Joint Multimodal Preference Optimization for Fine-Grained Visual-Textual Alignment

Jiwon Kim
Yonsei University
hanajibsa@yonsei.ac.kr

Hyunsoo Yoon†
Yonsei University
hs.yoon@yonsei.ac.kr

Abstract

Recent research has focused on addressing multimodal hallucinations in Large Vision-Language Models (LVLMs) by extending Direct Preference Optimization (DPO) to incorporate visual preference supervision. However, these methods often lack fine-grained visual contrast mechanisms and rely on single-margin optimization. This in turn limits their ability to capture precise visual semantics and results in weak multimodal alignment. To address these issues, we propose Joint Multimodal Preference Optimization (JoMPO), a novel optimization framework that symmetrically integrates a text-conditioned preference loss with a visual ranking-based objective. JoMPO leverages semantically contrastive image–text pairs and listwise ranking over multiple visual contexts, enabling fine-grained visual grounding and more robust cross-modal alignment. To support this framework, we introduce the Visual–Textual Contrast (VTC) dataset[1], consisting of image pairs that are semantically similar but visually distinct, each paired with a contextually grounded textual response. When trained with only 5k contrastive pairs, JoMPO consistently demonstrates superior performance across diverse benchmarks, highlighting its effectiveness in mitigating hallucinations and improving image-text alignment in LVLMs.

1 Introduction

Recent advances in large-scale AI models have led to the emergence of Large Vision-Language Models (LVLMs), which extend the capabilities of text-oriented Large Language Models (LLMs) by incorporating visual encoders through alignment modules (Zhang et al., 2024; Yin et al., 2024). Notable examples such as GPT-4o (Hurst et al., 2024), LLaVA, (Liu et al., 2023, 2024b), and Qwen-VL (Wang et al., 2024c) demonstrate strong multimodal capabilities in visual understanding and reasoning. Despite these achievements, LVLMs still exhibit fundamental limitations in visual-textual alignment (Sun et al., 2023). This often leads the model to neglect visual inputs and rely predominantly on language (Gupta et al., 2022; Zhu, 2023), resulting in *multimodal hallucinations* where the generated responses are inconsistent with the given visual content (Li et al., 2023; Bai et al., 2024; Zhou et al., 2023). This phenomenon poses a significant challenge to the reliability and practical deployment of LVLMs, and has led to increasing research efforts aimed at mitigating hallucinations (Liu et al.,

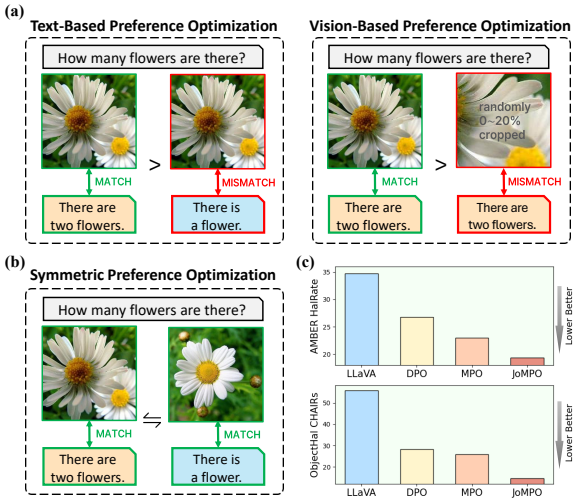

Figure 1: **Existing strategies**: Text-based preference optimization (PO) extends standard DPO by incorporating images, while vision-based PO, lacking contrastive pairs, treats randomly 0∼20% cropped images as dispreferred. (b) **Symmetric PO**: JoMPO employs a symmetric structure between contrastive image pairs and their associated text. (c) **Impact of JoMPO**: Applied to the LLaVA-1.5-7B model, JoMPO outperforms both text-based PO (DPO) and MPO, which incorporates vision-based PO into DPO, in reducing hallucination rates and demonstrates superior image-text alignment.

†Corresponding Author.

[1]Datasets are available at https://huggingface.co/datasets/arr-2025-jompo/VTC-5k.

Findings of the Association for Computational Linguistics: EACL 2026, pages 79–94
March 24-29, 2026

2024a; Lan et al., 2024).

Motivated by the success of Direct Preference Optimization (DPO) in aligning LLMs with human preferences (Rafailov et al., 2023), recent studies have extended DPO to the multimodal domain to mitigate hallucinations (Xiao et al., 2025; Yu et al., 2024a). A common approach is response-level preference optimization, which encourages the model to prefer non-hallucinated response y_w over hallucinated response y_l, given the input pair (m, x) of image and prompt. More recently, this formulation has been extended to incorporate visual preference optimization module by constructing visual preference pairs, such as (m_w, x, y_w) and (m_l, x, y_w), where m_w denotes a clean image and m_l is a visually perturbed version through methods like added noise or cropping (Wang et al., 2024a; Jiang et al., 2024; Fu et al., 2025), as depicted in Fig. 1(a). This enables the model to learn visually grounded preference signals instead of relying solely on textual priors (Yang et al., 2025; Lu et al., 2025).

While these multimodal preference optimization approaches effectively mitigate hallucinations, they still exhibit two fundamental limitations. First, using heavily degraded images as rejected samples makes contrasts between positive and negative examples overly explicit, leading the model to rely on superficial visual cues such as artifacts instead of understanding underlying visual semantics (Wu et al., 2025). Consequently, it fails to attend to fine-grained visual content. Second, existing approaches predominantly construct contrastive supervision only at the textual level while conditioning on the same or slightly perturbed visual input. This lack of semantically distinct image contrasts weakens direct visual supervision and leads to an asymmetric training signal dominated by textual differences, ultimately hindering the model's ability to learn balanced cross-modal associations and achieve visually grounded reasoning.

To overcome these limitations, we propose **Joint Multimodal Preference Optimization (JoMPO)**, a novel alignment framework that symmetrically integrates fine-grained visual and textual preference learning. JoMPO introduces a visual ranking-based objective that orders multiple visual conditions by their semantic compatibility with the associated text, employing a listwise ranking loss (Rafailov et al., 2023; Liu et al., 2024c) to capture fine-grained visual semantics. By jointly optimizing visual and textual preferences in a symmetric framework, JoMPO promotes context-aware alignment between images and text rather than relying on absolute input quality.

To support fine-grained multimodal alignment, we construct the **Visual-Textual Contrast (VTC) Dataset**, which provides semantically contrastive supervision across both visual and textual dimensions. Unlike prior datasets that contrast only textual responses, VTC includes visually similar yet subtly different image pairs selected via embedding discrepancies between a text-supervised vision-language model and a vision-only self-supervised model (Fig. 1(b)). Each pair shares a common prompt and includes tailored responses emphasizing their visual differences, facilitating joint optimization of visual and textual preferences for more precise multimodal grounding.

