Context Blindness in DPO: Mitigating Object Hallucination in MLLMs via Context-Calibrated Preference Optimization

Byungoh Ko[1], Jinyoung Park[2], Jongha Kim[2], Jeehye Na[2], Jaewon Cho[2], and Hyunwoo J. Kim[2]*

[1] Korea University, Seoul, Republic of Korea
ko990128@korea.ac.kr
[2] KAIST, Daejeon, Republic of Korea
hyunwoojkim@kaist.ac.kr

Abstract. Multimodal large language models (MLLMs) have made rapid progress, yet they still exhibit object hallucination, generating plausible but incorrect descriptions that are inconsistent with the visual input. Direct Preference Optimization (DPO) mitigates this by training models to prefer non-hallucinated responses over hallucinated ones, and recent efforts further enrich the preference data with relevant context. However, it remains unclear whether DPO actually leverages such context. To investigate this, we propose Contextual Preference Gain (CPG), a simple metric that measures how much a model's preference strengthens when relevant context is provided. We find that higher CPG consistently corresponds to lower hallucination, yet standard DPO and its variants exhibit only limited CPG, indicating that they underutilize contextual information and thus remain prone to hallucination. To address this, we propose Context-Calibrated DPO (C^2-DPO), which directly maximizes CPG while preserving the original preference ordering. Across multiple benchmarks, C^2-DPO substantially reduces hallucination without compromising general reasoning, relatively reducing the Object HalBench hallucination rate of Qwen2-VL-Instruct-2B by 36%. Code is available at https://github.com/mlvlab/C2-DPO

Keywords: Multimodal Large Language Models · Object Hallucination · Preference Optimization

1 Introduction

Large language models (LLMs) [2, 13, 28, 29, 45] have shown remarkable performance across a wide range of language tasks. Building on these advances, multimodal large language models (MLLMs) [5, 11, 20, 23, 24, 41] integrate visual encoders and multimodal training data, enabling unified understanding of images and text. Recent MLLMs achieve impressive results on core vision-language tasks such as image captioning [1, 22] and visual question-answering [3, 14, 26, 37, 47].

* Corresponding author

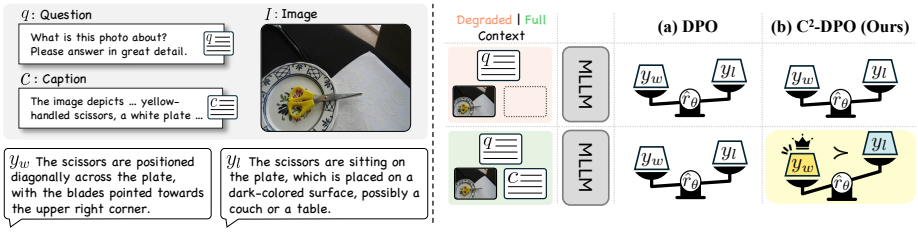


Fig. 1: Motivation of C^2-DPO. A model should yield higher preference scores when richer context is provided. However, standard DPO shows minimal change between full and degraded contexts, revealing *context-blind* behavior. C^2-DPO amplifies this gap, enabling models to better leverage contextual information and reduce hallucination.

Despite the progress, MLLMs remain prone to object hallucination, where the model generates semantically plausible descriptions that are not grounded in the visual input. This issue undermines the factual reliability of multimodal reasoning. To mitigate object hallucination, recent works have applied preference optimization, which trains the model to prefer non-hallucinated responses rather than hallucinated responses. Early studies [30,48,49] primarily focused on constructing high-quality preference data for preference optimization [34]. More recently, a follow-up study [33] enriches the input with non-hallucinated descriptions to strengthen contextual grounding, thereby making preferred and dispreferred responses more discriminative and improving alignment performance.

While these approaches modify the preference dataset for better optimization, a more fundamental question remains unanswered: "Does DPO encourage a model to leverage relevant context and learn contextual preference?" Motivated by the information-theoretic view that relevant information reduces uncertainty [7, 38], we argue that effective preference optimization should strengthen the model's preference for the correct response when richer contextual information is available. Structurally, DPO optimizes the preference margin for a single fixed input, so its gradient never rewards a larger margin when richer context is supplied. To investigate this, we introduce Contextual Preference Gain (CPG), a metric that measures how much the model's preference strengthens when an additional helpful context is provided. Using CPG, we identify two notable observations. First, CPG is strongly and negatively correlated with hallucination rate: models with higher CPG consistently achieve lower hallucination rates. Second, standard DPO and its variants show CPG distributions concentrated near zero, suggesting that their learned preferences are *context-blind*. Together, these findings indicate that existing preference optimization approaches inadequately exploit contextual information for grounding.

To address this limitation, we propose Context-Calibrated Direct Preference Optimization (C^2-DPO), a context-aware preference alignment framework designed to amplify grounding signals for mitigating object hallucination in MLLMs. C^2-DPO explicitly models how helpful context influences the relative preference between preferred and dispreferred responses, and encourages higher

preference scores of the input with the additional relevant context. Specifically, a calibration term directly rewards the preference margin obtained *with* context over the one obtained without it, turning contextual grounding into an explicit training signal rather than an emergent side effect. By calibrating preference scores with respect to contextual information, C^2-DPO enables MLLMs to better leverage input context and reduce object hallucination.

Comprehensive experiments across diverse benchmarks demonstrate that C^2-DPO effectively reduces object hallucinations while preserving general reasoning and instruction-following capabilities. On Object Halbench [35], hallucination rates decrease by 36% and 60% in response and mention level, respectively, with consistent improvements observed on HallusionBench [15]. Furthermore, C^2-DPO maintains its general reasoning performance on ScienceQA, MM-Vet, and TextVQA [26, 37, 47], indicating that mitigating object hallucination does not come at the cost of overall downstream task performance. Beyond the standard DPO, we further show that our context-aware preference gain formulation can be seamlessly incorporated into other preference-optimization variants such as SimPO [27] and RDPO [32], consistently yielding similar gains. Additionally, experiments in text-only settings confirm that C^2-DPO enhances factual alignment even without visual inputs, suggesting that context-aware preference modeling offers broad benefit beyond multimodal domains.

