From Squeezing to Grounding: Visual Guided DPO for Multimodal Hallucination Mitigation

Wenqi Liu[1*] **Hongxin He**[1*] **Yunxiao Wang**[1]
Xuemeng Song[2] **Yupeng Hu**[1] **Yinwei Wei**[1†]
[1]Shandong University [2]Southern University of Science and Technology
liuwq_bit@outlook.com, weiyinwei@hotmail.com

Abstract

Direct Preference Optimization (DPO) is widely used to reduce hallucination in Multimodal Large Language Models (MLLMs). Prior studies and our observations show that during DPO training, the log probabilities of chosen and rejected responses can decrease simultaneously, revealing the *squeezing effect*. For MLLMs, the redistributed probability mass may move toward responses driven by language priors rather than visual evidence, weakening the role of the image and amplifying hallucination. Existing squeezing-aware methods constrain rejected updates with $\mathcal{V}$-usable information or generation confidence, while methods with anchor terms regularize the DPO implicit reward of chosen responses. Neither explicitly grounds the control signal in visual evidence. We propose Visual Guided Direct Preference Optimization (VGDPO), which relaxes negative updates on rejected responses according to multimodal squeezing risk. VGDPO estimates this risk by combining token-level visual dependency with phrase-level hallucination localization, defining a hallucination ratio for each rejected response. We further apply a related reweighting principle to a visual contrastive objective, providing additional supervision for chosen responses under the original image. Experiments on four hallucination benchmarks and five MLLM backbones show that VGDPO improves DPO log probability dynamics and reduces multimodal hallucination while maintaining informative responses. Code is available at Link.

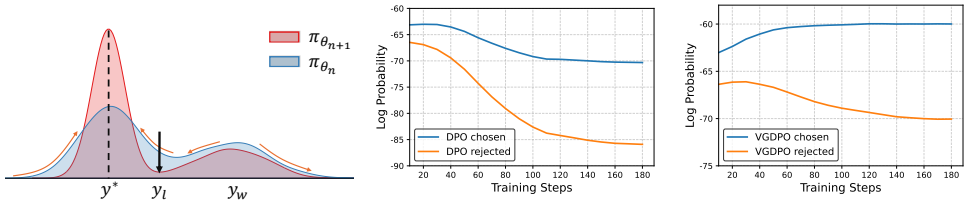
1 Introduction

Direct Preference Optimization (DPO) [1] has become a standard preference optimization method for aligning large language models without training an explicit reward model. By increasing the implicit reward gap between chosen and rejected responses, DPO provides a simple way to learn from preference pairs. It has also been widely adopted for Multimodal Large Language Models (MLLMs), where faithful responses are treated as chosen responses and hallucinated or weakly grounded responses are treated as rejected responses [2, 3, 4, 5, 6]. These multimodal DPO methods reduce content unsupported by the image and preserve accurate responses that follow instructions.

However, DPO suffers from preference likelihood decay due to an undesirable training dynamic known as the *squeezing effect* [7]. As illustrated in Figure 1(a), when a large negative gradient is applied to a low confidence rejected response, a softmax model can suppress a broad output region and move probability mass toward existing high confidence regions, which can also reduce the likelihood of the chosen response. This risk is amplified by off-policy preference data [8, 9, 4]. In DPO, fixed preference pairs are usually generated beforehand by another policy or an earlier checkpoint, while the model being optimized is the current policy. As their mismatch grows during training, both chosen

[*]Equal Contribution.
[†]Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Probability redistribution (b) DPO dynamics (c) VGDPO dynamics

Figure 1: Illustration of the *squeezing effect* and the motivation of VGDPO. (a) The *squeezing effect* redistributes probability mass away from both chosen (y_w) and rejected (y_l) responses toward high-confidence regions (y^*). (b) During DPO training, both log probabilities can decrease. (c) VGDPO increases the chosen log probability and decreases the rejected one.

and rejected responses can become relatively low-confidence under the current policy. Applying negative gradients to the rejected side in this case is more likely to trigger squeezing, redirecting probability mass toward high-confidence outputs already favored by the current policy [7, 10, 11].

In multimodal DPO, we observe the same phenomenon: the log probabilities of both chosen and rejected responses decrease simultaneously, as shown in Figure 1(b). The problem is more complex for MLLMs. On top of off-policy mismatch, they are influenced by both visual evidence and language priors [12, 13, 14, 15]. Since most MLLMs are built from a pretrained language model and then aligned with visual inputs [16], high-confidence outputs can be shaped by the language backbone. The suppressed probability mass may therefore move not to visually grounded outputs, but to responses favored by language priors. For example, given an image containing a red banana, an MLLM may still describe it as yellow or green because these colors are more associated with bananas in text. In this case, aggressively suppressing a rejected response can push the model away from low-confidence visual evidence and toward high-confidence language priors. As a result, DPO may reduce the absolute likelihood of visually grounded chosen responses while amplifying hallucinations caused by mismatches between generated content and image evidence.

Several recent studies alleviate the *squeezing effect* from a general preference optimization perspective. For example, VPO [10] constrains the rejected branch according to $\mathcal{V}$-usable information [17, 18], while CW-DPO [11] cools rejected responses using their average token log-probabilities. These methods reduce harmful updates on low-confidence rejected responses, but their reweighting criteria are not explicitly grounded in visual evidence. This is insufficient for MLLMs, where squeezing risk also depends on how much visually dependent hallucination-related content appears in the rejected response. Such content is repeatedly suppressed during DPO training, making the rejected response more likely to enter low-confidence regions that trigger squeezing. Other multimodal DPO variants notice abnormal dynamics, but they address them from a different angle. Methods with anchor terms [19, 4, 20, 21] observe that the chosen implicit reward can decrease together with the rejected implicit reward, and therefore add an anchor term to stabilize the chosen side. Although this encourages a higher implicit reward for chosen responses, it cannot guarantee that they receive higher policy probabilities. More importantly, it does not directly control the excessive negative update of the rejected branch, which is the source of the *squeezing effect* considered here.

These observations suggest that rejected responses should not be suppressed uniformly in multimodal DPO. As DPO training suppresses hallucination-related content, rejected responses with more visually dependent hallucination-related content are more likely to enter low-confidence regions and trigger squeezing, so they should receive weaker negative updates. Less risky rejected responses should stay close to standard DPO for preference discrimination. Based on this intuition, we propose **V**isual **G**uided **D**irect **P**reference **O**ptimization (VGDPO). To estimate the squeezing risk of each rejected response, VGDPO measures token-level visual dependency by comparing reference model likelihoods under the original image and a mosaic-corrupted image, and locates potentially hallucination-related phrases in rejected responses through semantic matching with chosen responses. These two signals define a hallucination ratio, used to relax the rejected branch according to multimodal squeezing risk. To further support the chosen response, VGDPO applies a similar reweighting principle to an auxiliary visual contrastive objective, which encourages higher likelihood under the original image than under the corrupted image. As shown in Figure 1(c), VGDPO improves the log probability dynamics by increasing the log probability of chosen responses while decreasing that of rejected responses.

Our contributions are as follows.

- We analyze the *squeezing effect* in multimodal DPO and show that it can reduce the absolute likelihood of visually grounded responses, especially when negative updates act on rejected responses with visually dependent hallucinations.
- We propose VGDPO, which estimates the squeezing risk of rejected responses using token-level visual dependency and phrase-level hallucination localization, and relaxes the negative update on high-risk rejected responses. We further introduce a reweighted visual contrastive objective that provides supervision for chosen responses under original visual evidence.
- We evaluate VGDPO on four hallucination benchmarks and five MLLM backbones spanning the LLaVA, Muffin, and Qwen families. The results show improved log probability dynamics, stronger gains under larger policy mismatch, and a better balance between hallucination reduction and response informativeness.

2 Preliminaries

We first review multimodal DPO and introduce the notation used by VGDPO. Given a multimodal preference sample (m, x, y_w, y_l), let m denote the image, x denote the textual query, and y_w and y_l denote the chosen and rejected responses. DPO [1] optimizes a policy from preference pairs without training an explicit reward model. It rewrites the reward of a response y to a prompt x as

$$r(x, y) = \beta \log \frac{\pi_\theta(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + \beta \log Z(x), \tag{1}$$

where π_θ is the policy model, π_{ref} is the reference model, $Z(x)$ is the partition function, and β controls the deviation from the reference policy. Combined with the Bradley-Terry model [22],

$$p(y_w \succ y_l \mid x) = \sigma\big(r(x, y_w) - r(x, y_l)\big), \tag{2}$$

DPO maximizes the likelihood of the observed preference pairs.

In the multimodal setting, the prompt is augmented with the image m. Therefore, for a chosen response y_w and a rejected response y_l, the standard multimodal DPO objective is

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E}_{(m,x,y_w,y_l)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid m, x)}{\pi_{\text{ref}}(y_w \mid m, x)} - \beta \log \frac{\pi_\theta(y_l \mid m, x)}{\pi_{\text{ref}}(y_l \mid m, x)} \right) \right]. \tag{3}$$

This objective increases the implicit reward gap between y_w and y_l, but it does not necessarily increase the absolute log probability of y_w. Existing methods with anchor terms [19, 4, 20, 21] further observe that the chosen implicit reward can decrease together with the rejected implicit reward during training, and therefore add a regularization term on the chosen implicit reward:

$$\mathcal{L}_{\text{anchor}} = -\mathbb{E}_{(m,x,y_w)\sim\mathcal{D}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y_w \mid m, x)}{\pi_{\text{ref}}(y_w \mid m, x)} \right) \right]. \tag{4}$$

This term helps stabilize the chosen response relative to the reference model. However, improving this relative reward does not directly control the absolute log probability of y_w, nor does it control the rejected branch that introduces strong negative updates. As discussed in Section 1, when preference-pair responses become relatively low-confidence under the current policy, negative updates on the rejected side may reduce the log probabilities of both y_w and y_l and move probability mass toward high-confidence outputs shaped by language priors. VGDPO therefore keeps the pairwise preference structure of DPO, but relaxes the negative update on the rejected branch according to visual dependency and hallucination localization.