To evaluate JoMPO, we run experiments on hallucination benchmarks including AMBER (Wang et al., 2024b), Object HalBench (Rohrbach et al., 2018), and MMHal-Bench (Sun et al., 2023), as well as on general-purpose and vision-centric multimodal benchmarks. Evaluated on LLaVA-1.5-7B and 13B (Liu et al., 2024b) backbones, JoMPO consistently outperforms the base models, reducing hallucination rates on Object HalBench by 77.9% and 74.5%, respectively, while also achieving improvements on general-purpose and vision-centric benchmarks. As shown in Fig. 1(c), it also surpasses both standard DPO and multimodal DPO variants using rejected images, achieving these gains with only 5k preference pairs, demonstrating strong data efficiency.

The main contributions of this work are:

- We propose JoMPO, a principled multimodal alignment framework that jointly optimizes listwise visual rankings and textual preferences in a symmetric architecture.
- We introduce the VTC Dataset, offering high-quality semantically contrastive image–text pairs for fine-grained multimodal understanding.
- JoMPO demonstrates strong data efficiency and achieves superior hallucination mitigation and performance across multiple benchmarks.

2 Preliminaries

2.1 Pairwise Preference Optimization

2.1.1 Direct Preference Optimization

DPO (Rafailov et al., 2023) is used for aligning Large Vision-Language Models (LVLMs) with-

out relying on explicit reward modeling or reinforcement learning. In contrast, Reinforcement Learning from Human or AI Feedback (RLHF/RLAIF) (Ouyang et al., 2022; Lee et al., 2023) typically optimize a policy model π_θ based on feedback from a learned reward model $r(x, m, y)$, where y is a response generated given a prompt x and an image m. These methods additionally enforce a constraint through a reference policy π_{ref}. The fundamental learning objective is formulated as follows:

$$\max_{\pi_\theta} \mathbb{E}_{\mathcal{D}}[r(x, m, y)] - \beta D_{KL}\big[\pi_\theta(\cdot|x, m) \| \pi_{ref}(\cdot|x, m)\big], \tag{1}$$

where $\mathcal{D}$ is the prompt-image dataset and β is a parameter that controls the regularization strength.

DPO derives the closed-form optimal solution to Eq. 1, yielding the implicit reward formulation

$$r(x, m, y) = \beta \log \frac{\pi_\theta(y|x,m)}{\pi_{ref}(y|x,m)} + \beta \log Z(x, m), \tag{2}$$

where $Z(x, m)$ is a partition function. Incorporating the Bradley–Terry model (Bong and Rinaldo, 2022) for pairwise preferences, DPO directly optimizes over a pairwise preference dataset $(y_w \succ y_l)$, where y_w denotes the preferred response and y_l is the dispreferred one, as follows:

$$\begin{aligned}\mathcal{L}_{DPO} &= -\mathbb{E}_{\mathcal{D}}\big[\log \sigma(r(x, m, y_w) - r(x, m, y_l))\big] \\ &= -\mathbb{E}_{\mathcal{D}}\big[\log \sigma(\beta \log \frac{\pi_\theta(y_w|x, m)}{\pi_{ref}(y_w|x, m)} - \beta \log \frac{\pi_\theta(y_l|x, m)}{\pi_{ref}(y_l|x, m)})\big],\end{aligned} \tag{3}$$

where $\sigma(\cdot)$ is the sigmoid function. The log-probability term $\log \pi(y|x, m) = \sum_{y_i \in y} \log p(y_i|x, m, y_{<i})$ denotes the auto-regressive log-likelihood of the token sequence y, conditioned on the prompt x and preceding tokens $y_{<i}$. Intuitively, DPO encourages the model to assign higher likelihood to the response y_w than to the response y_l under the same context.

2.1.2 Multimodal Direct Preference Optimization

However, text-conditioned preference optimization, where the image input m serves merely as an additional condition for textual data, fails to capture the complexity of multimodal scenarios. To overcome this, recent studies introduce image pairs (m_w, m_l), where m_w denotes the preferred visual input and m_l represents a less preferred image (Fu et al., 2025; Wang et al., 2024a). Given a multimodal input (x, m) and a preferred response y_w, the model is trained to assign higher probability to y_w when conditioned on m_w than on m_l. The training objective is defined as:

$$\mathcal{L}_{\text{MPO}} = -\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w|x,m_w)}{\pi_{\text{ref}}(y_w|x,m_w)} - \beta \log \frac{\pi_\theta(y_w|x,m_l)}{\pi_{\text{ref}}(y_w|x,m_l)}\Big), \tag{4}$$

where m_w is the reference image associated with the preferred response y_w, typically sampled from the dataset, and m_l is often generated by applying perturbations including cropping, rotation, or adding noise to m_w.

2.2 Listwise Preference Optimization

Listwise Preference Optimization (LPO) (Rafailov et al., 2023; Liu et al., 2024c; Zadeh et al., 2025) generalizes pairwise preference learning by optimizing over full rankings of candidate responses rather than individual pairs. The Plackett–Luce (PL) model (Luce et al., 1959), a generalization of the Bradley–Terry model, provides a probabilistic framework for modeling ranked preferences. Under this formulation, a language model is aligned to human preferences by maximizing the likelihood of the observed ranking. Given a prompt x and a ranked list of K candidate responses $y = \{y_1, \ldots, y_K\}$, the reward score for each response is defined as:

$$s_i = \beta \log \frac{\pi_\theta(y_i|x)}{\pi_{\text{ref}}(y_i|x)}, \quad \text{for } i = 1, \ldots, K. \tag{5}$$

Given the scores $s = \{s_1, \ldots, s_K\}$, the listwise loss is defined as the negative log-likelihood of the observed permutation τ under the PL model:

$$\mathcal{L}_{\text{LPO}} = -\log \left(\prod_{k=1}^{K} \frac{\exp(s_{\tau(k)})}{\sum_{j=k}^{K} \exp(s_{\tau(j)})} \right). \tag{6}$$

This objective enables learning from complete rankings, providing richer supervision and promoting globally consistent preference alignment beyond pairwise approaches.

3 Joint Multimodal Preference Optimization

Existing multimodal preference optimization methods effectively align LVLMs by incorporating visual preference supervision. However, these approaches often rely on treating visually corrupted images as rejected inputs, which can induce shortcut learning where the model simply learns to avoid degraded images rather than utilizing meaningful visual information. Moreover, the lack of semantic alignment between corrupted images and their corresponding text further limits the model's ability to learn cross-modal associations.