Our contributions are summarized as follows:

- We introduce Contextual Preference Gain (CPG), a diagnostic metric that quantifies how preference margins change with additional grounding information, and reveal a strong correlation between CPG and hallucination rate, exposing a *context blindness* phenomenon in existing DPO-style objectives.
- We propose C^2-DPO, a general preference alignment framework that explicitly optimizes contextual preference gain and integrates seamlessly with various preference optimization methods such as DPO, SimPO, and RDPO.
- We demonstrate the effectiveness and generality of C^2-DPO through extensive experiments across multimodal and text-only benchmarks, achieving substantial reduction in object hallucination while preserving the general reasoning performance.

2 Related Works

2.1 Hallucination in MLLMs

Multimodal large language models (MLLMs) often suffer from the "object hallucination problem", where generated responses appear natural but deviate from input conditions [35, 50]. To mitigate this, several inference-time approaches based on Contrastive Decoding (CD) [21] have been proposed [8, 10, 18, 19], controlling the model behavior during generation by contrasting multiple outputs. Still, CD-based approaches remain post-hoc remedies that fail to fundamentally correct the model's intrinsic behavior, yielding unsatisfactory performance. Also, they require additional inference cost as multiple forwardings are required during

generation. In parallel, Preference Optimization (PO) methods aim to directly train MLLMs to be less hallucinatory [16,30,33,46,48,49]. However, prior works focus on constructing better preference datasets, overlooking context blindness, a key limitation that prevents MLLMs from fully leveraging multimodal cues. Crucially, these methods alter *what* data the model sees but leave the optimization objective unchanged, so the enriched context is not guaranteed to be exploited during alignment. We identify this issue and propose a solution that further enhances the alignment with the same training data.

2.2 Direct Preference Optimization (DPO)

Reinforcement Learning from Human Feedback (RLHF) aligns model behavior with human preferences [9, 31, 36, 39]. By optimizing models toward preference data, RLHF has proven effective in reducing hallucination in MLLMs [16,30,33, 46, 48, 49]. Among its variants, Direct Preference Optimization (DPO) [34] has gained significant attention for its simplicity, removing the need for an explicit reward model. However, recent studies have revealed several limitations of DPO, including length bias [27], suboptimality of reference models [27, 42, 44], overfitting [4], and inaccurate reward estimation [43]. These efforts refine how preference is modeled for a *fixed* input, *e.g.*, by reshaping the reward margin or removing the reference model. Far less attention has been paid to how the preference should respond when the input context itself varies, which is precisely the regime where multimodal grounding matters. In this work, we highlight another overlooked issue, *context blindness*, where MLLMs trained with existing preference objectives fail to fully utilize or comprehend input contexts. To address this, we introduce a new DPO-based training objective that explicitly strengthens contextual understanding in multimodal alignment, offering a principled approach to further mitigate hallucination, complementary to dataset-scaling strategies.

3 Preliminaries

We briefly summarize Direct Preference Optimization for mitigating object hallucination in multimodal large language models (MLLMs).

Object hallucination in MLLMs refers to the generation of text that does not align with the visual content of the input image. To address this, preference alignment such as Direct Preference Optimization (DPO) [34] has emerged as a promising approach for mitigating the object hallucination problem. The preference alignment method is designed to train a policy model π_θ to reflect human preferences over candidate responses. Given an input x and a pair of responses (y_w, y_l) where y_w is the preferred sample while y_l represents the dispreferred sample, DPO optimizes the model to assign higher preference to y_w than y_l.

One of the representatives of preference modeling is the Bradley–Terry (BT) framework [6], which formulates pairwise preferences as probabilistic outcomes based on differences in underlying rewards. In the BT model, the probability of

preferring y_w over y_l under context x is defined as

$$p\left(y_w \succ y_l \mid x\right) = \sigma\left(r\left(x, y_w\right) - r\left(x, y_l\right)\right), \tag{1}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $r\left(x, y\right)$ is the reward associated with response y under context x.

Unlike the BT model, which requires an explicit reward function, DPO introduces an implicit reward $\hat{r}_\theta$ parameterized by the policy model itself:

$$\hat{r}_\theta\left(x, y\right) = \beta \log \frac{\pi_\theta\left(y \mid x\right)}{\pi_{\text{ref}}\left(y \mid x\right)}, \tag{2}$$

where π_θ is the trainable policy model, π_{ref} is a frozen reference model, and $\beta > 0$ controls the regularization strength. Substituting this implicit reward into the BT formulation yields the DPO loss:

$$\begin{aligned} \mathcal{L}_{\text{DPO}}(x, y_w, y_l) &= -\log p(y_w \succ y_l \mid x) \\ &= -\log \sigma\left(\hat{r}_\theta\left(x, y_w\right) - \hat{r}_\theta\left(x, y_l\right)\right) \\ &= -\log \sigma\left(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\right). \end{aligned} \tag{3}$$

Thus, DPO can be viewed as maximizing the probability that the model's pairwise preferences align with human judgments, while avoiding the need for an explicit reward model.

Following other DPO works [16, 30, 33, 46, 48, 49] for mitigating object hallucination, we consider $x = (v, q, c)$ as the input, where v denotes the input image, q is the text query, and c represents the auxiliary image description helpful for understanding the input image (*e.g.,* caption). Here, we treat a non-hallucinated response as the preferred response y_w and a hallucinated response as the dispreferred response y_l. Notably, the DPO objective in Eq. (3) treats x as a monolithic input and is agnostic to the role of c: it never explicitly requires that providing c alter the preference between y_w and y_l. This gap is precisely what we analyze in Section 4.2. For simplicity, since each training instance always includes (y_w, y_l), we denote the DPO loss as $\mathcal{L}_{\text{DPO}}(x)$ hereafter.