3 Method

In this section, we introduce VGDPO, which relaxes negative updates on the rejected branch according to multimodal squeezing risk. VGDPO first combines token-level visual dependency with phrase-level hallucination localization to define a hallucination ratio for each rejected response. Visual dependency identifies tokens whose likelihood relies on image evidence, while hallucination localization identifies phrases that may conflict with the chosen response. Their combination focuses the risk estimate on content that is both visually dependent and potentially erroneous. A higher ratio leads to stronger relaxation of the rejected update, while a lower ratio keeps the objective closer to standard DPO for preference discrimination. VGDPO further applies a similar reweighting principle to a visual contrastive objective, providing supervision for chosen responses under the original image. Figure 2 gives an overview of the method.

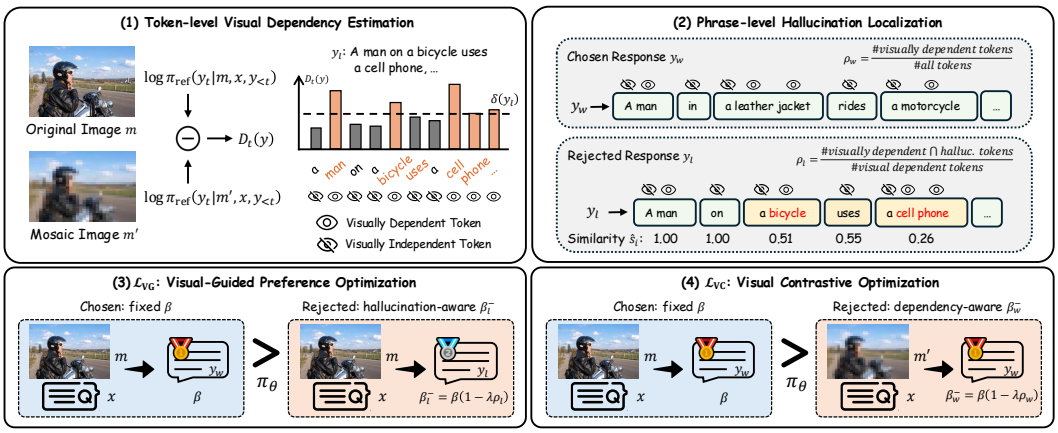


Figure 2: Overview of VGDPO. VGDPO estimates token-level visual dependency and phrase-level hallucination localization to relax high-risk rejected updates, and uses a related visual contrastive objective to supervise chosen responses.

3.1 Multimodal Squeezing Risk Estimation

Token-level Visual Dependency Estimation We estimate whether a token relies on visual evidence by comparing its likelihood under the original image and a corrupted image. We obtain the corrupted image m' by applying 32×32 mosaic corruption to m. This operation keeps the coarse spatial layout while removing fine-grained details such as object attributes and textures, which are often involved in visual hallucinations.

For a response sequence $y = (y_1, \ldots, y_L)$, we define the visual dependency score of token y_t as

$$D_t(y) = \log \pi_{\text{ref}}(y_t \mid m, x, y_{<t}) - \log \pi_{\text{ref}}(y_t \mid m', x, y_{<t}). \tag{5}$$

A larger $D_t(y)$ means that the token becomes much less likely after visual details are removed, indicating stronger dependence on the image. A smaller score means that the token is less affected by visual corruption and is more likely supported by language priors.

We use a response-specific threshold $\delta(y) = \frac{1}{L}\sum_{t=1}^{L} D_t(y)$ to identify visually dependent tokens. Tokens with $D_t(y) > \delta(y)$ are treated as visually dependent. This adaptive threshold avoids using a fixed global cutoff for responses with different lengths and visual contents.

Phrase-level Hallucination Localization We locate potentially hallucination-related content in the rejected response by using the chosen response as a semantic reference. Since the chosen response is expected to be better grounded in the image, rejected phrases that deviate from it are more likely to contain erroneous visual content. We segment both responses into phrase sets $P(y) = \{p_1, p_2, \ldots, p_n\}$ with spaCy*, and encode each phrase as $e_i = F_{\text{MiniLM}}(p_i)$ using `all-MiniLM-L6-v2` [23].

Let $P^+ = \{p_1^+, \ldots, p_{n_w}^+\}$ and $P^- = \{p_1^-, \ldots, p_{n_l}^-\}$ be the phrase sets of y_w and y_l. For each rejected phrase p_i^-, we compute its maximum semantic similarity to the chosen phrases:

$$\hat{s}_i = \max_{1 \le j \le n_w} \frac{(e_i^-)^\top e_j^+}{\|e_i^-\| \, \|e_j^+\|}, \quad i \in \{1, \ldots, n_l\}. \tag{6}$$

If $\hat{s}_i$ is larger than the similarity threshold η, the rejected phrase is considered semantically consistent with the chosen response. Otherwise, it is marked as potentially erroneous. We collect all tokens belonging to these potentially erroneous rejected phrases as $E_l = \{y_{l,t} \mid y_{l,t} \in p_i^-,\ \hat{s}_i \le \eta\}$.

3.2 Visual Guided Reweighting for Rejected Responses

The squeezing effect is mainly caused by the negative branch of DPO. As discussed in Appendix B, reducing the coefficient of this branch weakens excessive negative updates on rejected responses

*https://spacy.io/

and can also strengthen the optimization signal on chosen responses. In MLLMs, this reduction should depend on the content of the rejected response. We use the proportion of visually dependent hallucination-related tokens as a proxy for squeezing risk. Rejected responses with a high proportion receive stronger relaxation of the negative update, while responses with a low proportion stay closer to standard DPO to preserve the model's ability to distinguish them from chosen responses.

We then measure the proportion of visually dependent rejected tokens that are also potentially erroneous:

$$\rho_l = \frac{\sum_{t=1}^{L_l} \mathbb{I}\big(y_{l,t} \in E_l \ \wedge \ D_t(y_l) > \delta(y_l)\big)}{\sum_{t=1}^{L_l} \mathbb{I}\big(D_t(y_l) > \delta(y_l)\big)}, \tag{7}$$

where L_l is the length of y_l, $\mathbb{I}(\cdot)$ is the indicator function. The value ρ_l measures how much of the visually dependent part of the rejected response is potentially hallucination related. A larger ρ_l indicates higher squeezing risk, while a smaller ρ_l indicates that the response should remain closer to standard preference learning.

The hallucination-aware negative coefficient is defined as

$$\beta_l^- = \beta(1 - \lambda\rho_l), \tag{8}$$

where λ controls the reweighting strength. The resulting visual guided preference objective is

$$\mathcal{L}_{\mathrm{VG}} = -\mathbb{E}_{(m,x,y_w,y_l)\sim\mathcal{D}} \left[\log\sigma \left(\beta \log \frac{\pi_\theta(y_w \mid m, x)}{\pi_{\mathrm{ref}}(y_w \mid m, x)} - \beta_l^- \log \frac{\pi_\theta(y_l \mid m, x)}{\pi_{\mathrm{ref}}(y_l \mid m, x)} \right) \right]. \tag{9}$$

When ρ_l is large, VGDPO reduces β_l^- and weakens the negative update on the rejected branch. This avoids over suppressing rejected responses with higher squeezing risk. When ρ_l is small, β_l^- remains close to β, so the objective stays close to standard DPO. This keeps enough rejection pressure for samples with lower squeezing risk and preserves preference discrimination.

3.3 Visual Contrastive Objective with Reweighted Negative Branch

The visual guided preference objective alleviates squeezing by relaxing negative updates on high-risk rejected responses. However, we empirically observe that reweighting the rejected branch alone may not provide enough supervision for the chosen branch. The log probability of the chosen response can increase early in training and decrease again later. To supplement the positive branch gradient, we adopt the visual contrastive objective used in existing studies [19, 24, 25, 4], which compares the chosen response under the original image m and the corrupted image m'. On top of this objective, we introduce reweighted negative-gradient control on the corrupted-image branch. This objective encourages the policy to assign higher likelihood to y_w when the original visual evidence is available.

Following a related principle, we reweight the corrupted image branch according to the visual dependency of the chosen response. Unlike the standard preference objective, the negative branch here is also y_w, but conditioned on the corrupted image. Therefore, its coefficient should depend on how much the chosen response relies on visual evidence. Let $\rho_w = \frac{1}{L_w} \sum_{t=1}^{L_w} \mathbb{I}\big(D_t(y_w) > \delta(y_w)\big)$, where L_w is the length of y_w, and define $\beta_w^- = \beta(1 - \lambda\rho_w)$. The visual contrastive objective is then

$$\mathcal{L}_{\mathrm{VC}} = -\mathbb{E}_{(m,m',x,y_w)\sim\mathcal{D}} \left[\log\sigma \left(\beta \log \frac{\pi_\theta(y_w \mid m, x)}{\pi_{\mathrm{ref}}(y_w \mid m, x)} - \beta_w^- \log \frac{\pi_\theta(y_w \mid m', x)}{\pi_{\mathrm{ref}}(y_w \mid m', x)} \right) \right]. \tag{10}$$

The visual contrastive objective complements the visual guided preference objective. The preference objective relaxes high-risk rejected updates, while the contrastive objective supplies an additional gradient for the chosen response under the original image.