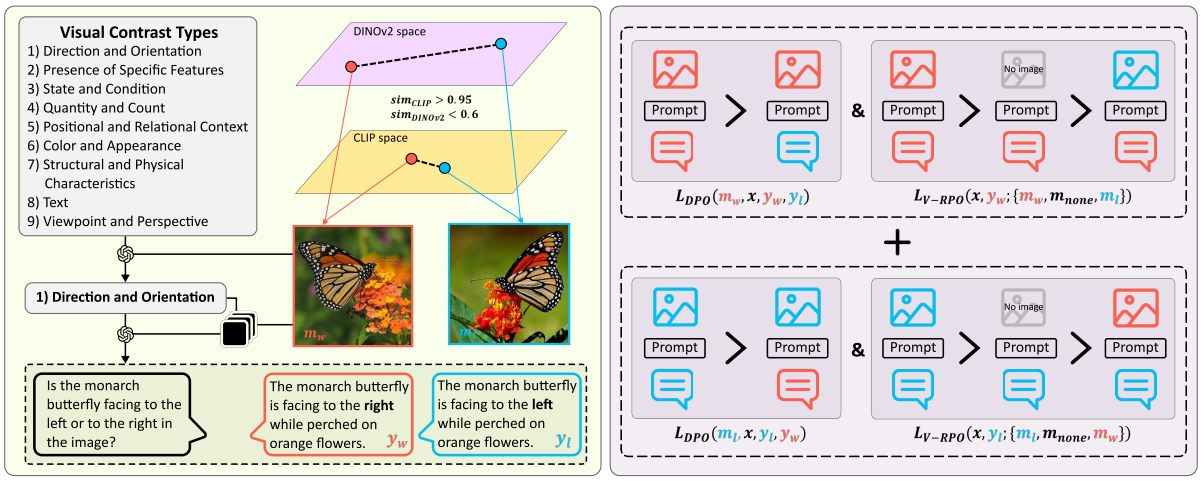

Figure 2: (a) Construction of VTC dataset by selecting image pairs that are similar in CLIP embedding but different in DINOv2 embedding, categorized into one of visual contrast types and annotated with textual responses. (b) Symmetric integration of text-based DPO and visual ranking-based preference optimization (V-RPO).

To address these limitations, JoMPO adopts a symmetric multimodal training objective with two components: (i) a text-based preference loss and (ii) a visual ranking-based objective. This symmetric design enables fine-grained multimodal alignment and encourages the model to attend to subtle visual cues, as illustrated in Figure 2(b).

3.1 Visual Ranking-Based Preference Optimization

To ensure that the model attends to visual content, JoMPO encourages higher rewards when a semantically matched image m_w is provided as a conditioning input compared to when no image is given, prompting the model to ground its responses in relevant visual cues. Conversely, when the same response y_w is conditioned on a contradictory image m_l, the model is trained to minimize the reward. This reduction in reward lowers the likelihood that the model will produce y_w under visually inconsistent conditions.

While conventional pairwise preference optimization captures only binary comparisons, limiting its ability to model complex relational structures, JoMPO leverages a listwise preference loss that jointly considers multiple visual conditions. To model the ordinal structure of visual conditions, we define a ranked list of three visual inputs for a given response: an aligned image m_w, a missing image m_{none}, and a misaligned image m_l. The model is trained to satisfy the preference ordering $s(m_w, x, y_w) > s(m_{none}, x, y_w) > s(m_l, x, y_w)$, where the score function is defined as $s(m, x, y) = \beta \log \frac{\pi_\theta(y|x,m)}{\pi_{\text{ref}}(y|x,m)}$. This score measures the relative preference for generating y under different visual contexts.

To enforce this ordering, we adopt a listwise ranking loss that minimizes the negative log-likelihood of the ground-truth permutation:

$$\mathcal{L}_{\text{V-RPO}}(x, y; \{m_w, m_{\text{none}}, m_l\}) = -\log \left(\prod_{k=1}^{3} \frac{\exp\left(s(m_k, x, y)\right)}{\sum_{j=k}^{3} \exp\left(s(m_j, x, y)\right)} \right), \quad (7)$$

where the visual inputs $\{m_k\}_{k=1}^{3}$ correspond to the predefined conditions $m_1 = m_w$, $m_2 = m_{none}$, $m_3 = m_l$. This ordering represents the target permutation $\tau = [m_w, m_{none}, m_l]$, defining the desired ranking of visual contexts for generating the response y. This listwise loss enables the model to capture interdependencies among ranked visual signals and to develop a more precise understanding of visual details.

3.2 Symmetric Multimodal Objective Integration

Prior approaches typically provide preference supervision only for a single fixed image–response pair (m_w, y_w), designating it as the 'winning' combination and leading to the other combinations being regarded as 'losing' cases. In contrast, JoMPO does not treat any image–text pair as inherently correct, but instead infers preference relations dynamically based on the semantic compatibility between the image and response. For example, when the image m_l is given, the model is trained to prefer y_l over y_w if y_l is semantically compatible with

m_l, guiding the model to better align with semantically coherent image–text pairs while suppressing incoherent ones.

To complement this formulation, JoMPO incorporates the reverse configuration to construct a symmetric training objective that enforces consistency across both directions of preference. The overall objective combines the text-conditioned loss Eq.3, defined as $\mathcal{L}_{DPO}(m, x, y_w, y_l) = -\mathbb{E}_{\mathcal{D}}[\log \sigma(r(x, m, y_w) - r(x, m, y_l))]$, and the visual ranking-based loss Eq.7 in a symmetrical manner:

$$\begin{aligned}\mathcal{L}_{\text{JoMPO}} = \mathcal{L}_{\text{DPO}}(m_w, x, y_w, y_l) \\ + \mathcal{L}_{\text{DPO}}(m_l, x, y_l, y_w) \\ + \mathcal{L}_{\text{V-RPO}}(x, y_w; \{m_w, m_{\text{none}}, m_l\}) \\ + \mathcal{L}_{\text{V-RPO}}(x, y_l; \{m_l, m_{\text{none}}, m_w\}).\end{aligned} \tag{8}$$

This joint multimodal alignment framework ensures that all image and response configurations contribute equally to the learning process, enabling balanced optimization of alignment signals across modalities.

4 Visual-Textual Contrast Dataset

Fine-grained visual–textual alignment requires contrastive supervision that distinguishes subtle visual differences, which is largely missing from existing resources. To address this gap, we propose the Visual-Textual Contrast (VTC) Dataset, which comprises image pairs that are visually similar yet differ in fine-grained visual features, each paired with contrastive textual responses. Unlike prior datasets that contrast only responses (y_w, y_l), VTC also incorporates contrastive image pairs (m_w, m_l), enabling joint optimization of both visual- and response-level preferences. While we employ widely used and well-validated models for image and text generation, the overall pipeline is fully automated and model-agnostic, ensuring scalability and applicability without dependence on specific architectures.

4.1 Construction of Contrastive Image Pairs

Most LVLMs rely on visual encoder pretrained with instance-level contrastive learning, such as CLIP (Radford et al., 2021), which tends to capture coarse-grained visual semantics due to its alignment with high-level captions (Yao et al., 2021; Varma et al., 2023). In contrast, self-supervised vision models such as DINOv2, trained without textual supervision, have demonstrated the ability to capture fine-grained visual information. Inspired by Tong et al. (2024b), we leverage this representational complementarity by selecting image pairs whose cosine similarity exceeds 0.95 in CLIP embeddings space but falls below 0.6 in DINOv2 embedding space, as illustrated in Fig. 2(a). These *CLIP-blind pairs* exhibit high semantic similarity yet contain subtle visual contrasts that current LVLMs struggle to distinguish.