4 Method

4.1 Overview

Our work aims to optimize multimodal large language models (MLLMs), via preference optimization, to generate outputs that align more closely with the input context. Most existing approaches [16, 30, 33, 46, 48, 49] modify the preference dataset, either by rewriting sentences with an external model or by appending enriched image descriptions to the input prompt. Such modifications make the contrast between hallucinated and non-hallucinated objects more explicit, helping the model understand the preference gap.

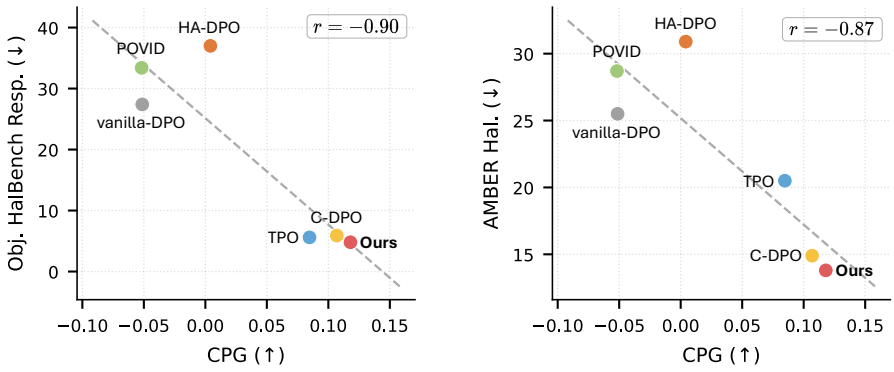


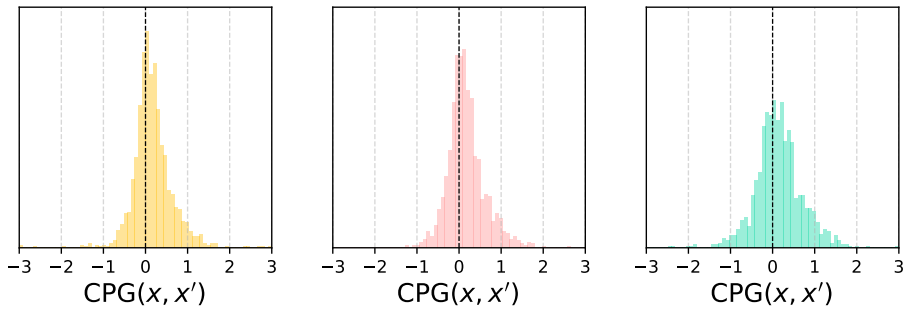
Fig. 2: Correlation between CPG and hallucination rates. We analyze the relationship between Contextual Preference Gain (CPG) and hallucination rate across multiple models. On both Object HalBench (left) and AMBER (right), CPG shows a strong negative correlation with hallucination rate, suggesting that increased CPG is closely associated with improved grounding.

However, recent approaches pay little attention to whether the model genuinely grounds this preference gap between preferred (non-hallucinated) and dispreferred (hallucinated) samples, leveraging the input context such as images or captions. We hypothesize that effective preference optimization should lead MLLMs to increasingly favor preferred response y_w over dispreferred response y_l as additional relevant context information is richer. In other words, the preference gap between y_w and y_l becomes larger as richer contexts (*e.g.*, captions) are given, while maintaining the preference ranking between them. This perspective suggests that the model should not only distinguish responses, but also calibrate the strength of this distinction according to the richness of context. With the preliminary experiments, we found that the standard DPO fails to satisfy these properties, revealing the *context blindness*.

To address this challenge, we present Context-Calibrated Direct Preference Optimization (C^2-DPO), a preference optimization method with a simple calibration term. Our calibration encourages the model to prefer responses grounded in rich context over degraded context by maximizing the contextual preference gain through a contrastive learning objective. This lets multimodal large language models better leverage input context during preference alignment, thereby mitigating object hallucination.

4.2 Observations

To take a deep dive into the preference alignment of the standard preference optimization, we measure the Contextual Preference Gain (CPG) according to the richness of the input. We first introduce the preference score as a measure of alignment strength via implicit reward margin between a preferred response



(a) DPO **(b)** SimPO **(c)** RDPO

Fig. 3: Distributions of Contextual Preference Gain (CPG) for three preference optimization methods: DPO, SimPO, and RDPO. In all cases, the CPG values are highly concentrated near zero, indicating that removing contextual information leads to only minimal changes in preference scores. This demonstrates that existing preference optimization methods are largely *context-blind*.

and a dispreferred response:

$$\Delta\hat{r}_\theta(x, y_w, y_l) = \hat{r}_\theta(x, y_w) - \hat{r}_\theta(x, y_l), \tag{4}$$

which represents how strongly the model prefers the non-hallucinated response y_w over the hallucinated one y_l given input x.

Contextual Preference Gain (CPG). Intuitively, providing additional relevant context should help a model more confidently distinguish a preferred response from a dispreferred one. As a result, the model should exhibit a stronger preference for the correct and grounded response when richer contextual information is available. This intuition aligns with the information-theoretic view that relevant information reduces uncertainty [7,38]. Motivated by this idea, we define Contextual Preference Gain (CPG) as the change in preference score when context is enriched. Given an input x and its degraded counterpart x', CPG is defined as

$$\mathrm{CPG}(x, x') = \Delta\hat{r}_\theta(x, y_w, y_l) - \Delta\hat{r}_\theta(x', y_w, y_l). \tag{5}$$

A positive CPG indicates that richer or more reliable input leads to a stronger preference for y_w over y_l.