3.4 Overall Objective

The final training objective combines these two objectives.

$$\mathcal{L}_{\mathrm{VGDPO}} = \mathcal{L}_{\mathrm{VG}} + \gamma\mathcal{L}_{\mathrm{VC}}, \tag{11}$$

where γ balances the two objectives. In this form, VGDPO keeps the pairwise preference learning of DPO, relaxes the negative update for rejected responses with high squeezing risk, and keeps low-risk rejected responses close to standard DPO. Both the visual dependency scores and the phrase-level semantic matching scores are computed during preprocessing, so its training cost is similar to methods that also introduce a visual contrastive objective.

Table 1: Main experimental results. The best and second-best results under the same setting are highlighted in boldface and underlined, respectively.

Model	Data Size	AMBER				MMHal-Bench		HallusionBench			POPE	
		CHAIR↓	Cover↑	Hal↓	Cog↓	Score↑	Hal↓	qAcc↑	fAcc↑	aAcc↑	Adv.↑	All↑
LLaVA-1.5-7B [26]	✗	8.8	50.3	39.9	4.9	2.48	54.2	3.07	10.69	37.37	84.8	87.0
+LLaVA-RLHF [27]	122k	10.3	50.3	49.7	5.2	2.00	66.7	5.05	9.82	34.45	81.0	84.8
+POVID [28]	17k	5.3	49.6	28.4	3.2	2.28	56.2	4.61	8.38	37.90	84.4	87.1
+HALVA [29]	21.5k	8.9	51.1	39.4	4.5	2.14	60.4	4.83	10.69	39.94	84.1	85.5
+HA-DPO [2]	6k	7.7	48.8	35.5	4.2	1.97	61.5	3.51	10.98	39.68	83.6	85.4
+RLAIF-V [30]	74.8k	2.9	50.5	15.8	0.9	3.04	39.6	3.51	2.89	29.67	83.3	84.8
+TPR [31]	21.4k	3.0	48.6	16.6	1.2	2.76	42.7	3.95	7.80	34.54	82.9	84.0
+OPA-DPO [4]	4.8k	2.6	45.5	12.4	1.0	2.83	46.9	5.05	8.67	36.58	82.7	83.7
+SymMPO [20]	21.4k	5.4	49.7	27.1	3.0	2.89	42.7	5.05	10.11	38.88	79.9	80.3
+DPO [1]	6k	**2.8**	45.0	**14.0**	**0.9**	2.71	47.9	2.63	9.82	34.18	84.5	**86.2**
+Anchor [19, 4, 20, 21]	6k	4.6	**50.4**	22.6	2.4	2.64	52.1	3.07	10.40	**37.82**	84.5	86.0
+CW-DPO [11]	6k	4.8	48.0	22.2	2.0	2.66	51.0	**3.51**	9.82	35.69	84.2	86.1
+VGDPO (Ours)	6k	3.7	49.3	18.0	1.8	**2.89**	**45.8**	**3.51**	**11.27**	37.11	**84.6**	**86.2**
LLaVA-1.5-13B [26]	✗	8.7	50.5	38.1	4.0	2.32	59.4	4.17	7.22	38.52	85.4	88.5
+LLaVA-RLHF [27]	122k	9.0	50.3	42.8	4.8	2.15	66.7	6.37	9.24	37.02	81.7	83.7
+HALVA [29]	21.5k	8.4	51.0	36.5	3.9	2.30	57.3	6.15	8.38	36.93	85.8	87.2
+HSA-DPO [3]	8k	2.8	45.4	14.6	1.5	2.55	50.0	3.95	6.06	36.58	85.1	87.3
+OPA-DPO [4]	4.8k	2.7	47.5	12.5	1.0	3.05	38.5	3.95	9.53	36.75	84.9	86.1
+SymMPO [20]	21.4k	4.3	52.7	23.7	1.9	3.01	39.6	6.81	9.24	37.11	82.3	82.9
+DPO [1]	6k	3.4	45.2	**14.8**	**1.0**	2.95	44.8	5.05	**7.22**	35.87	85.3	87.0
+Anchor [19, 4, 20, 21]	6k	3.9	**50.0**	19.1	1.6	2.86	46.9	5.05	6.64	37.55	85.5	87.5
+CW-DPO [11]	6k	5.1	48.6	24.4	1.9	2.86	44.8	4.61	5.49	33.03	85.6	**87.8**
+VGDPO (Ours)	6k	**3.1**	48.2	15.9	1.2	**3.01**	**42.7**	**5.49**	6.06	**38.88**	**85.7**	**87.8**
Muffin-13B [32]	✗	13.5	52.8	57.4	6.7	2.60	67.7	3.29	8.09	34.98	78.7	80.7
+RLHF-V [33]	1.4k	9.1	47.8	32.5	3.5	2.97	57.0	6.81	11.27	40.12	84.7	86.5
+DPO [1]	6k	7.4	50.5	31.4	3.1	2.94	58.0	5.05	8.09	35.25	73.8	74.1
+Anchor [19, 4, 20, 21]	6k	7.2	**50.7**	31.2	3.3	2.86	57.3	5.93	9.24	**37.20**	73.6	74.2
+CW-DPO [11]	6k	7.2	49.7	33.1	3.7	2.34	55.2	4.17	8.67	34.18	60.6	64.0
+VGDPO (Ours)	6k	**6.9**	50.5	**29.8**	**2.9**	**3.06**	**51.0**	**6.37**	**9.53**	36.93	**74.0**	**74.7**
Qwen2.5-VL-3B [34]	✗	8.1	**69.5**	51.9	5.6	3.70	40.6	10.76	19.94	44.99	86.2	87.3
+DPO [1]	6k	6.2	66.3	43.9	4.2	**3.89**	**34.4**	11.86	20.80	43.57	86.2	87.1
+Anchor [19, 4, 20, 21]	6k	4.3	56.6	22.0	1.7	3.40	37.5	14.72	23.12	48.62	86.3	87.5
+CW-DPO [11]	6k	5.1	50.2	24.6	2.4	2.98	47.9	15.60	22.83	48.98	86.5	**89.4**
+VGDPO (Ours)	6k	**4.0**	50.7	**19.5**	**1.6**	3.46	40.6	**16.26**	**23.41**	**49.15**	**86.7**	88.3
Qwen3.5-2B [35]	✗	9.3	**76.6**	69.3	4.7	4.61	32.3	13.40	22.54	43.93	87.8	**89.5**
+DPO [1]	6k	6.2	72.3	47.7	2.3	**4.82**	24.0	15.38	22.83	45.26	87.2	88.4
+Anchor [19, 4, 20, 21]	6k	**2.4**	57.4	15.7	**0.9**	3.82	28.1	17.36	**24.56**	46.58	87.8	**89.5**
+CW-DPO [11]	6k	2.6	51.0	13.6	1.3	3.71	27.1	18.90	23.12	47.82	87.7	89.2
+VGDPO (Ours)	6k	**2.4**	56.0	**13.1**	**0.8**	3.95	**21.9**	**19.34**	**24.56**	**48.09**	**88.0**	**89.5**

4 Experiments

4.1 Experimental Setup

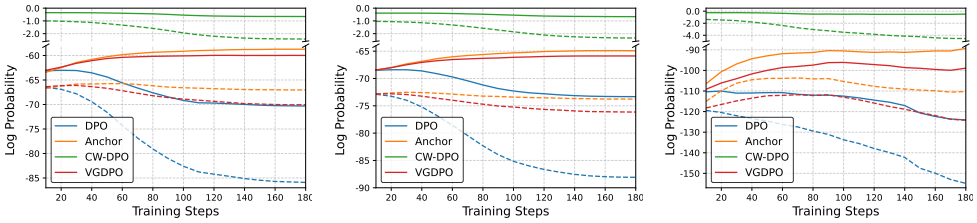
Implementation Details We evaluate VGDPO on five backbones: LLaVA-1.5-7B, LLaVA-1.5-13B [26], Muffin-13B [32], Qwen2.5-VL-3B [34], and Qwen3.5-2B [35], with details in Appendix C. We train all backbones on 6k instances sampled from the 21.4k TPR [31] dataset. For controlled training across all five backbones, we use $\beta = 0.1$, a learning rate of 1×10^{-6}, a batch size of 64, and 2 epochs on 8 NVIDIA RTX PRO 6000 GPUs. We tune VGDPO over $\eta \in \{0.6, 0.7, 0.8, 0.9\}$, $\lambda \in \{1.0, 0.8, 0.6, 0.4\}$, and $\gamma \in \{0.0, 0.2, 0.4, 0.6\}$.

Evaluation Metrics We evaluate on four multimodal hallucination benchmarks covering generative and discriminative settings: AMBER [36], MMHal-Bench [27], HallusionBench [14], and POPE [37]. Detailed benchmark descriptions and metrics are given in Appendix C.

Baselines We compare with representative multimodal alignment methods that provide public checkpoints, including LLaVA-RLHF [27], POVID [28], HALVA [29], HA-DPO [2], RLAIF-V [30], TPR [31], OPA-DPO [4], SymMPO [20], HSA-DPO [3], and RLHF-V [33]. These checkpoints use different data and training recipes. For a controlled comparison, we also train vanilla DPO, a DPO variant with an anchor term, and CW-DPO on the same 6k data as VGDPO. Since the official CW-DPO code is not publicly available, we implement its two-stage training procedure following the paper, including smoothing SFT and confidence-weighted DPO.