4.2 Categorization of Key Visual Contrast Types

Following the taxonomy proposed by Tong et al. (2024b), we categorize each image pair into one of nine key visual contrast types that are frequently overlooked by CLIP-based encoders. These categories encompass not only basic attributes such as object count and color but also subtle variations in orientation, state, and contextual configuration, capturing a broad spectrum of fine-grained visual differences essential for multimodal understanding. Examples of each type are illustrated in Figure 2(a), and detailed descriptions are provided in Appendix D.

To automate this categorization, we employ a vision–language model to assign the primary visual contrast type to each image pair, utilizing gpt-4o (Hurst et al., 2024), one of the most widely validated and high-performing models to date (Sun et al.; Zhao et al., 2023). For every category, the model estimates two scores: (1) *Importance*, indicating the relevance of the contrast type, and (2) *Hallucination Likelihood*, assessing the likelihood that an LVLM may misinterpret the contrast. Supporting rationales are generated for each score (Wei et al., 2022), and the category with the highest overall score is selected as the representative type. All selections undergo lightweight human verification to ensure annotation integrity and consistency.

4.3 Generation of Contrastive Text Data

Based on the identified visual contrast type, we generate a prompt–responses pair for each image using GPT-4o. Manually curated exemplars are provided as in-context demonstrations to maintain stylistic and structural consistency. The linguistic style remains uniform, with variations introduced only in segments reflecting specific visual differences. This design enables the model to focus on hallucination-prone regions and better capture fine-grained visual distinctions.

Algorithm	Data size	Object HalBench		AMBER						MMHal-Bench	
				Discriminative		Generative					
		CHAIRs↓	CHAIRi↓	Acc↑	F1↑	CHAIR↓	Cover↑	HalRate↓	Cog↓	Score↑	HalRate↓
LLaVA-1.5-7B		55.7	15.9	73.5	77.7	7.7	51.6	34.7	4.2	2.01	0.61
+LLaVA-RLHF	122k	55.4	27.3	68.7	74.7	9.7	53.2	46.6	5.3	1.88	0.71
+HA-DPO	6k	54.0	14.4	75.2	79.9	7.8	52.1	35.6	4.2	1.89	0.65
+POVID	17k	35.9	17.3	78.6	81.9	7.4	51.3	34.3	3.9	2.28	0.56
+HALVA	21.5k	41.4	11.7	78.0	83.5	6.6	53.0	32.2	3.4	2.25	0.54
+mDPO	10k	41.4	11.7	73.4	74.7	4.4	52.4	24.5	2.4	2.39	0.54
+RLAIF-V	16k	16.0	**3.70**	76.8	84.5	3.0	50.4	16.2	1.0	**3.00**	**0.38**
+OPA-DPO	4.8k	13.0	4.25	80.3	85.6	**2.8**	45.7	**12.9**	1.2	2.62	0.50
+DPO	5k	28.3	14.6	81.4	85.9	5.3	**59.0**	26.8	1.5	2.42	0.52
+MPO	5k	26.0	13.6	81.4	86.0	4.6	57.6	23.0	1.3	2.43	0.50
+JoMPO	5k	**12.3**	6.90	**82.8**	**87.3**	4.1	58.9	20.8	**0.9**	2.77	0.43
LLaVA-1.5-13B		51.0	13.7	71.4	73.2	6.8	51.9	31.8	3.3	2.48	0.52
+LLaVA-RLHF	122k	44.7	11.8	79.7	83.9	7.7	52.3	38.6	4.0	2.27	0.64
+HALVA	21.5k	45.4	12.8	82.9	86.5	6.4	52.6	30.4	3.2	2.58	0.45
+OPA-DPO	4.8k	16.3	5.50	84.1	87.5	**2.5**	48.3	**12.8**	**0.9**	2.85	0.40
+DPO	5k	36.0	19.1	84.7	87.9	5.3	58.4	27.8	2.0	2.59	0.47
+MPO	5k	30.3	16.6	85.0	88.2	4.6	56.7	23.4	1.5	2.64	0.48
+JoMPO	5k	**13.0**	**6.60**	**86.1**	**89.4**	4.5	**61.7**	24.1	**0.9**	**2.88**	**0.39**

Table 1: Main results of hallucination evaluation for LLaVA-1.5-7B and LLaVA-1.5-13B models trained with JoMPO. For each metric, the best performance is marked in **bold**, and the second-best is underlined. Rows shaded in green indicate models trained on the VTC dataset. DPO refers to standard text-only preference optimization, while MPO denotes multimodal preference optimization using 20% randomly cropped images as rejection examples.

To further enhance linguistic and functional diversity (Zhou et al., 2024), we construct two types of contrastive text data: (1) *Detailed Description*, in which prompts elicit comprehensive descriptions of each image, and responses explicitly convey both the primary visual contrast and relevant supporting details; (2) *Contrastive VQA*, comprising natural question–answer exchanges centered on the identified contrast type, suitable for modeling interactive, dialog-oriented scenarios. Examples of each task are shown in Figure 2 and discussed in Appendix E. These paired samples provide complementary supervision for learning discriminative visual grounding and achieving robust multimodal alignment.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets

To construct the VTC dataset, we utilized image sources from MSCOCO (Lin et al., 2014), ImageNet (Deng et al., 2009), CC3M (Sharma et al., 2018), LAION-Art (Schuhmann et al., 2022), and OpenImages v4 (Kuznetsova et al., 2020), computing CLIP and DINO embedding similarities across approximately 17 million images. To ensure computational efficiency, a batch size of 2048 was used, and pairwise similarities were calculated within each batch. Based on these similarity scores, we selected approximately 0.6k, 0.4k, 0.7k, 1.9k, and 1.4k images from each respective dataset, resulting in a curated set of about 5k contrastive pairs.

5.1.2 Evaluation

We evaluate our models on three representative hallucination benchmarks. **AMBER** (Wang et al., 2024b) provides a generative evaluation setting with 1,004 images, measuring CHAIR score, object coverage, hallucination rate, and human alignment. It also includes a discriminative component formulated as a binary classification task over more than 10,000 images, where accuracy indicates the model's ability to detect hallucinated responses. **Object HalBench** (Rohrbach et al., 2018) assesses object-level hallucinations in detailed image descriptions. Following prior work, we use eight diverse prompts per image to enhance evaluation stability and report both response-level and mention-level hallucination rates. **MMHal-Bench** (Sun et al., 2023) consists of 96 images across 12 object categories. Each model-generated response is rated by GPT-4 on a 0–6 scale, and those scoring below 3 are regarded as hallucinated; the hallucination rate is computed as the proportion of such low-scoring responses.

