P., Yan, M., Zhang, J., Huang, F., Sun, M., Liu, Y.: Browse and concentrate: Comprehending multimodal content via prior-LLM context fusion. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 11229–11245. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024). https://doi.org/10.18653/v1/2024.acl-long.605, https://aclanthology.org/2024.acl-long.605/
37. Xiao, W., Huang, Z., Gan, L., He, W., Li, H., Yu, Z., Shu, F., Jiang, H., Zhu, L.: Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. Proceedings of the AAAI Conference on Artificial Intelligence **39**(24), 25543–25551 (Apr 2025). https://doi.org/10.1609/aaai.v39i24.34744, https://ojs.aaai.org/index.php/AAAI/article/view/34744
38. Xie, Y., Li, G., Xu, X., Kan, M.Y.: V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 13258–13273. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.775, https://aclanthology.org/2024.findings-emnlp.775/
39. Xing, X., Xiong, P., Fan, L., Li, Y., Wu, Y.: Learning to ask denotative and connotative questions for knowledge-based VQA. In: Al-Onaizan, Y., Bansal,

M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 8301–8315. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024). https://doi.org/10.18653/v1/2024.findings-emnlp.487, https://aclanthology.org/2024.findings-emnlp.487/

40. Xu, P., Shao, W., Zhang, K., Gao, P., Liu, S., Lei, M., Meng, F., Huang, S., Qiao, Y., Luo, P.: Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
41. Yang, Z., Gan, Z., Wang, J., Hu, X., Lu, Y., Liu, Z., Wang, L.: An empirical study of gpt-3 for few-shot knowledge-based vqa. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 3081–3089 (2022)
42. Yang, Z., Luo, X., Han, D., Xu, Y., Li, D.: Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 10610–10620 (June 2025)
43. Yanuka, M., Ben-Kish, A., Bitton, Y., Szpektor, I., Giryes, R.: Bridging the visual gap: Fine-tuning multimodal models with knowledge-adapted captions. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 10497–10518. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025). https://doi.org/10.18653/v1/2025.naacl-long.527, https://aclanthology.org/2025.naacl-long.527/
44. Yu, T., Yao, Y., Zhang, H., He, T., Han, Y., Cui, G., Hu, J., Liu, Z., Zheng, H.T., Sun, M., et al.: Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13807–13816 (2024)
45. Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., Chen, D., Lu, X., Cui, G., Dang, Y., He, T., Feng, X., Song, J., Zheng, B., Liu, Z., Chua, T.S., Sun, M.: Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 19985–19995 (June 2025)
46. Zhang, C., Wang, S., Li, X., Zhu, Y., Qi, H., Huang, Q.: Enhancing the robustness of vision-language foundation models by alignment perturbation. IEEE Transactions on Information Forensics and Security (2025)
47. Zhang, J., Huang, J., Jin, S., Lu, S.: Vision-language models for vision tasks: A survey. IEEE transactions on pattern analysis and machine intelligence **46**(8), 5625–5644 (2024)
48. Zhang, M., Wu, W., Lu, Y., Song, Y., RONG, K., Yao, H., Zhao, J., Liu, F., Feng, H., Wang, J., Sun, Y.: Automated multi-level preference for MLLMs. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=woENr7FJaI
49. Zhang, X., Li, S., Shi, N., Hauer, B., Wu, Z., Kondrak, G., Abdul-Mageed, M., Lakshmanan, L.V.S.: Cross-modal consistency in multimodal large language models (2024), https://arxiv.org/abs/2411.09273
50. Zhang, Y., xiao, F., Huang, T., Fan, C.K., Dong, H., Li, J., Wang, J., Cheng, K., Zhang, S., Guo, H.: Unveiling the tapestry of consistency in large vision-language models. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2024), https://openreview.net/forum?id=tu1oC7zHGW

51. Zhao, Y., Yin, Y., Li, L., Lin, M., Huang, V.S.J., Chen, S., Chen, W., Yin, B., Zhou, Z., Zhang, W.: Beyond sight: Towards cognitive alignment in lvlm via enriched visual knowledge. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24950–24959 (2025)
52. Zhou, Y., Cui, C., Rafailov, R., Finn, C., Yao, H.: Aligning modalities in vision large language models via preference fine-tuning. arXiv preprint arXiv:2402.11411 (2024)
53. Zhu, K., Zhao, L., Ge, Z., Zhang, X.: Self-supervised visual preference alignment. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 291–300 (2024)