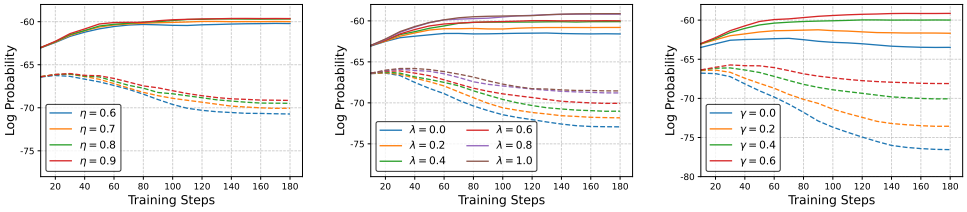
4.2 Main Results

Table 1 reports the results across five backbones. Under the controlled 6k-data setting, VGDPO provides the most consistent balance between hallucination reduction and response informativeness.



(a) LLaVA-1.5-7B (b) LLaVA-1.5-13B (c) Muffin-13B

Figure 3: Training dynamics on the probe set under DPO, Anchor, CW-DPO, and VGDPO across three backbones. Solid denotes chosen, dashed denotes rejected.



(a) Similarity threshold η (b) Reweighting strength λ (c) Visual contrast weight γ

Figure 4: Training dynamics of VGDPO on the probe set of LLaVA-1.5-7B with different hyperparameter settings. Solid denotes chosen, dashed denotes rejected.

On LLaVA and Muffin, it generally achieves the best results on MMHal-Bench and HallusionBench with competitive AMBER coverage and POPE accuracy. Although DPO sometimes gives lower AMBER hallucination scores, its lower coverage suggests that this improvement partly comes from less informative responses. The two Qwen models initially produce verbose, high-coverage responses with more unsupported content, whereas the chosen responses in our 6k data are shorter and better grounded. VGDPO shifts them toward this preferred style, reducing AMBER hallucination while retaining more coverage than CW-DPO and achieving the best overall HallusionBench results with competitive POPE accuracy. The base models or DPO occasionally score higher on MMHal-Bench because its GPT-based evaluation also rewards informativeness and may favor longer responses. Since the preference data are constructed with LLaVA-1.5-7B, the lower response probabilities of the other backbones in Figure 6 indicate stronger policy mismatch, under which VGDPO generally yields larger gains. These results support content-aware squeezing mitigation on recent MLLM architectures, although gains vary across benchmarks.

To study training dynamics, we sample 200 preference pairs from the training set as probe data and track the log probabilities of chosen and rejected responses during training. Figure 3 gives a closer look at the curves. With vanilla DPO, both chosen and rejected log probabilities decrease, matching the squeezing behavior discussed in Section 1. Anchor mainly regularizes the chosen response's implicit reward. The chosen log probability increases, but the rejected response is not sufficiently suppressed, so the preference contrast remains weak. CW-DPO has higher log probabilities for both branches because smoothing SFT raises both chosen and rejected responses. However, its second-stage DPO does not clearly enlarge their margin, making it difficult for the model to distinguish chosen responses from rejected ones. This may explain its weaker benchmark trade-off. VGDPO shows a clearer preference learning pattern: the chosen log probability increases, while the rejected log probability decreases. This suggests that VGDPO alleviates squeezing while preserving discriminative pressure.

4.3 Ablation Study

We study three hyperparameters of VGDPO on LLaVA-1.5-7B: the phrase similarity threshold η, the reweighting strength λ, and the visual contrast weight γ. Figure 4 shows how they affect the probe log probability curves. Table 2 reports quantitative results for γ, and Table 3 compares different ratio definitions for controlling the rejected branch. Results for η and λ are provided in Appendix D.

Figure 4(a) shows that η controls phrase matching strictness. Larger values mark more rejected phrases as potentially erroneous and weaken the rejected update, while smaller values may miss

Table 2: Ablation results of VGDPO with different visual contrast weights γ. The best and second-best results are highlighted in boldface and underlined, respectively.

VGDPO	AMBER				MMHal-Bench		HallusionBench			POPE	
	CHAIR↓	Cover↑	Hal↓	Cog↓	Score↑	Hal↓	qAcc↑	fAcc↑	aAcc↑	Adv.↑	All↑
$\gamma = 0.0$	**2.9**	46.2	**13.7**	**1.1**	2.79	49.0	**3.95**	10.69	36.31	84.0	85.3
$\gamma = 0.2$	3.4	49.0	16.0	1.8	**2.90**	**44.8**	3.51	10.69	35.78	84.2	85.6
$\gamma = 0.4$	3.7	49.3	18.0	1.8	2.89	45.8	3.51	**11.27**	37.11	**84.6**	**86.2**
$\gamma = 0.6$	4.5	**49.9**	22.1	2.3	2.68	47.9	3.73	10.40	**38.44**	84.3	86.0

Table 3: Ablation results of different definitions of ρ_l in Equation 8. **Halluc./Visual** uses hallucination-related and visually dependent tokens over visually dependent tokens. **Halluc./All** changes the denominator to all tokens, while **Visual/All** uses visually dependent tokens over all tokens.

Ratio	AMBER				MMHal-Bench		HallusionBench			POPE	
	CHAIR↓	Cover↑	Hal↓	Cog↓	Score↑	Hal↓	qAcc↑	fAcc↑	aAcc↑	Adv.↑	All↑
Halluc./Visual	**3.7**	49.3	18.0	1.8	**2.89**	**45.8**	**3.51**	**11.27**	**37.11**	84.6	86.2
Halluc./All	**3.7**	**49.6**	**17.1**	**1.7**	2.71	47.9	3.07	10.69	37.02	84.5	86.0
Visual/All	3.8	49.5	18.1	2.1	2.54	51.0	3.29	10.40	37.02	**84.7**	**86.3**

hallucination-related phrases. We use $\eta = 0.7$ to balance localization and semantic tolerance. Figure 4(b) studies the reweighting strength λ. Smaller values preserve stronger rejection pressure, while larger values relax the rejected branch more aggressively, showing the trade-off between alleviating squeezing and maintaining discrimination. We use $\lambda = 0.6$ to keep a moderate level of regularization on the rejected response. Figure 4(c) illustrates the role of the visual contrast weight γ. When $\gamma = 0.0$, VGDPO removes the visual contrastive objective, and the chosen log probability rises early but later declines. This verifies the effect of the visual contrastive term. A moderate weight sustains chosen response likelihood, while too large a weight can overemphasize the auxiliary objective. Table 2 shows that $\gamma = 0.4$ gives the best trade-off.

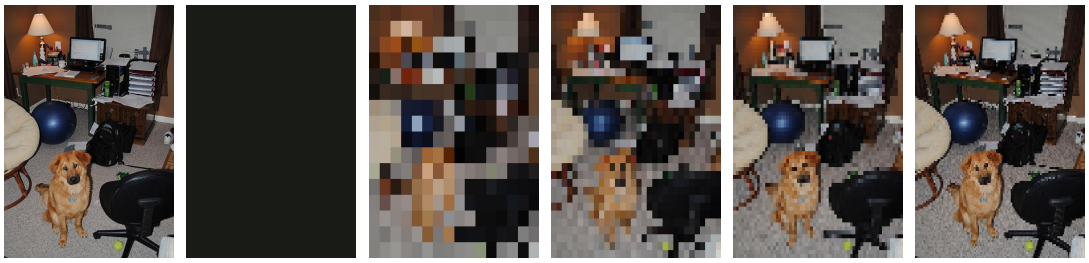
Table 3 compares different definitions of ρ_l by changing the numerator or denominator in Equation 7. The default **Halluc./Visual** ratio uses hallucination-related visually dependent tokens over visually dependent tokens and gives the best overall balance, especially on MMHal-Bench and HallusionBench. **Halluc./All** changes the denominator to all tokens and is competitive on AMBER, but it can dilute visually dependent errors in long responses. **Visual/All** uses visually dependent tokens over all tokens, ignores whether visual tokens are erroneous, and performs worse on most benchmarks. These results support combining hallucination localization with visual dependency when estimating squeezing risk.

4.4 Effect of Visual Corruption Strategy

We first study how mosaic granularity affects visual dependency estimation. Figure 5 shows the corrupted images under different granularities. The amount of removed visual detail directly affects which tokens are identified as visually dependent. The upper block of Table 4 reports the results.

Table 4: Ablation results of mosaic granularities and visual corruption strategies. The best and second-best results within each comparison block are highlighted in boldface and underlined, respectively.

Setting	AMBER				MMHal-Bench		HallusionBench			POPE	
	CHAIR↓	Cover↑	Hal↓	Cog↓	Score↑	Hal↓	qAcc↑	fAcc↑	aAcc↑	Adv.↑	All↑
Mosaic granularity											
1×1	**3.7**	48.7	**17.4**	**1.6**	2.78	47.9	3.29	10.11	37.55	**84.6**	86.1
16×16	3.8	49.1	17.6	1.7	2.70	51.0	2.63	10.11	36.58	84.4	85.8
32×32	**3.7**	49.3	18.0	1.8	**2.89**	**45.8**	**3.51**	**11.27**	37.11	**84.6**	**86.2**
64×64	3.8	**50.2**	18.9	2.0	2.70	50.0	**3.51**	10.98	**37.99**	84.5	86.0
128×128	3.9	49.7	19.1	2.1	2.77	46.9	3.07	10.69	37.11	84.5	86.1
Corruption strategy											
Mosaic	3.7	**49.3**	18.0	1.8	2.89	45.8	**3.51**	**11.27**	**37.11**	**84.6**	**86.2**
Blur	**3.2**	47.3	**14.2**	**1.3**	**2.91**	**43.8**	3.29	9.82	36.75	84.4	85.7
Noise	3.3	48.8	15.4	1.5	2.85	44.8	**3.51**	**11.27**	**37.11**	84.2	85.6



(a) Original (b) 1×1 (c) 16×16 (d) 32×32 (e) 64×64 (f) 128×128

Figure 5: Original image and mosaic images with different granularities.