5.1.3 Baselines

We primarily compare JoMPO with a range of existing approaches rooted in either the RLHF/RLAIF or DPO paradigms. Most previous methods such as HA-DPO (Zhao et al., 2023), POVID (Zhou et al.,

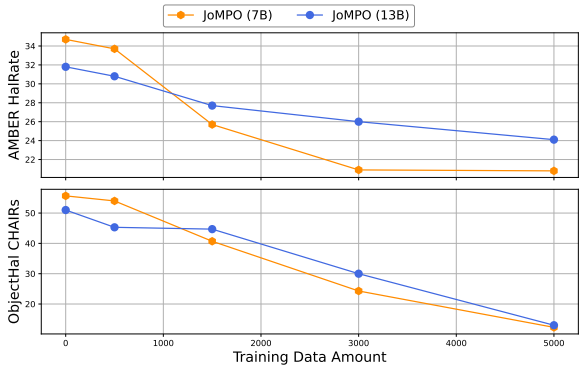

Figure 3: Performance of the JoMPO framework on LLaVA models (7B and 13B) evaluated on the AMBER and Object HalBench with varying training set sizes.

	AMBER (Gen.)		MMHalBench	
	CHAIR↓	HalRate↓	Score↑	HalRate↓
JoMPO	**4.1**	**20.8**	**2.77**	**0.43**
w/o ranking	4.8	24	2.61	0.47
w/o sym	4.6	21.3	2.55	0.49
w/o ranking, sym	5.1	25.5	2.36	0.53
w/o ranking, text	4.9	24.3	2.55	0.49
w/o ranking, no-image	4.7	23.7	2.52	0.49

Table 2: Ablation study of the JoMPO components using the LLaVA-1.5-7B model.

2024), HALVA (Sarkar et al., 2024), mDPO (Wang et al., 2024a), RLAIF-V (Yu et al., 2024b), and OPA-DPO (Yang et al., 2025) adopt the DPO framework, leveraging contrastive preference signals to align model behavior with human or synthetic feedback. In contrast, LLaVA-RLHF (Sun et al., 2023) employs PPO as its optimization strategy, representing a reinforcement learning-based alternative.

5.1.4 Implementation Details

We evaluate the proposed JoMPO framework using two Large Vision-Language Models, LLaVA v1.5 7B and LLaVA v1.5 13B. Both models adopt CLIP ViT-L/336px as the vision encoder and are built upon the Vicuna 7B and Vicuna 13B backbones (Chiang et al., 2023), respectively. We focus our main experiments on the LLaVA-1.5 family to ensure fair and methodologically consistent comparison with prior multimodal preference optimization methods developed under the same backbone. To assess the generalization of JoMPO beyond LLaVA-1.5, we additionally report results on a more recent LVLM in Appendix B. We first performed Supervised Fine-tuning (SFT) for 2 epochs, followed by 2 epochs of training with the proposed JoMPO framework. For all experiments, we set the hyperparameter $\beta = 0.05$, the learning rate to 5e-7, and the batch size to 4. Training was conducted on 8 A100 GPUs and completed within 2 hours. Details of the training setup and additional experiments are discussed in Appendix C.

5.2 Experimental Results and Observations

Table 1 presents a comprehensive evaluation of hallucination reduction across multiple benchmarks, comparing the proposed JoMPO method with prior preference optimization strategies. The results highlight three key findings: (1) JoMPO consistently demonstrates superior hallucination reduction performance, achieving 77.9% and 74.5% relative reductions on Object HalBench for the 7B and 13B variants, respectively. The strong gains on Object HalBench highlight JoMPO's effectiveness in mitigating *object-level hallucinations* through improved visual grounding. Furthermore, its superior performance on the *Discriminative* subset of AMBER indicates enhanced capability in both open-ended generation and discriminative reasoning, demonstrating an enhanced ability to answer binary (yes/no) questions grounded in visual evidence. (2) JoMPO outperforms both text-only preference optimization (DPO) and its multimodal extension (MPO) on nearly all evaluation metrics. This consistent advantage suggests that incorporating semantically contrastive image–text pairs into the training objective enables the model to capture more granular visual details. (3) To investigate the impact of training data size, we evaluate JoMPO under varying amounts of preference supervision, as shown in Figure 3. Both the 7B and 13B models demonstrate a clear downward trend in hallucination rates on the AMBER and Object HalBench as the training set grows. Notably, JoMPO performs strongly even with limited data. Notably, the 7B model variant often surpasses the 13B model across hallucination benchmarks. This suggests that smaller models may benefit more from preference optimization, as their lower capacity allows stronger and more targeted alignment signals, whereas larger models can become over-regularized or less responsive to limited supervision. One possible explanation is that larger models may require more diverse preference signals or different regularization schedules to effectively incorporate alignment updates when supervision is limited.

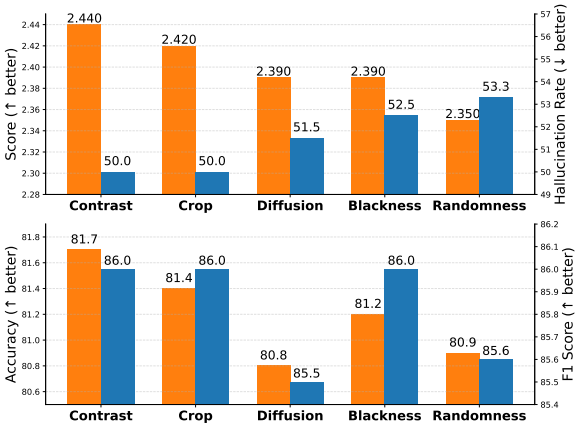

Figure 4: Performance comparison of rejected image construction strategies on MMHal-Bench (Score↑, Hallucination Rate↓) and AMBER (Accuracy↑, F1 Score↑) benchmarks.

5.3 Ablation Studies

To assess the contribution of each component in JoMPO, we perform ablation experiments using LLaVA-1.5-7B as the backbone, with results reported on selected metrics from the AMBER generative task and MMHal-Bench (Table 2). This analysis shows that incorporating both the visual ranking objective and the symmetric loss structure is essential for effective hallucination reduction. The model achieves consistently stronger performance when both components are included, indicating that listwise comparison of images and bidirectional preference supervision jointly contribute to more precise multimodal alignment. Variants without ranking and text-based supervision suggest that image-only optimization may not sufficiently capture the semantic grounding needed for hallucination mitigation, highlighting the importance of integrating both visual and textual signals. Furthermore, removing both ranking and the no-image condition emphasizes the complementary role of contrastive supervision across modalities, as the full JoMPO setup more robustly distinguishes relevant visual cues, leading to improved image–text alignment.

