Mitigating Hallucination Through Theory-Consistent Symmetric Multimodal Preference Optimization

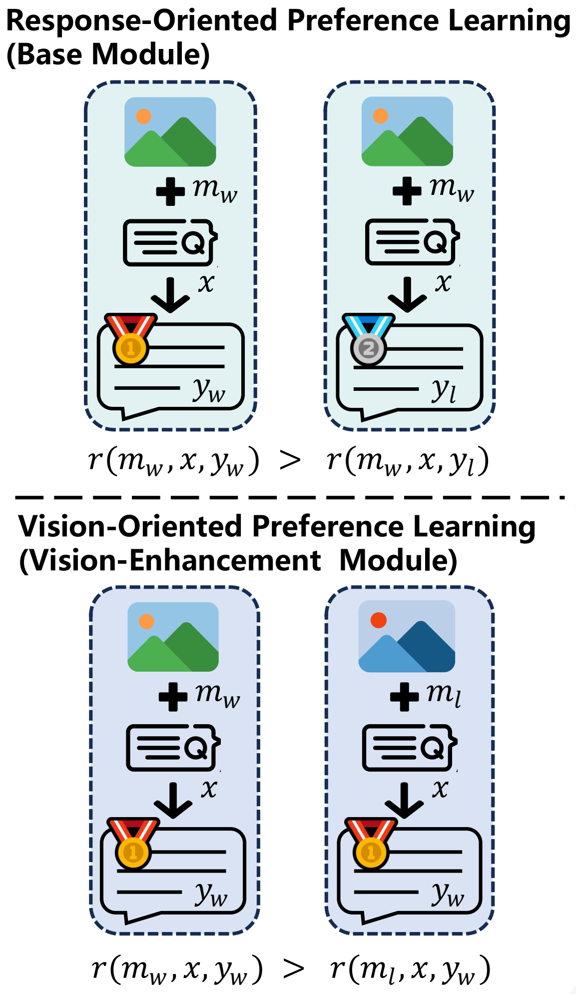
Wenqi Liu[1], Xuemeng Song[2], Jiaxi Li[3], Yinwei Wei[1], Na Zheng[4], Jianhua Yin[1], Liqiang Nie[5]

[1]Shandong University, [2]Southern University of Science and Technology, [3]University of Georgia, [4]National University of Singapore, [5]Harbin Institute of Technology (Shenzhen)

The Challenge: MLLM Hallucination

Multimodal Large Language Models (MLLMs) demonstrate impressive capabilities but often generate outputs that fail to align with the provided image, a phenomenon referred to as **hallucination**.

Direct Preference Optimization (DPO) is a crucial technique for aligning models and mitigating this issue. To better adapt DPO to multimodal tasks, researchers have extended the original **response-oriented preference learning** to incorporate **vision-oriented preference learning**. However, existing approaches to vision-oriented preference learning still suffer from significant unresolved challenges.



Core Limitations of Existing Methods

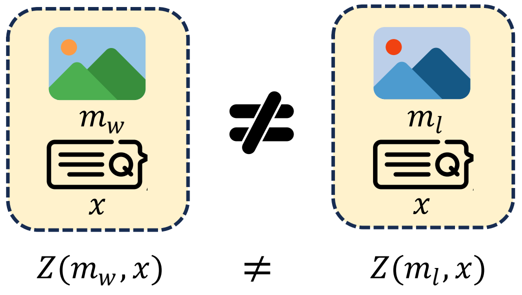
Non-Rigorous Objective

Existing vision-oriented DPO methods compare different images (m_w vs. m_l) but fail to account for canceling out the partition function $Z(m, x)$. This approach is theoretically flawed as it deviates from the standard DPO derivation.