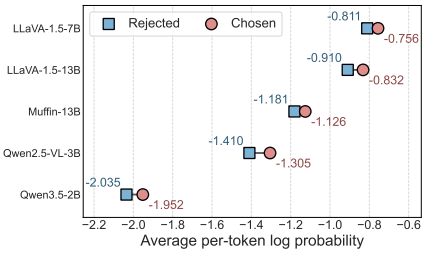


Figure 6: Per-token log probabilities of preference responses across five models.



(a) Original (b) Mosaic (c) Blur (d) Noise

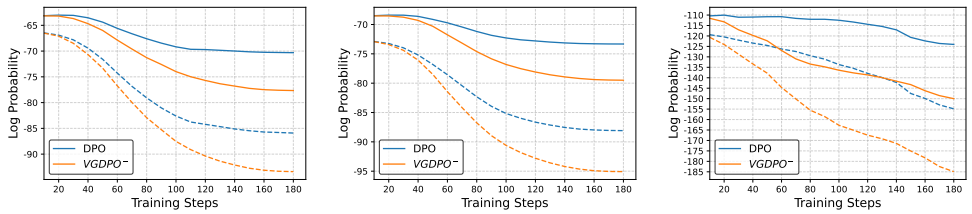
Figure 7: Examples of the original image and different visual corruptions.

Very coarse corruption such as 1×1 and 16×16 removes too much visual information. Many tokens then become sensitive to the corrupted image, and the dependency signal loses selectivity. Very fine corruption such as 64×64 and 128×128 leaves the corrupted image too similar to the original, making it hard to detect tokens that rely on fine-grained visual evidence. The 32×32 setting gives the best overall balance across benchmarks. It preserves coarse layout while removing enough local detail to identify visually dependent tokens, yielding the strongest MMHal-Bench and POPE results while remaining competitive on AMBER.

To test robustness to the corruption choice, we replace the default 32×32 mosaic with Gaussian blur and Gaussian noise at severity level 5 following ImageNet-C [38], as illustrated in Figure 7. The lower block of Table 4 shows comparable overall performance across all three strategies, with blur and noise further improving several hallucination metrics. This confirms that VGDPO remains stable across different corruption operators. We use mosaic by default because it is simple to apply and control while preserving coarse layout and removing local visual details.

4.5 Reversed Reweighting Analysis

Our method assumes that visually dependent hallucination-related rejected content has higher squeezing risk and should receive weaker negative updates. To directly test this assumption, we construct a



(a) LLaVA-1.5-7B (b) LLaVA-1.5-13B (c) Muffin-13B

Figure 8: Training dynamics on the probe set under DPO and VGDPO$^-$ across three backbones. Solid denotes chosen, dashed denotes rejected.

reversed variant of VGDPO by changing the sign in Equation 8:

$$\beta_l^- = \beta(1+\lambda\rho_l). \tag{12}$$

Unlike VGDPO, this variant increases the negative coefficient for rejected responses with larger hallucination ratios. It intentionally strengthens the update that VGDPO relaxes. Since this comparison focuses on the rejected branch, we do not include the visual contrastive objective. Figure 8 shows the log probability curves under DPO and this reversed variant. We denote this variant as VGDPO$^-$. Table 5 reports the corresponding benchmark results with full VGDPO included for comparison.

Figure 8 shows that VGDPO$^-$ makes the log probabilities of both chosen and rejected responses decrease more strongly than vanilla DPO. This suggests that increasing the negative coefficient on visually dependent hallucination-related content amplifies the *squeezing effect*. Table 5 gives a consistent message at the benchmark level. As the training data become more off-policy from LLaVA-1.5-7B to Muffin-13B, squeezing is more likely to occur, and VGDPO shows a clearer advantage. It lowers MMHal-Bench hallucination across all backbones and improves AMBER hallucination metrics in the most off-policy Muffin-13B setting, supporting our design choice of relaxing negative updates for responses with higher squeezing risk.

Table 5: DPO, VGDPO$^-$, and VGDPO results.

Backbone	Algo.	AMBER			MMHal
		CHAIR↓	Cover↑	Hal↓	Hal↓
LLaVA-1.5-7B	DPO	2.8	45.0	14.0	47.9
	VGDPO$^-$	2.9$_{(+0.1)}$	43.0$_{(-2.0)}$	14.1$_{(+0.1)}$	51.0$_{(+3.1)}$
	VGDPO	3.7$_{(+0.9)}$	49.3$_{(+4.3)}$	18.0$_{(+4.0)}$	45.8$_{(-2.1)}$
LLaVA-1.5-13B	DPO	3.4	45.2	14.8	44.8
	VGDPO$^-$	3.2$_{(-0.2)}$	44.5$_{(-0.7)}$	15.0$_{(+0.2)}$	46.9$_{(+2.1)}$
	VGDPO	3.1$_{(-0.3)}$	48.2$_{(+3.0)}$	15.9$_{(+1.1)}$	42.7$_{(-2.1)}$
Muffin-13B	DPO	7.4	50.5	31.4	58.0
	VGDPO$^-$	8.8$_{(+1.4)}$	51.2$_{(+0.7)}$	38.7$_{(+7.3)}$	63.5$_{(+5.5)}$
	VGDPO	6.9$_{(-0.5)}$	50.5$_{(+0.0)}$	29.8$_{(-1.6)}$	51.0$_{(-7.0)}$

4.6 General Multimodal Evaluation

We further evaluate whether hallucination mitigation affects general multimodal capabilities on four benchmarks: MMBench [39], MM-Vet [40], LLaVA-Bench [26], and MME [41]. As shown in Table 6, VGDPO improves the LLaVA-1.5-7B backbone on the first three benchmarks, while remaining competitive with DPO on MME. These results show that alleviating squeezing preserves general multimodal performance. They also suggest that the benefits of VGDPO extend beyond hallucination-specific evaluation.

Table 6: Results on multimodal benchmarks.

Method	MMBench	MM-Vet	LLaVA-Bench	MME
LLaVA-1.5-7B	65.2	31.5	60.7	1784.1
+DPO	66.3	31.6	65.1	**1793.3**
+VGDPO	**66.4**	**35.3**	**65.7**	1787.4

5 Conclusion

This paper studied the *squeezing effect* in multimodal DPO, where preference optimization can reduce the likelihood of both chosen and rejected responses and shift probability mass toward language priors rather than visual evidence. We proposed VGDPO, which estimates the squeezing risk of rejected responses from visual dependency and hallucination localization, then relaxes high-risk rejected updates and adds a visual contrastive objective for chosen responses. Experiments on four hallucination benchmarks and five MLLM backbones spanning three model families show that VGDPO improves DPO training dynamics and achieves a better balance between hallucination reduction and response informativeness.

Limitations. Despite its effectiveness, VGDPO still has several limitations. The response-specific mean threshold is purely relative, so a response with uniformly low visual dependency may still assign some tokens above the mean due to minor noise. A hybrid threshold with an absolute lower bound may better avoid over-relaxing such cases. The phrase-level hallucination localization module relies on the chosen response as a semantic reference, so brief or incomplete chosen responses may omit valid details, produce false positives, and favor overly concise outputs, introducing a faithfulness–informativeness trade-off. Future work can explore stronger visual consistency matching methods and more localized visual interventions.

Acknowledgments and Disclosure of Funding

This work was supported by the Key R&D Program of Shandong Province, China (Grant No. 2025CXGC020101) and the National Natural Science Foundation of China (Grant Nos. 62572282, 62376137, 62576195 and 62276155). Beyond the funding disclosed above, the authors declare no competing financial or non-financial interests that could be perceived as influencing this work.

References

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

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

[3] Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 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, 2025.

[4] Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10610–10620, 2025.

[5] Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, Yinwei Wei, and Xuemeng Song. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. In *Forty-third International Conference on Machine Learning*, 2026.

[6] Wenqi Liu, Shijie Ma, Yunxiao Wang, Meng Liu, Qile Su, Han Liu, Bohan Hou, Zeyu Wang, Xuanyu Zheng, Changyi Liu, et al. Thinking beyond videos: Unifying video reasoning and deep research for open-world video agents. *arXiv preprint arXiv:2608.23329*, 2026.

[7] Yi Ren and Danica J. Sutherland. Learning dynamics of LLM finetuning. In *The Thirteenth International Conference on Learning Representations*, 2025.

[8] Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

[9] Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

[10] Zecheng Wang, Chunshan Li, Yupeng Zhang, Han Liu, Bingning Wang, Dianhui Chu, and Dianbo Sui. VPO: Reasoning preferences optimization based on $\mathcal{V}$-usable information. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[11] Jusheng Zhang, Kaitong Cai, Jing Yang, and Keze Wang. Learning dynamics of vlm finetuning. *arXiv preprint arXiv:2510.11978*, 2025.

[12] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[13] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853, 2024.

[14] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385, 2024.

[15] Shijie Ma, Yuying Ge, Teng Wang, Yuxin Guo, Yixiao Ge, and Ying Shan. Genhancer: Imperfect generative models are secretly strong vision-centric enhancers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 24402–24412. IEEE, 2025.

[16] Yunxiao Wang, Meng Liu, Wenqi Liu, Xuemeng Song, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, Guorui Zhou, and Liqiang Nie. Time: Temporal-sensitive multi-dimensional instruction tuning and robust benchmarking for video-llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10323–10331, 2026.

[17] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR, 2022.

[18] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. In *International Conference on Learning Representations*, 2020.

[19] Fei Wang, Wenxuan Zhou, James Y Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mdpo: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8078–8088, 2024.