We evaluate CPG under a context-removal setting where $x' = (v, q, \varnothing)$ corresponds to the input with the auxiliary caption c removed. As shown in Figure 2, CPG exhibits a strong negative correlation with hallucination rates on Object HalBench [35] and AMBER [40], where models with higher CPG consistently achieve lower hallucination rates. This suggests that CPG captures a model's ability to leverage contextual information for grounding. However, as shown in Figure 3, existing preference optimization methods (DPO, SimPO, and RDPO [27,32,34]) show CPG distributions highly concentrated near zero,

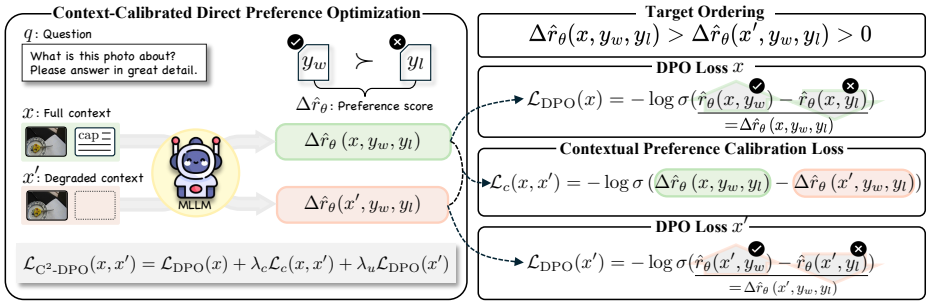


Fig. 4: Context-Calibrated Direct Preference Optimization via Contextual Preference Calibration (C^2-DPO). Given a full context x and a degraded context x', C^2-DPO calibrates the preference score by (1) applying the standard DPO loss on x, (2) contrasting preference scores between x and x' via a Contextual Preference Calibration loss to maximize contextual preference gain, and (3) applying DPO on x' to preserve correct preference ordering under degraded context. This joint objective encourages the model to strengthen preference for grounded responses when richer context is available, effectively reducing hallucination.

indicating that their learned preferences are largely insensitive to contextual information. Given the strong association between CPG and hallucination rate, this *context-blind* behavior reveals a fundamental limitation in existing preference optimization methods.

4.3 Contextual Preference Calibration

Section 4.2 shows that the CPG indicates how well a model grounds its responses in context, and thus how prone it is to hallucinate. Standard preference optimization leaves the CPG concentrated near zero, a failure we term *context blindness*. A context-calibrated model should instead satisfy the ordering

$$\underbrace{\Delta\hat{r}_\theta(x, y_w, y_l)}_{\text{full context}} > \underbrace{\Delta\hat{r}_\theta(x', y_w, y_l)}_{\text{degraded context}} > 0, \tag{6}$$

where each term is the preference score from Eq. (4) under the corresponding context. The first inequality requires the preference score under the full context to exceed the degraded-context one, which is exactly a positive CPG. The second requires it to stay positive on its own, so that y_w remains favored over y_l even without the auxiliary cues. Together, these require the model to preserve the preference ranking across contexts while maximizing the gap as the context becomes richer.

We formulate these desiderata as a primary objective subject to two constraints, taking the full-context DPO loss as the objective and the two inequalities of Eq. (6) as the constraints:

$$\min_\theta \ \mathcal{L}_{\text{DPO}}(x) \quad \text{s.t.} \quad \text{CPG}(x, x') \geq 0, \qquad \Delta\hat{r}_\theta(x', y_w, y_l) \geq 0. \tag{7}$$

Rather than solving Eq. (7) directly, we relax each constraint into a differentiable surrogate that decreases monotonically with its margin, scaled by coefficients $\lambda_c, \lambda_u > 0$. We derive the two surrogates below.

Contextual Preference Calibration Loss. For the first constraint, we contrast the two preference scores with a binary NCE-style objective, treating the full context x as the positive and the degraded context x' as the negative:

$$\begin{aligned}\mathcal{L}_c(x, x') &= -\log \frac{\exp\left(\Delta\hat{r}_\theta\left(x, y_w, y_l\right)\right)}{\exp\left(\Delta\hat{r}_\theta\left(x, y_w, y_l\right)\right) + \exp(\Delta\hat{r}_\theta(x', y_w, y_l))} \\ &= -\log\sigma\left(\Delta\hat{r}_\theta\left(x, y_w, y_l\right) - \Delta\hat{r}_\theta\left(x', y_w, y_l\right)\right) \\ &= -\log\sigma\left(\mathrm{CPG}(x, x')\right). \end{aligned} \tag{8}$$

Minimizing $\mathcal{L}_c$ monotonically increases the CPG, driving the full-context preference score above the degraded-context one. Since a grounded response is favored more strongly as the context grows richer, the model is encouraged to leverage the additional context rather than ignore it.

Preference under the degraded context. For the second constraint, we apply the DPO objective to the degraded context x':

$$\mathcal{L}_{\mathrm{DPO}}(x') = -\log\sigma(\Delta\hat{r}_\theta(x', y_w, y_l)), \tag{9}$$

a monotone surrogate that pushes the degraded-context score positive, keeping y_w favored even when the auxiliary cues are removed. Anchoring this score also prevents $\mathcal{L}_c$ from widening the CPG degenerately by lowering $\Delta\hat{r}_\theta(x', y_w, y_l)$ rather than improving grounding.

Training objective. Combining the primary objective in Eq. (7) with the two surrogate terms gives our full training objective:

$$\mathcal{L}_{\text{C}^2\text{-DPO}}(x, x') = \mathcal{L}_{\mathrm{DPO}}(x) + \lambda_c\,\mathcal{L}_c(x, x') + \lambda_u\,\mathcal{L}_{\mathrm{DPO}}(x'), \tag{10}$$

where $\lambda_c, \lambda_u > 0$ weight the two surrogates against the full-context DPO loss. Jointly optimizing the three terms encourages the model to satisfy the ordering in Eq. (6), thereby increasing CPG and reducing object hallucination. This allows MLLMs to better leverage the input context, improving robustness to hallucination while maintaining a stable preference across input conditions.

5 Experiments

In this section, we empirically investigate the effectiveness of C^2-DPO in enhancing context awareness and reducing hallucinations. We introduce the experimental setup in Sec. 5.1, present the main results in Sec. 5.2, conduct ablation studies in Sec. 5.3, and provide a qualitative analysis in Sec. 5.4.