5.4 Impact of Contrastive Visual Pair Construction

To validate the effectiveness of using *Contrast* images that are globally similar to the chosen image but contain localized semantic differences, we conduct a comparative experiment. As shown in Fig. 4, the proposed Contrast method consistently outperforms the others across metrics. We consider four

	General				Vision-Centric	
	TextVQA Acc.	ScienceQA Acc.	LLaVA Score	MMVet Score	CVBench Acc.	MMVP Acc.
LLaVA	62.9	68.5	69.2	31.5	61.6	59.3
+ JoMPO	**64.2**	**63.7**	**72.5**	**33.3**	**63.0**	**62.3**

Table 3: Evaluation results on general-purpose and vision-centric benchmarks.

baseline strategies for constructing rejected images: (1) *Diffusion*: perturbs the original image by adding Gaussian noise over approximately 500 steps following a forward diffusion process. (2) *Randomness*: selects an image randomly from the training set. (3) *Blackness*: removes all visual content by setting the RGB values of the image to zero. (4) *Crop*: randomly cropped from the original image. All methods are evaluated under the same multimodal preference optimization framework without JoMPO. Among them, *Crop*, which preserves most of the original visual content, achieves relatively better performance, suggesting that losing visual information weakens supervision signals.

5.5 General and Visual Capability Evaluation

To examine whether JoMPO introduces any trade-offs between hallucination mitigation and general capability (i.e., an alignment tax), we evaluate its performance on a diverse set of general-purpose and vision-centric benchmarks. Specifically, we assess open-ended reasoning and language-centric visual question answering using LLaVA-Bench, MMVet (Yu et al., 2023), TextVQA (Singh et al., 2019), and ScienceQA (Lu et al., 2022), and evaluate visual grounding and perception using CVBench (Tong et al., 2024a) and MMVP (Tong et al., 2024b) (Table 3).

Overall, the results show that JoMPO preserves general capabilities while improving performance on three of the four general benchmarks. The only exception is ScienceQA, which contains a substantial portion of non-visual and knowledge-centric questions. This behavior is consistent with the design of JoMPO, which emphasizes fine-grained visual distinctions through structured multimodal preference optimization, making its benefits more pronounced on vision-grounded tasks. Consequently, JoMPO improves or maintains performance on benchmarks that require visual grounding, without introducing a systematic degradation in general capabilities.

6 Conclusion

In this work, we introduced Joint Multimodal Preference Optimization (JoMPO), a novel alignment framework that symmetrically integrates textual and visual supervision to mitigate hallucinations in Large Vision-Language Models (LVLMs). Unlike prior approaches that rely on synthetically degraded images and asymmetric preference signals, JoMPO employs bidirectional preference optimization and listwise visual ranking over semantically contrastive image–text pairs, enabling the model to learn fine-grained multimodal associations. To support this framework, we curated the Visual–Textual Contrast (VTC) Dataset, comprising visually similar but subtly different image pairs annotated with textual responses. This dataset provides rich multimodal supervision, facilitating precise visual grounding. Experiments on standard hallucination benchmarks and selected general multimodal tasks indicate that JoMPO reduces hallucination rates and improves visual alignment while preserving general capability, with gains observed using 5k contrastive pairs. Overall, JoMPO provides a data-efficient approach for aligning LVLMs, supporting fine-grained visual grounding and multimodal alignment.

Limitations

While this work establishes a unified multimodal preference optimization framework and demonstrates consistent improvements across benchmarks, several limitations remain.

(1) Reliance on Pretrained Models for Dataset Construction. The construction of the proposed Visual–Textual Contrast (VTC) dataset relies on pretrained models such as CLIP, DINOv2, and GPT-4o for feature extraction and annotation assistance. While this design enables largely automated data construction and ensures reproducibility, it may also partially reflect representational biases or pretraining distributions inherent to these models. Although the data construction pipeline itself is designed to be architecture-agnostic, in this work we adopt widely validated and publicly available pretrained models to ensure experimental stability and reproducibility.

(2) Scope of Evaluation and Generalization. While JoMPO was evaluated on diverse hallucination and general multimodal benchmarks, its effectiveness was primarily tested on visual–textual datasets in the English domain. Extending the framework to multilingual, domain-specific, or non-visual modalities (e.g., audio or video) would further validate its generality. Moreover, although the dataset construction and training process are fully automated, lightweight human verification was employed to ensure annotation integrity, which may still introduce minor subjectivity.

References

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

Heejong Bong and Alessandro Rinaldo. 2022. Generalized results for the existence and consistency of the mle in the bradley-terry-luce model. In *International Conference on Machine Learning*, pages 2160–2177. PMLR.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Jinlan Fu, Shenzhen Huangfu, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. 2025. Chip: Cross-modal hierarchical direct preference optimization for multimodal llms. *arXiv preprint arXiv:2501.16629*.

Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille. 2022. Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5078–5088.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Modality-fair preference optimization for trustworthy mllm alignment. *arXiv preprint arXiv:2410.15334*.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov,

and 1 others. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981.

Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. 2024. A survey of hallucination in large visual language models. *arXiv preprint arXiv:2410.15359*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and 1 others. 2023. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, and 1 others. 2024c. Lipo: Listwise preference optimization through learning-to-rank. *arXiv preprint arXiv:2402.01878*.

Jinda Lu, Jinghan Li, Yuan Gao, Junkang Wu, Jiancan Wu, Xiang Wang, and Xiangnan He. 2025. Adavip: Aligning multi-modal llms via adaptive vision-enhanced preference optimization. *arXiv preprint arXiv:2504.15619*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

R Duncan Luce and 1 others. 1959. *Individual choice behavior*, volume 4. Wiley New York.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arik, and Tomas Pfister. 2024. Mitigating object hallucination via data augmented contrastive tuning. *CoRR*.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and 1 others. Parrot: Multilingual visual instruction tuning. In *Forty-second International Conference on Machine Learning*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024a. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024b. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

Maya Varma, Jean-Benoit Delbrouck, Sarah Hooper, Akshay Chaudhari, and Curtis Langlotz. 2023. Villa: Fine-grained vision-language representation learning from real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22225–22235.

Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2024a. mdpo: Conditional preference optimization for multimodal large language models. *arXiv preprint arXiv:2406.11839*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024b. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024c. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. 2025. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30284–30297, Vienna, Austria. Association for Computational Linguistics.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025. 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*, pages 10610–10620.

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and 1 others. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv e-prints*, pages arXiv–2405.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Fatemeh Pesaran Zadeh, Yoojin Oh, and Gunhee Kim. 2025. Lpoi: Listwise preference optimization for vision language models. *arXiv preprint arXiv:2505.21061*.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

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

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Jingbo Zhu. 2023. Overcoming language priors with counterfactual inference for visual question answering. *Chinese Computational Linguistics*, page 58.

A Related works

A.1 Multimodal Hallucination in LVLMs

Multimodal hallucination refers to the phenomenon where a model generates text that is inconsistent with the given visual input, and is often referred to as object hallucination (Li et al., 2023; Bai et al., 2024; Zhou et al., 2023). Prior studies categorize object hallucinations into three major types:

- **Object Existence Hallucination**: mentioning objects that are not present in the image.
- **Object Attribute Hallucination**: describing an existing object with incorrect attributes such as color, size, direction, or texture.
- **Object Relation Hallucination**: misrepresenting the spatial position or relationship of an object within the scene.