[20] Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Yuanshuai Li, Yuping Yan, Jirui Han, Fei Ming, Lingjuan Lv, and Yaochu Jin. Iris: Implicit reward-guided internal sifting for mitigating multimodal hallucination. *arXiv preprint arXiv:2602.01769*, 2026.

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

[23] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788, 2020.

[24] Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-dpo: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13258–13273, 2024.

[25] Songtao Jiang, Yan Zhang, Ruizhe Chen, Tianxiang Hu, Yeying Jin, Qinglin He, Yang Feng, Jian Wu, and Zuozhu Liu. Modality-fair preference optimization for trustworthy mllm alignment. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 403–411, 2025.

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[27] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110, 2024.

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

[29] Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan O Arik, and Tomas Pfister. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *The Thirteenth International Conference on Learning Representations*, 2025.

[30] Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, et al. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19985–19995, 2025.

[31] Lehan He, Zeren Chen, Zhelun Shi, Tianyu Yu, Lu Sheng, and Jing Shao. Systematic reward gap optimization for mitigating VLM hallucinations. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[32] Tianyu Yu, Jinyi Hu, Yuan Yao, Haoye Zhang, Yue Zhao, Chongyi Wang, Shan Wang, Yinxv Pan, Jiao Xue, Dahai Li, et al. Reformulating vision-language foundation models and datasets towards universal multimodal assistants. *arXiv preprint arXiv:2310.00653*, 2023.

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

[34] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[35] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[36] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[37] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305, 2023.

[38] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

[39] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024.

[40] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning*, 2024.

[41] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026.

[42] Jinlan Fu, huangfushenzhen, Hao Fei, Xiaoyu Shen, Bryan Hooi, Xipeng Qiu, and See-Kiong Ng. Chip: Cross-modal hierarchical direct preference optimization for multimodal LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025.

[43] Yuanshuai Li, Yuping Yan, Junfeng Tang, Yunxuan Li, Zeqi Zheng, and Yaochu Jin. Mitigating visual hallucinations via semantic curriculum preference optimization in mllms. *arXiv preprint arXiv:2509.24491*, 2025.

[44] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[45] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[46] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

[47] Kevin Lu and Thinking Machines Lab. On-policy distillation. *Thinking Machines Lab: Connectionism*, 2025. https://thinkingmachines.ai/blog/on-policy-distillation.

[48] Shijie Ma, Fei Zhu, Yuxin Guo, Meng Wang, Wenzhuo Liu, Zhun Zhong, Xu-Yao Zhang, and Cheng-Lin Liu. Happy++: Towards stable and unified continual generalized category discovery. 2026.

[49] Biqing Qi, Pengfei Li, Fangyuan Li, Junqi Gao, Kaiyan Zhang, and Bowen Zhou. Online dpo: Online direct preference optimization with fast-slow chasing. *arXiv preprint arXiv:2406.05534*, 2024.

[50] Chengzhi Yu, Yifan Xu, Yifan Chen, and Wenyi Zhang. Optimizing lvlms with on-policy data for effective hallucination mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4733–4743, 2026.

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

[52] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

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

[54] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[55] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023.

[56] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

A Related Work

A.1 Multimodal Hallucination Mitigation

Direct Preference Optimization (DPO) [1] has been widely used to reduce hallucination in Multimodal Large Language Models (MLLMs). One line of work focuses on preference data construction. RLHF-V [33] uses human annotation to revise erroneous model responses and build preference pairs. RLAIF-V [30] reduces annotation cost with an automatic pipeline that uses open source MLLMs to generate and filter preference data. TPR [31] constructs preference pairs at the semantic topic level through clustering and rewriting, which provides more informative preference supervision.

Another line of work modifies the preference optimization objective for multimodal alignment. For example, mDPO [19] adds visual contrast and anchor terms to facilitate visual discrimination and regularize the implicit reward of the chosen response. CHiP [42] uses preference signals at multiple granularities to improve cross-modal alignment. SymMPO [20] studies the inconsistency in multimodal DPO and introduces symmetric image-response pairs. SCPO [43] builds preference data with different difficulty levels and uses curriculum-based preference alignment. Methods with anchor terms [19, 4, 20, 21] study a related but different training issue: the chosen implicit reward can decrease together with the rejected implicit reward. These methods stabilize the chosen side relative to the reference model, but they mainly improve implicit reward and do not directly address the decline of chosen log probability or the excessive negative update on visually mismatched rejected responses. VCD [12] compares model distributions under original and degraded images to adjust next-token logits during inference. VGDPO connects multimodal squeezing dynamics to content-aware risk estimation and selective rejected-side optimization by combining offline visual contrast with phrase-level hallucination localization, while using standard decoding after training.

A.2 Preference Optimization under Policy Mismatch

Existing studies [44, 45, 46, 47, 48] show that the data distribution and sampling policy strongly affect preference optimization. Online methods collect new preference labels during training, while offline methods learn from a fixed preference dataset. On-policy methods optimize responses sampled from the current policy, while off-policy methods use responses generated by a different behavior policy. Standard DPO is usually offline and off-policy. The preference labels and responses are fixed before training, and the optimized policy gradually moves away from the policy that generated the data. This mismatch can make static preference pairs less effective as training proceeds [8].

Several methods try to reduce this mismatch by moving training closer to an on-policy process. For example, OAIF [9] uses an LLM judge to compare responses sampled from the current model and produce preference labels during training. OFS-DPO [49] uses fast and slow modules to adapt to newly sampled data while reducing forgetting. In the multimodal setting, Yu et al. [50] construct preference pairs from sampled responses during training. OPA-DPO [4] adds an early SFT stage to facilitate on-policy alignment. These methods can improve alignment, but they often need extra sampling, judging, or training stages.

The *squeezing effect* [7] gives another view of why off-policy DPO can fail. With a softmax output head, strong negative gradients on low confidence rejected responses can push probability mass toward high confidence outputs. As a result, the log probabilities of both chosen and rejected responses may decrease. This effect is related to policy mismatch because fixed rejected responses can become lower confidence samples as the current policy moves away from the behavior policy. VPO [10] addresses this issue in text reasoning tasks by constraining the negative gradient according to $\mathcal{V}$-usable information, while CW-DPO [11] cools VLM negatives according to their average token log probability. VPO targets text reasoning, while CW-DPO reweights by confidence without explicitly modeling visual evidence or hallucination degree. VGDPO estimates rejected-side squeezing risk from visual dependency and hallucination localization to guide selective negative updates.

B Gradient Analysis of VGDPO

We analyze the gradients of VGDPO to clarify two design choices. The first is why VGDPO reweights the rejected branch. The second is how the visual contrastive term supplies additional chosen response supervision while contrasting the original and corrupted images.

B.1 Gradient Analysis of the Visual Guided Preference Loss

We analyze how constraining the rejected branch affects both branches. Given image m, text query x, chosen response y_w, and rejected response y_l, define

$$r = \beta \log \frac{\pi_\theta(y_w \mid m, x)}{\pi_{\text{ref}}(y_w \mid m, x)} - \beta(1-w) \log \frac{\pi_\theta(y_l \mid m, x)}{\pi_{\text{ref}}(y_l \mid m, x)}, \tag{13}$$

and

$$\mathcal{L} = -\log \sigma(r), \tag{14}$$

where $w \in (0, 1)$ controls the negative coefficient of the rejected branch. For the chosen response, taking the derivative gives

$$\frac{\partial \mathcal{L}}{\partial \pi_\theta(y_w \mid m, x)} = \frac{\partial \mathcal{L}}{\partial r} \cdot \frac{\partial r}{\partial \pi_\theta(y_w \mid m, x)} = (\sigma(r) - 1) \cdot \frac{\beta}{\pi_\theta(y_w \mid m, x)}. \tag{15}$$

For the rejected response, we have

$$\frac{\partial \mathcal{L}}{\partial \pi_\theta(y_l \mid m, x)} = \frac{\partial \mathcal{L}}{\partial r} \cdot \frac{\partial r}{\partial \pi_\theta(y_l \mid m, x)} = (1 - \sigma(r)) \cdot \frac{\beta(1-w)}{\pi_\theta(y_l \mid m, x)}. \tag{16}$$

The chosen-branch derivative is negative, so gradient descent increases the chosen response probability, while the rejected-branch derivative is positive and decreases the rejected response probability. Let $\Delta_l = \log \frac{\pi_\theta(y_l|m,x)}{\pi_{\text{ref}}(y_l|m,x)}$. For low-confidence rejected samples, Δ_l is usually negative, and introducing w changes the margin from the DPO value r_{DPO} to $r = r_{\text{DPO}} + \beta w \Delta_l < r_{\text{DPO}}$. Since $\sigma(\cdot)$ is increasing, the factor $1 - \sigma(r)$ becomes larger, which enlarges the magnitude of the chosen derivative and gives stronger positive supervision to y_w.

At the same time, the rejected branch is multiplied by the explicit coefficient $\beta(1 - w)$. This coefficient directly relaxes the negative update on y_l, preventing low-confidence rejected responses from receiving overly large negative gradients that can trigger squeezing. The full rejected gradient also contains the sigmoid factor $1 - \sigma(r)$, so the final magnitude is determined by both terms. Since the sigmoid factor is bounded and changes smoothly, its increase only partially offsets the linear reduction induced by $1-w$. The key effect is that the harmful negative branch is explicitly constrained, while the chosen branch can receive a stronger positive update. A very large constraint may weaken preference discrimination, so the constraint should be applied selectively rather than uniformly.

This analysis matches the multimodal motivation. Not all rejected content should receive the same negative update. If a rejected response has a high proportion of visually dependent hallucination-related content, it has higher squeezing risk and should receive stronger relaxation of the negative update on the rejected branch. If this proportion is low, the squeezing risk is weaker and the model should still learn to distinguish the rejected response from the chosen one. In the main paper, β_l^- implements this idea by setting the rejected branch coefficient according to the hallucination ratio of visually dependent tokens.