Indirect Supervision

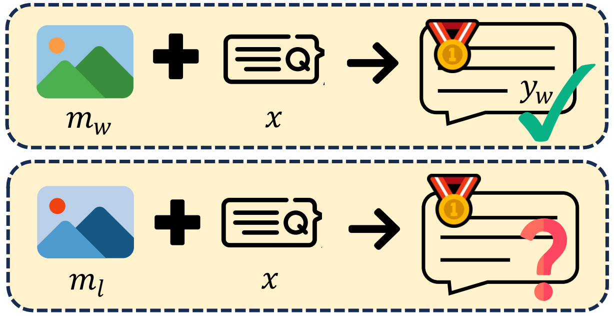
Existing vision-oriented DPO methods contrast images (m_w vs. m_l) while relying on the same response (y_w). This contradicts the fundamental design of DPO, which is explicitly intended to learn from preferences between two responses (y_w vs. y_l).

Limitation1: Non-Rigorous Objective Function



- **Intractable partition functions cannot be eliminated due to the difference in the image input.**

Limitation2: Indirect Preference Supervision



- **Only utilize contrastive images as indirect preference supervision, fail to explore the preferred responses of these images.**

SymMPO

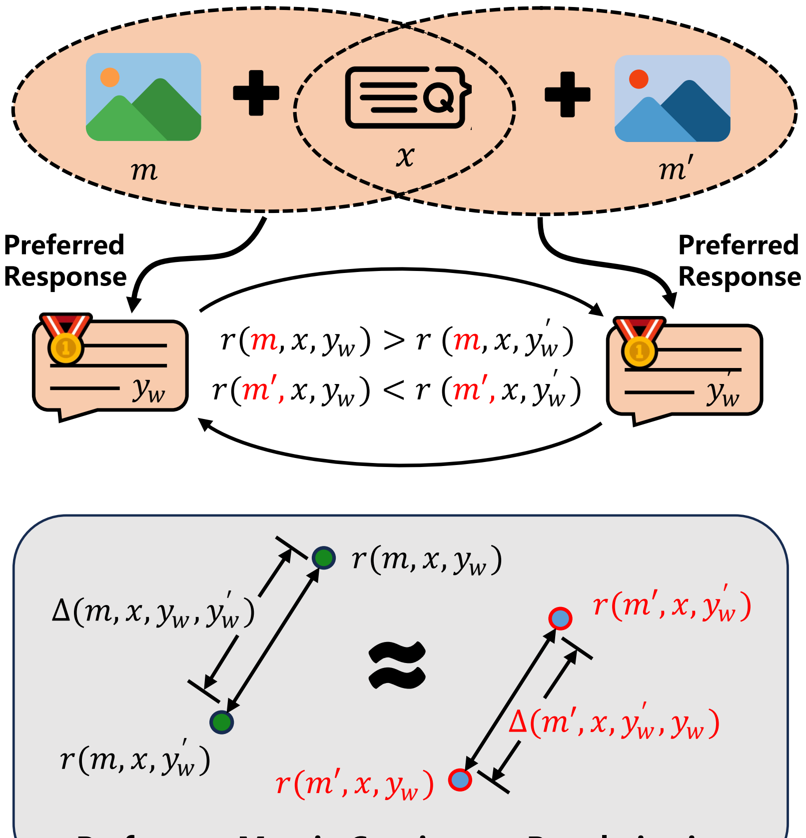
We propose **Sym**metric **M**ultimodal **P**reference **O**ptimization **(SymMPO)**.

Instead of contrasting images, SymMPO contrasts their own preferred responses within a symmetric framework.

- For input (m, x), the preference is $y_w \succ y_w'$.
- For input (m', x), the preference is $y_w' \succ y_w$.

This approach is both **theory-consistent** (correctly cancels the partition functions) and leverages **direct preference supervision**.