5.1 Experimental Setup

Models and training settings. We evaluate our method on two representative MLLMs of different scales: LLaVA-v1.5-7B [23] and Qwen2-VL-Instruct-2B [41].

For a fair comparison, we adopt the same training data as C-DPO [33], using the datasets released for LLaVA-v1.5-7B[3] and Qwen2-VL-Instruct-2B[4], respectively. Each preference data consists of an image, a query, and an auxiliary caption, paired with a non-hallucinated (preferred) and a hallucinated (dispreferred) response. For the degraded input x', we remove the auxiliary caption while keeping the image and query, following the setting in Section 4.2; results under alternative degradations (*e.g.*, noisy images, randomly substituted captions) are reported in the supplementary material. We set $\beta = 0.1$, with (λ_c, λ_u) set to $(0.3, 0.5)$ for LLaVA-v1.5-7B and $(0.5, 0.3)$ for Qwen2-VL-Instruct-2B, selected from the $[0.3, 0.5]$ range based on the sensitivity analysis in Section 5.3. All models are trained for one epoch using LoRA [17] with AdamW [25] at a learning rate of 2×10^{-6}. Training is conducted with a global batch size of 64.

Evaluation benchmarks. We evaluate the hallucination rates and general multimodal reasoning ability of our C^2-DPO approach across a wide range of benchmarks. To evaluate hallucination, we use Object HalBench [35], AMBER [40], and HallusionBench [15]. To assess general multimodal reasoning performance, we further evaluate on ScienceQA [26], MM-Vet [47], and TextVQA [37].

Baselines. We compare C^2-DPO with a wide range of state-of-the-art methods. Specifically, VCD [19], OPERA [18], and DoLa [10] apply contrastive decoding strategies, whereas HA-DPO [48], POVID [49], CLIP-DPO [30], RLAIF-V [46], TPO [16], and C-DPO [33] adopt preference optimization approahces. We additionally include a vanilla-DPO baseline, obtained by training with the original DPO [34] objective on the C-DPO dataset [33], without using any auxiliary captions. Further details on training, datasets, and evaluation protocols are provided in the supplementary material.

5.2 Main Results

Comparison with baselines. As shown in Table 1, C^2-DPO delivers consistent and substantial improvements over prior methods across a wide range of hallucination benchmarks. On Object HalBench, our approach achieves the strongest performance among all compared methods. For Qwen2-VL-Instruct-2B, C^2-DPO reduces response-level and mention-level hallucinations to 1.6 and 1.0, outperforming C-DPO [33] by large margins and yielding a 36% and 60% relative reduction, respectively. A similar trend holds on AMBER, where C^2-DPO achieves a hallucination score of 16.1, lower than both vanilla-DPO (39.9) and C-DPO (17.5). The improvements generalize to the larger model as well: on LLaVA-v1.5-7B, our method attains a response-level hallucination of 4.8 and a mention-level score of 2.7, along with clear reductions across all AMBER dimensions. These hallucination reductions come without sacrificing general multimodal reasoning performance, as C^2-DPO maintains strong results on ScienceQA, MM-Vet, and

[3] https://huggingface.co/datasets/psp-dada/SENTINEL/blob/main/LLaVA_v1_5_7b_SENTINEL_8_6k.json

[4] https://huggingface.co/datasets/psp-dada/SENTINEL/blob/main/Qwen2_VL_2B_Instruct_SENTINEL_12k.json

Table 1: Comparison of hallucination mitigation methods in MLLMs. The best and second-best results are highlighted in **bold** and underline, respectively. All comparisons are conducted under identical model size and architecture. "Rsp." and "Men." denote response-level and mention-level hallucination degrees, while "Hal." and "Cog." represent the Hallucination Score and Cognitive Score, respectively.

Model	Method	Hallucination Benchmarks						General Benchmarks		
		Object HalBench		AMBER			HallusionBench	ScienceQA	MM-Vet	TextVQA
		Rsp. ↓	Men. ↓	CHAIR ↓	Hal. ↓	Cog. ↓	Question Acc. ↑	Image Acc. ↑	Overall ↑	Acc. ↑
LLaVA-v1.5-7B [23]	Base	52.7	28.0	8.4	35.5	4.0	46.86	66.8	31.0	**58.2**
	VCD [19]	51.3	25.9	9.1	39.8	4.2	–	68.7	29.8	56.1
	OPERA [18]	45.3	22.9	6.5	28.5	3.1	–	68.2	30.3	**58.2**
	DoLa [10]	44.0	25.1	6.2	27.7	2.9	–	67.5	30.8	56.6
	HA-DPO [48]	37.0	20.9	6.7	30.9	3.3	47.74	**69.7**	30.6	56.7
	POVID [49]	33.4	16.6	5.3	28.7	3.0	46.59	68.8	31.8	56.6
	CLIP-DPO [30]	–	–	3.7	16.6	1.3	–	67.6	–	56.4
	RLAIF-V [46]	7.8	4.2	**2.8**	15.7	**0.9**	35.43	68.2	29.9	55.1
	TPO [16]	5.6	3.2	3.6	20.5	1.6	40.12	67.1	25.7	55.3
	vanilla-DPO [34]	27.4	15.9	5.4	25.5	2.5	**47.83**	69.3	32.0	**58.2**
	C-DPO [33]	5.9	3.3	3.0	14.9	1.3	47.48	69.4	**33.4**	**58.2**
	C^2-DPO	**4.8**	**2.7**	**2.8**	**13.8**	1.2	47.03	69.5	**33.4**	**58.2**
Qwen2-VL-Instruct-2B [41]	Base	16.1	8.3	6.2	35.5	2.7	51.11	76.9	**49.9**	78.2
	vanilla-DPO [34]	6.0	3.6	4.2	39.9	3.2	51.37	77.0	49.8	**78.4**
	C-DPO [33]	2.5	1.6	2.7	17.5	**0.8**	51.55	**77.3**	48.2	**78.4**
	C^2-DPO	**1.6**	**1.0**	**2.7**	**16.1**	0.9	**52.08**	77.2	49.1	**78.4**

Table 2: Ablation on loss components. We ablate the contribution of each loss component in C^2-DPO.