B.2 Gradient Analysis of the Visual Contrastive Loss

The visual contrastive objective is introduced for a different but related reason. Reweighting the rejected branch can alleviate squeezing, but it may not provide enough direct supervision for the chosen response. The visual contrastive term adds another gradient that increases the chosen response probability under the original image and contrasts it with the corrupted image.

We write the visual contrastive loss as

$$\mathcal{L}_{\text{VC}} = -\log \sigma(r_v), \tag{17}$$

where

$$r_v = \beta \log \frac{\pi_\theta(y_w \mid m, x)}{\pi_{\text{ref}}(y_w \mid m, x)} - \beta(1-v) \log \frac{\pi_\theta(y_w \mid m', x)}{\pi_{\text{ref}}(y_w \mid m', x)}. \tag{18}$$

Here m is the original image, m' is the corrupted image, and $v \in (0, 1)$ controls the negative coefficient of the corrupted-image branch. The objective prefers the chosen response under the original image over the same response under the corrupted image.

Table 7: Ablation results of VGDPO with different similarity thresholds η. The best and second-best results are highlighted in boldface and underlined, respectively.

VGDPO	AMBER				MMHal-Bench		HallusionBench			POPE	
	CHAIR↓	Cover↑	Hal↓	Cog↓	Score↑	Hal↓	qAcc↑	fAcc↑	aAcc↑	Adv.↑	All↑
$\eta = 0.6$	**3.7**	49.0	**17.7**	**1.8**	<u>2.77</u>	<u>49.0</u>	**3.73**	10.40	**38.26**	84.1	85.4
$\eta = 0.7$	**3.7**	49.3	<u>18.0</u>	**1.8**	**2.89**	**45.8**	<u>3.51</u>	**11.27**	<u>37.11</u>	<u>84.6</u>	<u>86.2</u>
$\eta = 0.8$	<u>4.0</u>	**49.6**	19.4	<u>2.0</u>	2.70	50.0	3.29	9.53	36.66	**84.7**	**86.3**
$\eta = 0.9$	<u>4.0</u>	<u>49.5</u>	19.5	<u>2.0</u>	2.66	52.1	<u>3.51</u>	<u>10.98</u>	36.75	84.4	85.8

For the original-image branch, taking the derivative gives

$$\frac{\partial \mathcal{L}_{\mathrm{VC}}}{\partial \pi_\theta(y_w \mid m, x)} = (\sigma(r_v) - 1) \cdot \frac{\beta}{\pi_\theta(y_w \mid m, x)}. \tag{19}$$

For the corrupted-image branch, we have

$$\frac{\partial \mathcal{L}_{\mathrm{VC}}}{\partial \pi_\theta(y_w \mid m', x)} = (1 - \sigma(r_v)) \cdot \frac{\beta(1 - v)}{\pi_\theta(y_w \mid m', x)}. \tag{20}$$

The original-image gradient has the same direction as the chosen branch in the visual guided preference loss, so it supplies additional positive supervision for y_w. The corrupted-image branch suppresses the same response when visual evidence is removed, which encourages the model to distinguish intact visual input from degraded visual input. Meanwhile, this branch is explicitly multiplied by $\beta(1 - v)$, so increasing v relaxes the negative update on the corrupted-image branch and avoids overly strong suppression under a low-confidence corrupted visual condition.

Together, the visual guided DPO term and the visual contrastive term are complementary. The former relaxes harmful negative updates on rejected responses, while the latter supplements the chosen response gradient and improves sensitivity to visual evidence.

C Experimental Details

Backbones. We evaluate VGDPO on five backbones: LLaVA-1.5-7B, LLaVA-1.5-13B [26], Muffin-13B [32], Qwen2.5-VL-3B [34] and Qwen3.5-2B [35]. Specifically, LLaVA-1.5 uses CLIP ViT-L-336px [51] as the vision encoder and Vicuna [52, 53] as the language backbone. Muffin-13B uses BEiT-3 [54, 55] as the vision encoder and Vicuna-13B as the language backbone. These five backbones cover different model sizes and vision-language architectures, allowing us to evaluate the effectiveness and robustness of VGDPO across diverse multimodal settings.

Baseline Implementations. For the controlled baselines, DPO and Anchor use the same training setting as VGDPO, including $\beta = 0.1$, a learning rate of 1×10^{-6}, a batch size of 64, and 2 training epochs. For CW-DPO, we implement its two-stage procedure on the same 6k training data following the original paper: stage 1 smoothing SFT for 2 epochs, followed by stage 2 confidence-weighted DPO for another 2 epochs. The DPO training hyperparameters in stage 2 are the same as VGDPO.

Benchmarks. We evaluate on four multimodal hallucination benchmarks. **1) AMBER** [36] provides object annotations for 1,004 images. We report object hallucination (CHAIR), coverage, response level hallucination, and cognitive hallucination. **2) MMHal-Bench** [27] contains 96 images across 12 categories. We report information score and hallucination rate, evaluated by GPT 4.1 [56]. **3) HallusionBench** [14] evaluates visual illusion and knowledge hallucination with 346 images and 1,129 prompts. We report question level accuracy (qAcc), figure level accuracy (fAcc), and overall accuracy (aAcc), using GPT 4.1 mini [56] to reduce API cost. **4) POPE** [37] is a binary object hallucination benchmark. We report adversarial set accuracy and overall accuracy.

D Additional Ablation Results

Tables 7 and 8 give the full quantitative ablation results for the similarity threshold η and the reweighting strength λ. All experiments use LLaVA-1.5-7B and follow the same training setup as the

Table 8: Ablation results of VGDPO with different reweighting strengths λ. The best and second best results are highlighted in boldface and underlined, respectively.

VGDPO	AMBER				MMHal-Bench		HallusionBench			POPE	
	CHAIR↓	Cover↑	Hal↓	Cog↓	Score↑	Hal↓	qAcc↑	fAcc↑	aAcc↑	Adv.↑	All↑
$\lambda = 0.0$	**3.3**	48.7	**15.8**	**1.4**	2.76	49.0	2.85	10.11	36.75	84.2	85.8
$\lambda = 0.2$	3.9	49.0	<u>17.5</u>	<u>1.7</u>	2.66	50.0	<u>3.07</u>	10.40	37.02	84.0	85.3
$\lambda = 0.4$	<u>3.7</u>	<u>49.7</u>	17.7	1.8	<u>2.81</u>	<u>46.9</u>	**3.51**	10.11	37.11	**84.6**	<u>86.1</u>
$\lambda = 0.6$	<u>3.7</u>	49.3	18.0	1.8	**2.89**	**45.8**	**3.51**	**11.27**	37.11	**84.6**	**86.2**
$\lambda = 0.8$	<u>3.7</u>	**49.8**	18.4	1.8	2.69	49.0	<u>3.07</u>	10.40	<u>37.46</u>	<u>84.3</u>	85.7
$\lambda = 1.0$	4.5	<u>49.7</u>	21.2	2.3	2.79	49.0	<u>3.07</u>	<u>10.69</u>	**37.64**	**84.6**	<u>86.1</u>

main experiments. These two hyperparameters affect different parts of VGDPO: η determines which rejected phrases are treated as potentially erroneous, while λ controls how strongly the detected risk changes the negative coefficient.

Table 7 shows that η controls the strictness of phrase matching. A smaller value such as 0.6 marks fewer phrases as potentially erroneous, so the method behaves closer to vanilla DPO and may miss some hallucination-related phrases. A larger value such as 0.9 marks more phrases as potentially erroneous and relaxes the rejected branch more often, which can weaken hallucination suppression. The best overall result appears at $\eta = 0.7$, where hallucination localization and semantic tolerance are better balanced.

Table 8 shows the effect of λ. A smaller value such as 0.0 keeps β_l^- larger and preserves stronger suppression of rejected responses, which is useful for low-risk rejected samples but may be insufficient for high-risk rejected samples. A larger value such as 1.0 reduces the negative coefficient more aggressively, which can protect high-risk rejected samples from squeezing but may relax low-risk rejected samples too much and weaken preference discrimination. The best overall balance is obtained at $\lambda = 0.6$, supporting the need for moderate rather than uniform relaxation.

E Case Study

Figures 9–14 show qualitative examples of model responses, where hallucinated content is marked in red and more accurate or richer VGDPO outputs are marked in green. Compared with the base models and DPO, VGDPO produces fewer hallucinated descriptions, showing that relaxing high-risk rejected updates and strengthening visual supervision can improve grounding. These cases further demonstrate the effectiveness of VGDPO in improving visual grounding without sacrificing response informativeness.

F Broader Impacts

This work may improve visual grounding and reduce hallucination in MLLMs, which can benefit downstream visual question answering, educational assistance, accessibility tools, human–AI interaction, and other image-based assistance scenarios. However, stronger multimodal generation can also be misused, intentionally or unintentionally, to produce more convincing but incorrect content, and models trained with VGDPO may still make mistakes on ambiguous images, fine-grained visual details, or knowledge-intensive queries. Therefore, hallucination mitigation should not be treated as a guarantee of factual correctness or as a substitute for human verification in high-stakes settings. Practical deployment should combine methods such as VGDPO with careful evaluation, user-facing uncertainty communication, and safeguards against misleading visual content.



Describe this image.

LLaVA-1.5-7B

The image features **a clear blue sky** with two colorful kites flying high in the air. The kites are shaped like comic book characters, with one resembling Spiderman and the other resembling Superman. The kites are positioned close to each other, creating a visually appealing scene.
In the background, there is a person standing on the ground, likely enjoying the sight of the kites soaring in the sky. The overall atmosphere of the image is lively and fun, capturing the essence of outdoor leisure activities.