These hallucinations often arise when LVLMs rely excessively on linguistic priors while failing to sufficiently attend to the visual input (Gupta et al., 2022; Zhu, 2023). Common failure cases include describing non-existent animals in the background, misreporting the color of a fruit, or reversing the spatial relation between entities in a scene. Root causes include weak cross-modal alignment (Sun et al., 2023), coarse-grained vision-language pretraining (Yao et al., 2021), and a lack of supervision on fine-grained visual distinctions. Surveys by (Bai et al., 2024; Liu et al., 2024a) provide comprehensive taxonomies of hallucination phenomena, while benchmarks such as Object HalBench (Rohrbach et al., 2018) and AMBER (Wang et al., 2024b) have been introduced to evaluate hallucination severity and model reliability.

To mitigate these issues, prior work has explored architectural modifications and additional supervision strategies during pretraining (Tong et al., 2024a; Varma et al., 2023), as well as fine-tuning methods that enhance the model's sensitivity to visual details (Sun et al., 2023; Sarkar et al., 2024). However, many of these approaches lack supervision over subtle visual contrasts, which are essential for ensuring visually faithful outputs. In this work, we address this gap by explicitly leveraging semantically contrastive yet visually similar image–text pairs, facilitating more precise visual grounding and multimodal alignment.

Model	AMBER (Gen.)	MMHal
	CHAIR↓ / Cover↑ / Hal↓ / Cog↓	Score↑ / Hal↓
Qwen2-VL-7B	6.8 / 72.1 / 54.5 / 5.1	3.26 / 0.33
+ JoMPO	**4.3** / 62.0 / **23.5** / **1.2**	**3.43** / **0.31**

Table 4: Hallucination evaluation results on Qwen2-VL-7B-Instruct.

A.2 Preference Optimization for Hallucination Mitigation

Inspired by the success of Direct Preference Optimization (DPO) in aligning LLMs with human preferences (Rafailov et al., 2023), recent studies have extended DPO to the multimodal domain (Yu et al., 2024a; Xiao et al., 2025). Standard DPO relies on pairwise comparisons between preferred and dispreferred responses under shared visual prompts. While effective in textual domains, this formulation is limited in multimodal settings due to its asymmetry in visual supervision.

To address this, multimodal extensions such as mDPO (Wang et al., 2024a) and CHiP (Fu et al., 2025) incorporate visual preference by perturbing images (e.g., via cropping or noise injection) to serve as negative examples. Other works, like RLAIF-V (Yu et al., 2024b) and OPA-DPO (Yang et al., 2025), explore synthetic AI feedback or on-policy data to improve reward quality. However, these methods typically generate overly explicit positive–negative image pairs, which risk shortcut learning and fail to capture nuanced visual distinctions.

In contrast, our proposed JoMPO framework symmetrically integrates text-based DPO with a visual ranking-based preference objective. Specifically, it constructs semantically contrastive yet visually similar image–text pairs, and jointly trains the model using a visual ranking-based loss alongside the standard text-side preference loss in a symmetric fashion. This enables the model to capture fine-grained semantic distinctions. By addressing the visual supervision asymmetry present in prior approaches, JoMPO demonstrates strong generalization performance across various hallucination benchmarks.

B Additional Experiments

Most prior multimodal DPO-based hallucination mitigation studies adopt LLaVA-1.5 as the primary backbone for both training and evaluation. To ensure a fair and controlled comparison under identical preference optimization protocols, our main experiments are therefore conducted on LLaVA-1.5 models. We report results on both the 7B and 13B variants to demonstrate the robustness of JoMPO across different parameter scales within the same model family.

We additionally apply JoMPO to a more recent large vision-language model, Qwen2-VL-7B-Instruct, to assess the generalization of the proposed framework beyond the LLaVA-1.5 backbone. As shown in Table 4, JoMPO consistently reduces hallucination-related metrics on both fine-grained and generative benchmarks, including substantial improvements in generative hallucination rate and cognitive hallucination scores on AMBER, as well as improved performance on MMHal-Bench.

We note that most existing multimodal preference optimization baselines were originally developed and evaluated on LLaVA-1.5, and thus direct numerical comparison on Qwen2-VL would not be methodologically aligned. Nevertheless, the consistent improvements over the base Qwen2-VL model demonstrate that JoMPO is not tied to a specific backbone, but instead generalizes effectively to stronger and more recent LVLM architectures. We leave a broader evaluation on additional state-of-the-art LVLMs as future work to further strengthen the empirical validation of JoMPO.

C Experimental Details

C.1 Training Setup

We trained both LLaVA-1.5 7B and 13B models on the VTC-5k dataset using 8 NVIDIA A100 GPUs. During the supervised fine-tuning (SFT) phase, we employed a per-device batch size of 4 and gradient accumulation steps of 4. The model was trained for 2 epochs using bfloat16 precision and FlashAttention, with LoRA-based parameter-efficient tuning (rank 256, alpha 512) applied to the base model, vision tower, and multimodal projector. For the subsequent JoMPO fine-tuning, we used a per-device batch size of 4 with gradient accumulation steps of 8, again using bfloat16 precision and FlashAttention. Optimization was performed using a cosine learning rate scheduler with a peak learning rate of 5×10^{-7} and a warmup ratio of 0.03. The temperature scaling coefficient β was set to 0.05 throughout JoMPO training.

C.2 Hyperparameter Experiments

We conduct a hyperparameter study on the scaling coefficient β and training epochs using the *LLaVA-1.5-7B* model with JoMPO. As shown in Table 5, performance remains stable across settings, with minor fluctuations in both discriminative and generative metrics. The configuration of $\beta = 0.05$ and two epochs yields the most balanced results, achieving the lowest hallucination rates.

Model	epoch	AMBER Disc.		AMBER Generative			
		Acc.	F1	CHAIR↓	Cover.	Hal↓	Cog.↓
β=0.05	epoch2	82.8	87.3	4.1	58.9	20.8	0.9
	epoch3	82.9	87.3	4.3	59.2	22.4	1.1
β=0.1	epoch2	83.2	88.5	4.5	60.1	23.6	1.1
	epoch3	82.6	87.2	4.5	60.2	23.6	1.2
β=0.3	epoch2	82.1	86.9	4.6	60.4	23.8	1.1
	epoch3	82.1	87.0	4.7	60.3	24.0	1.2
β=0.5	epoch2	81.8	86.7	4.4	60.2	23.0	1.2
	epoch3	82.0	86.8	4.4	60.1	22.9	1.2

Table 5: Hyperparameter experiments of β and epoch on **LLaVA-1.5-7B + JoMPO**.

C.3 Evaluation Protocol

To ensure deterministic inference, we set `do_sample` to `False` and `temperature` to 0.0 across all evaluations. This guarantees consistent outputs for identical inputs, allowing for reliable and reproducible comparisons. These settings were used for all hallucination benchmarks in Tab. 1, including AMBER, ObjectHal Bench, and MMHal-Bench, as well as for the additional benchmarks reported in Tab. 3. In the data scaling experiments shown in Fig. 3, we fixed the random seed to 1234 and varied only the number of training samples to isolate the effect of dataset size. Most baseline results in Tab. 1 were sourced from the original (Yang et al., 2025) paper, and we conducted additional re-evaluations using the same deterministic inference settings to ensure comparability.