Model	Method			Object HalBench		AMBER		
	$\mathcal{L}_{\text{DPO}}(x)$	$\mathcal{L}_c(x, x')$	$\mathcal{L}_{\text{DPO}}(x')$	Rsp. ↓	Men. ↓	CHAIR ↓	Hal. ↓	Cog. ↓
LLaVA-v1.5-7B	✓			5.9	3.3	3.0	14.9	1.3
	✓		✓	7.1	4.1	3.1	15.9	1.2
	✓	✓		5.9	3.2	3.1	15.0	1.3
	✓	✓	✓	**4.8**	**2.7**	**2.8**	**13.8**	**1.2**
Qwen2-VL-Instruct-2B	✓			2.5	1.6	2.7	17.5	**0.8**
	✓		✓	3.3	2.6	2.7	22.5	1.1
	✓	✓		3.5	2.3	3.2	18.6	1.0
	✓	✓	✓	**1.6**	**1.0**	**2.7**	**16.1**	0.9

TextVQA. Overall, these results highlight that C^2-DPO provides a scalable and model-agnostic method for mitigating object hallucination in MLLMs.

5.3 Ablations and Further Analysis

In this section, we provide a comprehensive analysis of C^2-DPO through ablations and additional studies. We first examine the contribution of each loss component, and then show that our contextual preference calibration generalizes beyond DPO to variants such as SimPO [27] and RDPO [32], and even to text-only LLM settings. We further analyze the learning dynamics of CPG and its behavior under varying context lengths, assess robustness to noisy context, and study the sensitivity to loss coefficients, collectively highlighting the robustness and versatility of our approach.

Table 3: Effectiveness of combining the proposed Contextual Preference Calibration with SimPO and RDPO. The results show that Contextual Preference Calibration generalizes well across preference optimization variants effectively.

Method	Object HalBench		AMBER		
	Rsp. ↓	Men. ↓	CHAIR ↓	Hal. ↓	Cog. ↓
SimPO	3.9	2.6	**2.9**	17.8	1.0
C^2-SimPO	**3.3**	**2.0**	3.0	**14.4**	**0.8**
RDPO	3.4	2.3	2.9	38.7	2.5
C^2-RDPO	**1.7**	**1.2**	**2.5**	**29.0**	**1.6**

Table 4: Performance on AlpacaEval 2 for LLMs. Applying our Contextual Preference Calibration consistently improves both DPO and SimPO.

Method	AlpacaEval 2 [12]	
	LC (%) ↑	WR (%) ↑
DPO	24.1	27.4
C^2-DPO	**25.9**	**29.7**
SimPO	33.5	36.5
C^2-SimPO	**34.1**	**37.4**

Ablation on loss components. We analyze how the two components of the C^2-DPO objective, $\mathcal{L}_c(x, x')$ and $\mathcal{L}_{\text{DPO}}(x')$, affect hallucination on Object HalBench and AMBER. As shown in Table 2, optimizing either component alone leads to weaker performance. In contrast, jointly optimizing both losses produces a clear synergistic effect, yielding substantial improvements over the original baseline.

Effects on different DPO variants. To examine whether the Contextual Preference Calibration approach is specific to DPO or can generalize to various preference optimization schemes, we apply our formulation to two representative alternatives: SimPO [27] and RDPO [32]. As shown in Table 3, incorporating our formulation—denoted as C^2-SimPO and C^2-RDPO—consistently reduces hallucination rates. These results indicate that contextual preference gain is orthogonal to the underlying preference-learning objective and can be seamlessly integrated with different optimization recipes. This general applicability suggests our method provides a broadly compatible improvement to preference-based multimodal alignment, rather than being limited to a DPO-specific remedy.

Applicability on text-only LLMs. To assess whether our Contextual Preference Calibration (CPC) strategy remains effective in text-only LLM settings, we conduct experiments following the experimental protocol of SimPO [27]. Using Qwen2.5-Instruct-1.5B [41], we assess performance on the widely used instruction-following benchmark, AlpacaEval 2 [12]. To adapt CPC to the text-only scenario, we instantiate it following Eq. 10 by replacing both x and x' using only the query, *i.e.*, $x = q, x' = \varnothing$. As shown in Table 4, CPC yields consistent improvements across both DPO and SimPO-based models, highlighting the versatility of our formulation and suggesting that CPC provides a unified and effective strategy for improving reliability in preference optimization.

Learning dynamics of contextual preference gain. Figure 5 illustrates how Contextual Preference Gain (CPG) evolves during training for C^2-DPO compared to the baseline, C-DPO. C-DPO exhibits near-zero or even negative CPG throughout training, indicating that it fails to leverage contextual signals. In contrast, C^2-DPO shows a clear and steady increase in CPG. This pattern suggests that C^2-DPO becomes progressively more sensitive to input context

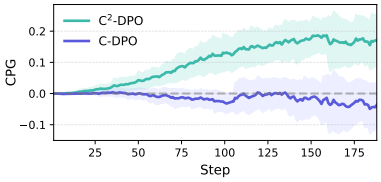


Fig. 5: Contextual Preference Gain (CPG) comparison between C^2-DPO and C-DPO. Our C^2-DPO steadily increases CPG throughout training, while C-DPO stays near or below zero, indicating its limited ability to utilize contextual signals.