+DPO

The image features a beautiful scene of **two Spider-Man kites** flying high in the sky against a backdrop of a cloudy blue sky. The kites are designed to resemble the popular comic book character, Spider-Man, adding a fun and playful element to the scene. The kites are soaring gracefully in the air, creating a sense of excitement and enjoyment for those flying them or watching the spectacle. The combination of the kites and the cloudy sky creates a visually appealing and dynamic atmosphere.

+VGDPO (Ours)

The image features a cloudy blue sky with two large, colorful kites flying high in the air. The kites are shaped like comic book characters, adding a fun and vibrant touch to the scene. One kite is positioned more towards the left side of the sky, while the other is located more towards the center. The kites appear to be enjoying the breeze, soaring gracefully in the sky.

Figure 9: Case study on LLaVA-1.5-7B, comparing responses from the base model, DPO, and VGDPO. Hallucinated content is highlighted in **red**, and more accurate or richer VGDPO output is highlighted in **green**.



Describe this image.

LLaVA-1.5-7B

The image features a computer desk with a large computer monitor sitting on top of it. The monitor displays an Apple logo, indicating that it is an Apple computer. Next to the monitor, there is a keyboard and a mouse on the desk, ready for use.
In addition to the main computer setup, **there are two other keyboards and two mice placed on the desk**. This suggests that the desk might be used by multiple people or for various tasks that require multiple input devices.

+DPO

The image features a desktop computer setup with a large monitor prominently displaying an Apple logo. The monitor takes up a significant portion of the screen, making it the focal point of the scene. The Apple logo appears to be a colorful and visually appealing design.
The computer setup includes a keyboard placed in front of the monitor, which is likely used for input and navigation. The overall appearance of the setup suggests a well-organized and functional workspace.

+VGDPO (Ours)

The image features a large computer monitor sitting on a desk, displaying an Apple logo on its screen. The monitor is positioned towards the center of the desk, and it appears to be the main focus of the scene.
In addition to the monitor, there is a keyboard and a mouse placed on the desk, indicating that the setup is ready for use. **The keyboard is located towards the left side of the desk, while the mouse is positioned towards the right side**. The overall scene suggests a well-equipped workspace for computer-related tasks.

Figure 10: Case study on LLaVA-1.5-7B, comparing responses from the base model, DPO, and VGDPO. Hallucinated content is highlighted in **red**, and more accurate or richer VGDPO output is highlighted in **green**.



Describe this image.

LLaVA-1.5-7B

The image features a black dog sitting on a sidewalk next to a bicycle. The dog is wearing a yellow bandana, and it appears to be tied to the bicycle. The bicycle is parked on the side of the road, and another bicycle can be seen further back in the scene.
In addition to the dog and bicycles, **there is a person visible in the background, possibly the dog's owner or a passerby**. The scene captures a moment of everyday life where a dog is tied to a bicycle on the sidewalk.

+DPO

The image features a black dog sitting on the sidewalk next to a bicycle. The dog appears to be tied to the bicycle, possibly for safety or to keep it close to its owner. The bicycle is an old-fashioned one, adding a sense of nostalgia to the scene. The dog seems to be patiently waiting, possibly for its owner to return or for someone to pass by. **The scene captures a moment of companionship between the dog and the bicycle.**

+VGDPO (Ours)

The image features a black dog sitting on a sidewalk next to a bicycle. The dog is wearing a yellow bandana, adding a touch of color to the scene. The bicycle is parked close to the dog, occupying a significant portion of the sidewalk.
In the background, **there is another bicycle further away from the dog and the main bicycle**. The scene appears to be a casual, everyday moment captured on the sidewalk.

Figure 11: Case study on LLaVA-1.5-7B, comparing responses from the base model, DPO, and VGDPO. Hallucinated content is highlighted in **red**, and more accurate or richer VGDPO output is highlighted in **green**.



Describe this image.

Muffin-13B

In the image, there is a person wearing an orange jumpsuit and a black helmet riding a red motorcycle on a track. They are skillfully leaning into a turn, showcasing their expertise in motorcycle racing. **Several spectators can be seen in various positions around the track, watching the race and cheering on the rider. There are at least eight people visible**, spread out across different areas of the scene, focused on the ongoing competition.

+DPO

In the image, there is a person wearing an orange outfit, riding a red motorcycle on a street near a fence. The motorcycle appears to be speeding or taking a sharp turn, as the rider leans into it closely. **Several other motorcycles are visible in the background**, possibly indicating that the scene takes place near a racing track or gathering of motorcycle enthusiasts.

+VGDPO (Ours)

In the image, there is a person wearing an orange jumpsuit and a helmet, riding a red motorcycle on a track. They are leaning down close to the ground while making a sharp turn, showcasing their skill and agility on the bike. The scene captures the excitement of motorcycle racing and the rider's focus on navigating the track successfully.

Figure 12: Case study on Muffin-13B, comparing responses from the base model, DPO, and VGDPO. Hallucinated content is highlighted in **red**, and more accurate or richer VGDPO output is highlighted in **green**.



Describe this image.

Muffin-13B

In the image, there is a cat laying on a desk, sleeping peacefully with its head resting on a keyboard. **The cat occupies the left side of the scene**, while the keyboard spans across the entire width of the desk.
A computer mouse can be found towards the right side of the desk, close to the edge. Additionally, **there are several books scattered around the desk in various positions**, adding elements of interest to the scene.

+DPO

The image features a brown tabby cat resting on a computer keyboard in a relaxed manner. The cat is lying down with its head and paw propped up on the keyboard, seemingly napping. Next to the keyboard, there is a mouse which the cat could have been playing with earlier. The scene takes place on a desk, where **a TV can also be seen nearby**, adding to the domestic setting.

+VGDPO (Ours)

The image features a cat resting on a computer desk near a keyboard. The cat is laying down and has its head close to the keyboard, possibly dozing off or just relaxing. The keyboard covers most of the desk, and a computer mouse can be seen towards the right side of the desk. The cat's presence adds a sense of comfort and companionship to the workspace.

Figure 13: Case study on Muffin-13B, comparing responses from the base model, DPO, and VGDPO. Hallucinated content is highlighted in **red**, and more accurate or richer VGDPO output is highlighted in **green**.



Describe this image.

Muffin-13B

The image features a red and black train traveling on a track that runs along a mountain side, giving a sense of climbing a hill. The train appears small in size due to the aerial view, but it is nevertheless making its way through the mountains.
In addition to the train, **there are several people scattered throughout the scene, possibly hikers or tourists enjoying the beautiful mountain landscape. Some of them are located closer to the train while others can be seen further away, likely admiring the views or exploring the area.**

+DPO

The image features a red train with a black roof traveling on a track that appears to be on top of a mountain or near hills. The train is situated in a scenic area, providing picturesque views of the landscape. **There are a few people in the scene, likely enjoying the view or possibly partaking in some outdoor activities like hiking or sightseeing. One person can be seen towards the left side of the image, while three others are located closer to the right side, further ahead along the track.**

+VGDPO (Ours)

The image features a red trolley train traveling on a track that is situated along a mountain top. The train appears to be climbing uphill, offering picturesque views of the valley below. This scenic location makes it an ideal spot for tourists to enjoy the breathtaking landscape while riding the trolley train.

Figure 14: Case study on Muffin-13B, comparing responses from the base model, DPO, and VGDPO. Hallucinated content is highlighted in **red**, and more accurate or richer VGDPO output is highlighted in **green**.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction match the paper's core claims: the squeezing effect analysis, the VGDPO method, and the empirical results on hallucination benchmarks. The scope and limitations stated in the introduction and conclusion are consistent with the experiments and appendix.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Appendix B provides the gradient derivations for the rejected-branch reweighting and the visual contrastive loss, along with the assumptions used in the analysis. These derivations support the design choices described in the main paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendix C specify the datasets, backbone models, data sampling, optimizer, learning rate, batch size, number of epochs, tuning ranges, evaluation metrics, and test protocols needed to reproduce the main results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The paper describes the sources of training data and implementation details. The relevant implementation code and instructions are publicly available on GitHub at `https://github.com/Liuwq-bit/VGDPO`.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix C specify the model backbones, dataset size, optimizer, learning rate, batch size, number of epochs, and hardware setup used for the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We do not report error bars or statistical significance tests. The main claims are supported by results across five backbones and four benchmarks, but the paper does not quantify run-to-run variability.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4.1 specifies the model backbones, dataset size, learning rate, batch size, number of epochs, and hardware setup used for the experiments. The main experiments were run on 8 NVIDIA RTX PRO 6000 GPUs, which provides the key compute context for reproduction.

   Guidelines: We provide the specific computation resources used for experiments in Section 4.1.

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public benchmarks and pretrained models, does not involve human subjects, and follows the standard evaluation and reporting practices of the field. We do not identify any aspect that conflicts with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the broader impacts of our work in Section F.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release a new high-risk model or a scraped dataset in this paper. The experiments rely on public benchmarks and pretrained backbones, so no additional misuse safeguards are required for the released assets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and use the public datasets, pretrained models, and codebases listed in the paper according to their original licenses and terms of use. The implementation and evaluation follow the documented usage conditions of these assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We will release the training and evaluation code, together with documentation for preprocessing, hyperparameters, and benchmark evaluation, to support reproduction of the reported results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not include any crowdsourcing experiment or human-subject study.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human participants, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We do not use LLMs as part of the core method. LLMs are used only for language polishing and as automatic evaluators for standard benchmark scoring, without affecting the methodology or experimental results.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.