Symmetric Multimodal Preference Optimization (SymMPO)



Preference Margin Consistency Regularization

Objective Function

The full loss function integrates standard DPO with our novel symmetric losses, designed to enhance multimodal alignment:

$$\mathcal{L}_{SymMPO} = \mathcal{L}_{DPO_m} + \lambda\mathcal{L}_{Pair} + \gamma\mathcal{L}_{Margin} + \eta\mathcal{L}_{AncPO}$$

- $\mathcal{L}_{DPO_m}$ (Standard DPO) : Aligns response quality using (preferred, less-preferred) pairs.

$$\mathcal{L}_{DPO_m} = -\mathbb{E}_{(x,m,y_w,y_l)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|m,x)}{\pi_{ref}(y_w|m,x)} - \beta\log\frac{\pi_\theta(y_l|m,x)}{\pi_{ref}(y_l|m,x)}\right)\right]$$

- $\mathcal{L}_{Pair}$ (Symmetric Loss) : Serves as the core vision-oriented loss, enforcing symmetric in preference.

$$\begin{aligned}\mathcal{L}_{Pair} &= -\mathbb{E}_{(x,m,m',y_w,y'_w)\sim\mathcal{D}}\left[\log\sigma\left(r(m,x,y_w) - r(m,x,y'_w)\right) + \log\sigma\left(r(m',x,y'_w) - r(m',x,y_w)\right)\right] \\ &= -\mathbb{E}_{(x,m,m',y_w,y'_w)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|m,x)}{\pi_{ref}(y_w|m,x)} - \beta\log\frac{\pi_\theta(y'_w|m,x)}{\pi_{ref}(y'_w|m,x)}\right)\right. \\ &\qquad \left. + \log\sigma\left(\beta\log\frac{\pi_\theta(y'_w|m',x)}{\pi_{ref}(y'_w|m',x)} - \beta\log\frac{\pi_\theta(y_w|m',x)}{\pi_{ref}(y_w|m',x)}\right)\right].\end{aligned}$$

Objective Function

The full loss function integrates standard DPO with our novel symmetric losses, designed to enhance multimodal alignment:

$$\mathcal{L}_{SymMPO} = \mathcal{L}_{DPO_m} + \lambda\mathcal{L}_{Pair} + \gamma\mathcal{L}_{Margin} + \eta\mathcal{L}_{AncPO}$$

- $\mathcal{L}_{Margin}$ (Margin Consistency) : Ensures preference gap remain consistent in both directions.

$$\begin{cases} \mathcal{L}_{Margin} = \mathbb{E}_{(x,m,m',y_w,y'_w)\sim\mathcal{D}}\Big(\Delta(m,x,y_w,y'_w) - \Delta(m',x,y'_w,y_w)\Big)^2, \\ \Delta(m,x,y_w,y'_w) = r(m,x,y_w) - r(m,x,y'_w) = \log\dfrac{\pi_\theta(y_w|m,x)}{\pi_{ref}(y_w|m,x)} - \log\dfrac{\pi_\theta(y'_w|m,x)}{\pi_{ref}(y'_w|m,x)}, \end{cases}$$

- $\mathcal{L}_{AncPO}$ (Anchored Loss) : Stabilizes training by anchoring the likelihood of preferred response.

$$\mathcal{L}_{AncPO} = -\mathbb{E}_{(x,m,m',y_w,y'_w)\sim\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|m,x)}{\pi_{ref}(y_w|m,x)} - \delta\right) + \log\sigma\left(\beta\log\frac{\pi_\theta(y'_w|m',x)}{\pi_{ref}(y'_w|m',x)} - \delta\right)\right]$$