D Dataset Details

D.1 Visual Contrast Types

To ensure diversity and fine-grained supervision in visual-textual alignment, we adopt the visual contrast taxonomy proposed in (Tong et al., 2024b) and apply it systematically during the construction of the Visual–Textual Contrast (VTC) dataset. Specifically, every image pair in VTC is associated with a contrast instance grounded in one of nine predefined visual dimensions:

- **Orientation and Direction:** Refers to differences in the facing or movement direction of subjects, such as left versus right orientation of people, animals, or vehicles.
- **Presence of Specific Features:** Captures whether specific elements or attributes are present or absent in the image.
- **State and Condition:** Involves dynamic or situational properties, such as whether a flag is waving or the ground appears wet.
- **Quantity and Count:** Focuses on numerical distinctions, including the number of visible objects or entities.
- **Positional and Relational Context:** Emphasizes spatial relationships among elements, such as above/below or near/far positioning.
- **Color and Appearance:** Addresses perceptual qualities like color, shading, or surface texture.
- **Structural and Physical Characteristics:** Concerns the shape, size, material, or structural integrity of objects.
- **Text and Symbols:** Relates to the presence and recognition of textual content or signage within the image.
- **Viewpoint and Perspective:** Refers to the camera angle or framing, such as top-down views versus side views.

For each image pair, a prompt–response pair is generated such that the contrast between the two images reflects a semantic difference along a single designated dimension. This ensures that each instance in the dataset provides focused supervision aligned with a specific type of visual reasoning. By covering all nine dimensions across the dataset, VTC enables models to learn to discriminate subtle visual cues and align them precisely with corresponding textual semantics.

D.2 Distribution of Visual Contrast Types in VTC

To evaluate the diversity and balance of visual reasoning dimensions in the VTC dataset, we analyze the distribution of the nine predefined contrast types across all 5,025 examples. Each instance in

Visual Contrast Type	# of Instances
Orientation and Direction	195
Presence of Specific Features	1,424
State and Condition	91
Quantity and Count	318
Positional and Relational Context	282
Color and Appearance	1,745
Structural and Physical Characteristics	511
Text	131
Viewpoint and Perspective	328
Total	5,025

Table 6: Distribution of visual contrast types in VTC.

VTC is explicitly annotated with one of these visual contrast dimensions. The overall distribution is intentionally designed to reflect a broad spectrum of visual reasoning capabilities. Categories such as *Presence of Specific Features* and *Color and Appearance* appear more frequently, as they are commonly encountered in real-world visual understanding tasks. This dimension-aware design promotes robust training across diverse visual phenomena and enables fine-grained analysis of model performance by contrast type. A detailed breakdown of the distribution is provided in Tab. 6.

LLaVA-1.5-7B

The image depicts a man riding a horse in a vast, open field with mountains in the background. The man is positioned towards the left side of the scene, while the horse is located more towards the center. The field appears to be a mix of grass and dirt, creating a picturesque landscape. In addition to the man and the horse, there are two other people visible in the scene, one near the center and the other towards the right side. A backpack can also be seen on the ground, likely belonging to one of the people in the field.

JoMPO

The image depicts a vast, open plain with a person riding a horse in the foreground. In the distance, majestic snow-capped mountains rise, creating a dramatic backdrop. The landscape is a mix of smooth, dry grass and rocky terrain, giving a sense of openness. The sky is partly cloudy, adding depth to the scene, and the overall mood is serene and expansive.

Figure 5: Qualitative results with hallucinated content are marked in red, and missing details are marked in blue.

E Qualitative Results

To qualitatively assess the impact of our proposed JoMPO framework, we present example outputs comparing the baseline model (LLaVA-1.5-7B) and our JoMPO-aligned model across a subset of images from the AMBER benchmark. For consistency and to isolate the effect of visual grounding, we standardize the prompt to a single instruction across all examples: *"Describe the image."* Figures 5–7 illustrate the model responses under identical visual and textual inputs. Each figure contains the original image, followed by the responses from LLaVA-1.5-7B and the JoMPO-aligned model. Overall, the JoMPO model consistently produces more visually grounded descriptions. It demonstrates greater sensitivity to fine-grained visual cues, while reducing hallucinated content that is not visually evident. In contrast, the baseline LLaVA-1.5-7B model tends to rely more heavily on prior knowledge, often generating plausible but ungrounded or speculative details. These examples complement our quantitative findings and highlight the effectiveness of joint multimodal preference optimization in aligning model outputs with visual inputs.

LLaVA-1.5-7B

The image depicts a group of people enjoying themselves in a large body of water, likely a hot spring or a swimming pool. There are five people in the scene, with three of them sitting in the water and two others standing nearby. They are all smiling and seem to be having a great time together. In the background, there are several benches placed around the area, providing a place for people to sit and relax. Some of these benches are closer to the water, while others are situated further away. The overall atmosphere appears to be one of leisure and enjoyment as the group of friends spends time together in the water.

JoMPO

The image depicts a group of four people in a natural hot spring, surrounded by smooth, clear blue water. They are standing and sitting in the water, enjoying the warmth. The sky is partly cloudy, and steam rises from nearby geothermal vents, creating a serene and inviting atmosphere.

Figure 6: Qualitative results.

LLaVA-1.5-7B

In the image, a person is holding a small piglet in their arms. The piglet appears to be sleeping or resting comfortably in the person's embrace. There are two other people in the scene, one standing to the left of the person holding the piglet and another person on the right side. A chair can be seen in the background, and a handbag is placed on the floor near the right side of the scene. The overall atmosphere of the image suggests a warm and caring interaction between the person holding the piglet and the piglet itself.

JoMPO

The image shows a close-up view of a person holding a small piglet in their arms. The piglet is resting its head on the person's arm, and its eyes are visible. In the background, another person is partially visible, wearing a light-colored shirt. The focus is on the interaction between the person holding the piglet and the viewer, creating a warm and intimate atmosphere.

Figure 7: Qualitative results.

F License and Terms of Use

All artifacts created and used are handled in accordance with their respective licenses and terms of use. The VTC dataset and its accompanying resources are released under the Creative Commons Attribution–NonCommercial 4.0 International (CC BY-NC 4.0) license, permitting sharing and adaptation for non-commercial research and educational purposes with appropriate attribution. Publicly available datasets, including MSCOCO, ImageNet, CC3M, LAION-Art, and OpenImages, are used solely for research purposes and in compliance with their original licenses. The released dataset is intended exclusively for non-commercial academic research, and any use beyond this scope should adhere to the licenses of the original data sources and applicable regulations. All external datasets and pretrained models are used consistently with their intended research purposes. The dataset does not intentionally include personally identifiable information. Images are sourced from publicly available datasets, and textual annotations are automatically generated without targeting sensitive attributes. Documentation of the dataset construction, contrast types, and intended usage is provided to ensure transparency and reproducibility.