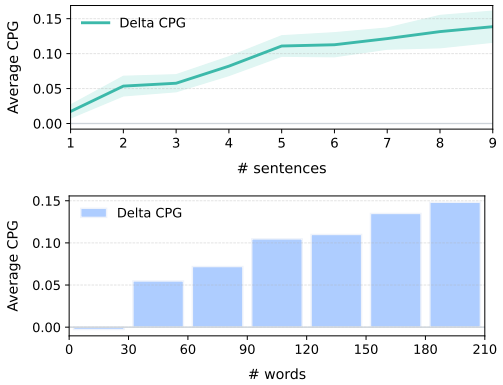


Fig. 6: Average Contextual Preference Gain (CPG) at the sentence (top) and word (bottom) level. CPG increases consistently as more relevant context is added, indicating that richer contextual signals lead to stronger preference separation.

during preference optimization, which in turn enables the model to better align its choices with the provided context and leads to reduced object hallucination.

Contextual preference gain according to the context lengths. Figure 6 illustrates how Contextual Preference Gain (CPG) evolves as the amount of non-hallucinated contextual information increases. The top figure reports the average CPG as a function of the number of sentences added to the input. We observe a clear upward trend: CPG steadily increases with more sentences, indicating that richer contextual grounding amplifies the preference difference between grounded and hallucinated responses. The bottom figure provides a complementary analysis at the word level, showing that CPG also grows consistently as additional words are incorporated. Together, these results demonstrate that stronger contextual signals yield greater preference separation, highlighting the sensitivity of CPG to the granularity and richness of added context.

Robustness to noisy context. We further evaluate robustness when the auxiliary context is noisy. During training, we randomly mask the auxiliary captions c with probability $p \in \{0.0, 0.3, 0.5\}$. Table 5 shows that C^2-DPO consistently achieves lower hallucination rates than C-DPO across all masking levels on Object HalBench. Importantly, the performance gap widens as the masking probability increases, suggesting that contextual preference calibration enables the model to remain robust even when contextual information is partially corrupted.

Sensitivity analysis on loss coefficients. Table 6 presents the sensitivity analysis for λ_c and λ_u, which control the strength of $\mathcal{L}_c(x, x')$ and $\mathcal{L}_{\text{DPO}}(x')$. The results show that the losses are most effective with weights in the 0.3 to 0.5 range, while 0.1 provides comparatively weaker influence. This narrow region of strong performance remains consistent across both models, indicating that C^2-DPO works reliably without extensive hyperparameter tuning.

Table 5: Robustness to context quality.

Method	Noise p	Object HalBench	
		Rsp. ↓	Men. ↓
C-DPO	0.0	2.5	1.6
C^2-DPO		**1.6**	**1.0**
C-DPO	0.3	9.6	5.2
C^2-DPO		**8.5**	**5.0**
C-DPO	0.5	13.0	7.2
C^2-DPO		**8.4**	**4.6**

Table 6: Sensitivity analysis on loss coefficients.

Model	λ	Value	Object HalBench	
			Rsp. ↓	Men. ↓
LLaVA-v1.5-7B	λ_c	0.1	7.4	4.3
		0.3	**4.8**	**2.7**
		0.5	5.5	3.1
	λ_u	0.1	5.5	3.2
		0.3	5.6	3.1
		0.5	**4.8**	**2.7**
Qwen2-VL-Instruct-2B	λ_c	0.1	4.1	2.6
		0.3	2.0	1.3
		0.5	**1.6**	**1.0**
	λ_u	0.1	2.4	1.5
		0.3	**1.6**	**1.0**
		0.5	2.0	1.3

5.4 Qualitative Analysis

Figure 7 shows qualitative comparisons between the baseline model and ours. The baseline exhibits hallucinations ("a hot dog with *mustard*"). Although it correctly identifies the object as a hot dog, it mistakenly describes it as having *mustard*, revealing a failure to attend to fine-grained visual details. In contrast, C^2-DPO produces an accurate description ("the hot dog is covered in *chili*"), showing that it inspects such details more carefully. It also identifies additional cues, such as a "bottle of water", which the baseline overlooks. This reflects our objective, which calibrates preference toward grounded evidence rather than frequent co-occurrences, a common source of hallucination.

Figure 8 further illustrates how each model uses contextual information provided alongside the image. CPG reveals this directly: the baseline C-DPO shows a low score, indicating that it underutilizes the key cue (*e.g.*, "a *bus stop* on the left side of the image"). In contrast, our C^2-DPO attains a much higher CPG score, correctly referencing "the *bus stop*" and grounding its response in the described surroundings. Here CPG tracks observable behavior: its higher value coincides with explicit use of the caption, illustrating at the instance level what Section 4.2 shows in aggregate. By enabling stronger context-aware reasoning, C^2-DPO mitigates hallucinations more faithfully in multimodal understanding.

6 Conclusion

In this work, we investigated whether existing preference optimization methods truly leverage contextual information for grounding, and found that they do not. To quantify this, we introduced Contextual Preference Gain (CPG), which revealed two key findings: CPG is strongly and negatively correlated with hallucination rate, and existing DPO-style methods exhibit *context-blind* behavior with CPG concentrated near zero. Motivated by this, we proposed Context-Calibrated DPO (C^2-DPO), which directly maximizes CPG through a simple

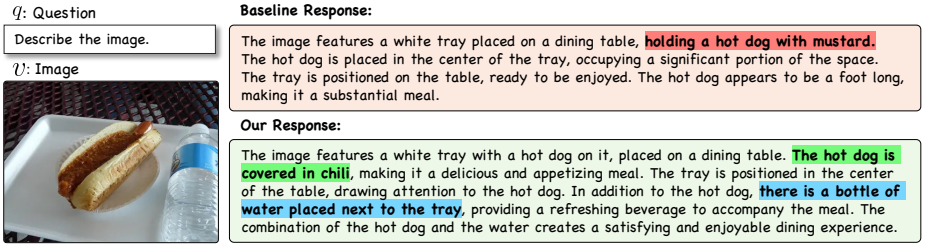


Fig. 7: Qualitative example on detailed captioning. We compare the outputs from models trained with the baseline and our C^2-DPO. Red, green, and blue denote hallucinated details, correct details, and additional correct details captured only by our model, respectively. Visualization is done on AMBER benchmark with LLaVA-1.5-7B.