Experiment

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

Model	Data Size	Feedback	HallusionBench			Object-HalBench		MMHal-Bench		AMBER		MMStar
			qAcc↑	fAcc↑	aAcc↑	Resp.↓	Ment.↓	Score↑	Hall↓	Acc↑	F1↑	Overall↑
Muffin-13B [37]	✗	✗	6.15	12.71	41.89	53.0	24.3	2.06	66.7	74.2	80.0	25.4
+RLHF-V [10]	1.4k	Human	9.67	13.87	45.79	8.5	4.9	2.60	56.2	82.0	86.7	31.0
LLaVA-1.5-7B [3]	✗	✗	3.95	11.56	41.71	56.5	27.9	2.26	56.2	71.8	74.5	33.3
+LLaVA-RLHF [14]	122k	Self-Reward	5.49	12.13	38.26	55.4	27.3	2.00	66.7	68.7	74.7	31.4
+POVID [38]	17k	GPT-4V	7.03	9.53	43.31	35.9	17.3	2.28	56.2	78.6	81.9	34.4
+HALVA [39]	21.5k	GPT-4V	5.49	11.27	42.42	49.1	24.6	2.14	60.4	78.0	83.5	32.3
+HA-DPO [13]	6k	GPT-4	5.49	11.56	42.16	44.9	21.8	1.97	61.5	74.2	78.0	32.6
+RLAIF-V [11]	74.8k	LLaVA-Next	5.93	5.49	36.75	9.9	4.9	3.04	39.6	72.7	84.4	34.6
+TPO [26]	21.4k	LLaVA-Next	7.03	11.27	41.62	5.0	4.7	2.76	42.7	82.2	87.2	34.2
+OPA-DPO [19]	4.8k	LLaVA-Next	6.37	11.84	42.69	6.1	3.7	2.83	46.9	81.3	85.6	33.1
+DPO [9]	21.4k	DeepSeek-V3	**7.25**	7.80	40.21	**12.9**	**8.8**	2.44	49.0	71.3	82.6	33.4
+mDPO [15]	21.4k	DeepSeek-V3	6.81	9.53	42.78	19.9	10.1	2.71	50.0	80.6	86.3	34.2
+SymMPO (Ours)	21.4k	DeepSeek-V3	**7.25**	**13.58**	**44.28**	19.5	9.7	**2.89**	**42.7**	**82.6**	**87.7**	**34.8**
LLaVA-1.5-13B [3]	✗	✗	6.59	9.53	43.48	51.2	25.1	2.16	59.4	71.3	73.2	33.1
+LLaVA-RLHF [14]	122k	Self-Reward	8.57	10.11	43.48	45.3	21.5	2.15	66.7	79.7	83.9	33.5
+HALVA [39]	21.5k	GPT-4V	8.79	10.11	42.24	47.0	22.9	2.30	57.3	82.9	86.5	33.1
+HSA-DPO [40]	8k	GPT-4/4V	6.15	8.95	41.62	5.4	2.9	2.55	50.0	79.8	82.8	33.7
+OPA-DPO [19]	4.8k	LLaVA-Next	6.81	12.13	42.60	7.7	4.4	3.05	38.5	84.1	87.5	32.3
+DPO [9]	21.4k	DeepSeek-V3	10.32	10.69	39.50	**15.4**	**8.5**	2.65	45.8	69.2	84.6	33.0
+mDPO [15]	21.4k	DeepSeek-V3	9.23	10.69	39.85	20.9	10.8	2.93	43.8	83.8	88.8	35.0
+SymMPO (Ours)	21.4k	DeepSeek-V3	**10.54**	**10.98**	**44.55**	20.4	10.0	**3.01**	**39.6**	**84.9**	**89.1**	**35.2**

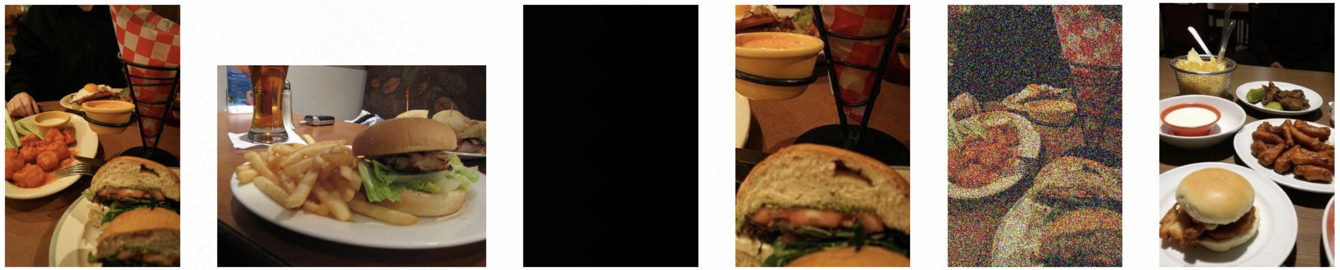
Table 2: Ablation studies with LLaVA-1.5-7B.

Model	HallusionBench			Object-HalBench		MMHal-Bench		AMBER		MMStar
	qAcc↑	fAcc↑	aAcc↑	Resp.↓	Ment.↓	Score↑	Hall↓	Acc↑	F1↑	Overall↑
SymMPO	**7.25**	**13.58**	44.28	19.5	**9.7**	**2.89**	**42.7**	**82.6**	**87.7**	34.8
w/o-$\mathcal{L}_{Pair}$	6.59	11.84	43.22	**18.1**	10.6	2.53	50.0	81.7	87.1	33.8
w/o-$\mathcal{L}_{Margin}$	7.03	10.98	**44.46**	21.1	11.0	2.40	54.2	82.0	87.3	34.5
w/o-$\mathcal{L}_{AncPO}$	6.81	11.84	40.83	21.6	11.6	2.39	59.4	79.5	87.4	**36.2**

- **SymMPO** consistently outperforms both standard **DPO** and **mDPO** (the previous vision-oriented method) across key hallucination benchmarks.

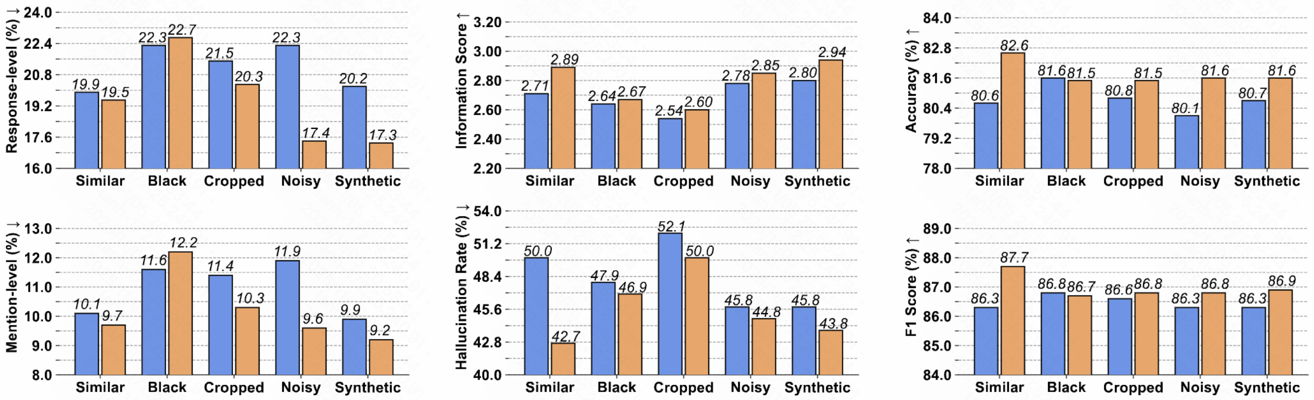
- Ablation experiments further validate the effectiveness of each component within SymMPO, proving the effectiveness of individual components.

Experiment



(a) Original (b) Similar (c) Black (d) Cropped (e) Noisy (f) Synthetic

Figure 3: Samples of the original image and its related contrastive images.



(a) Object-HalBench (b) MMHal-Bench (c) AMBER

Figure 4: Results of SymMPO and mDPO using different types of contrastive images (↑/↓: higher/lower is better). Orange represents SymMPO, and **blue** represents mDPO.

- To investigate the impact of different types of **contrastive image pairs** on the optimization performance of SymMPO, we constructed various types of contrastive image pairs and conducted experiments using SymMPO.
- Based on the experimental results, we analyzed how different image pair data influence the symmetric preference optimization effectiveness of SymMPO.

Thank You