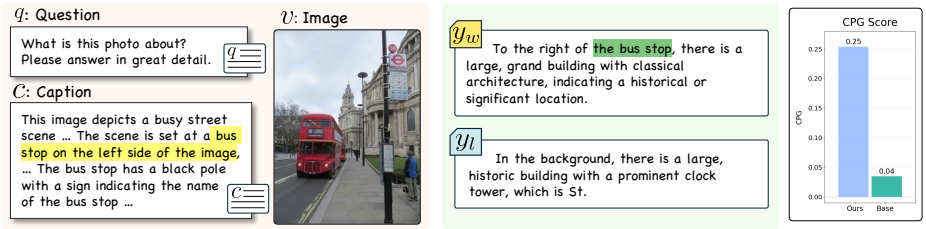


Fig. 8: Qualitative comparison of context utilization between our C^2-DPO and the baseline C-DPO. Yellow and green highlights mark context-relevant details from the provided caption that are correctly reflected in the positive response. The bar chart (right) reports the corresponding Contextual Preference Gain (CPG), which is substantially higher for C^2-DPO.

contrastive regularization while preserving correct preference ordering under degraded context. Extensive experiments demonstrate that C^2-DPO substantially reduces object hallucination without compromising general reasoning ability, and generalizes to other preference optimization variants and text-only settings. Beyond object hallucination, our results suggest that explicitly modeling how context shapes preference is a broadly useful principle for preference optimization. As C^2-DPO relies only on the contrast between a full and a degraded input, it extends naturally to other modalities and tasks where context can be defined, which we leave for future work.

Acknowledgment. This work was supported by the InnoCORE program of the Ministry of Science and ICT (N10250156, 30%), Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korean government(MSIT) (RS-2025-25442149, LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain, 30%), and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government(MSIT) (No. RS-2024-00457882, AI Research Lab Project, 40%).

References

1. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: Nocaps: Novel object captioning at scale. In: ICCV (2019)
2. Anthropic: Claude 3.7 sonnet. `https://www.anthropic.com/news/claude-3-7-sonnet` (2025)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV (2015)
4. Azar, M.G., Guo, Z.D., Piot, B., Munos, R., Rowland, M., Valko, M., Calandriello, D.: A general theoretical paradigm to understand learning from human preferences. In: AISTATS (2024)
5. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv:2309.16609 (2023)
6. Bradley, R.A., Terry, M.E.: Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika (1952)
7. Burgin, M.: The essence of information: Paradoxes, contradictions, and solutions. In: Electronic Conference on Foundations of Information Science: The nature of information: Conceptions, misconceptions, and paradoxes (FIS 2002). Retrieved September (2002)
8. Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: Halc: Object hallucination reduction via adaptive focal-contrast decoding. In: ICML (2024)
9. Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences. In: NeurIPS (2017)
10. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J., He, P.: Dola: Decoding by contrasting layers improves factuality in large language models. In: ICLR (2024)
11. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. In: NeurIPS (2023)
12. Dubois, Y., Galambosi, B., Liang, P., Hashimoto, T.B.: Length-controlled alpacaeval: A simple way to debias automatic evaluators. In: COLM (2024)
13. Google DeepMind: Gemini 2.5. `https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/` (2025)
14. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
15. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: CVPR (2024)
16. He, L., Chen, Z., Shi, Z., Yu, T., Shao, J., Sheng, L.: A topic-level self-correctional approach to mitigate hallucinations in mllms. arXiv:2411.17265 (2024)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
18. Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR (2024)
19. Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR (2024)

20. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. TMLR (2024)
21. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T.B., Zettlemoyer, L., Lewis, M.: Contrastive decoding: Open-ended text generation as optimization. In: ACL (2023)
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2017)
26. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: NeurIPS (2022)
27. Meng, Y., Xia, M., Chen, D.: Simpo: Simple preference optimization with a reference-free reward. In: NeurIPS (2024)
28. Meta AI: Meta-ai. the llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/` (2025)
29. OpenAI: Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/` (2024)
30. Ouali, Y., Bulat, A., Martinez, B., Tzimiropoulos, G.: Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In: ECCV (2024)
31. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. In: NeurIPS (2022)
32. Park, R., Rafailov, R., Ermon, S., Finn, C.: Disentangling length from quality in direct preference optimization. In: ACL Findings (2024)
33. Peng, S., Yang, S., Jiang, L., Tian, Z.: Mitigating object hallucinations via sentence-level early intervention. In: ICCV (2025)
34. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: NeurIPS (2023)
35. Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: EMNLP (2018)
36. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv:1707.06347 (2017)
37. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019)
38. Soni, J., Goodman, R.: A mind at play: how Claude Shannon invented the information age. Simon and Schuster (2017)
39. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.F.: Learning to summarize with human feedback. In: NeruIPS (2020)
40. Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan, M., Zhang, J., et al.: Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv:2311.07397 (2023)
41. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191 (2024)
42. Wu, J., Wang, X., Yang, Z., Wu, J., Gao, J., Ding, B., Wang, X., He, X.: Alphadpo: Adaptive reward margin for direct preference optimization. In: ICML (2025)

43. Xiao, T., Yuan, Y., Zhu, H., Li, M., Honavar, V.G.: Cal-dpo: Calibrated direct preference optimization for language model alignment. In: NeurIPS (2024)
44. Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., Kim, Y.J.: Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. In: ICML (2024)
45. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv:2505.09388 (2025)
46. Yu, T., Zhang, H., Li, Q., Xu, Q., Yao, Y., Chen, D., Lu, X., Cui, G., Dang, Y., He, T., et al.: Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In: CVPR (2025)
47. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: ICML (2024)
48. Zhao, Z., Wang, B., Ouyang, L., Dong, X., Wang, J., He, C.: Beyond multimodal hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. In: ICME (2025)
49. Zhou, Y., Cui, C., Rafailov, R., Finn, C., Yao, H.: Aligning modalities in vision large language models via preference fine-tuning. arXiv:2402.11411 (2024)
50. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. In: ICLR (2024)