P^2-DPO: Grounding Hallucination in Perceptual Processing via Calibration Direct Preference Optimization

Ruipeng Zhang[1,3], Zhihao Li[1,2,3], Haozhang Yuan[1,2,3], C.L. Philip Chen[1,2,3], Tong Zhang[1,2,3*]

[1]Guangdong Provincial Key Laboratory of Computational AI Models and Cognitive Intelligence, School of Computer Science & Engineering, South China University of Technology

[2]Pazhou Lab, Guangzhou, China

[2]Engineering Research Center of the Ministry of Education on Health Intelligent Perception and Paralleled Digital-Human, Guangzhou, China

ICLR International Conference On Learning Representations

Background & Motivation

- LVLM hallucinations are not pure **perception failures**. In many cases, the model already attends to the **correct visual region**, yet still produces the wrong answer.
- We identify this overlooked issue as **Perceptual Processing failure**, which mainly appears as **perceptual bottlenecks** in attended regions and **poor robustness** under image degradation.
- Existing DPO preference pairs are often **off-policy**, **vision-agnostic**, and costly to annotate, making them poorly matched to this problem.

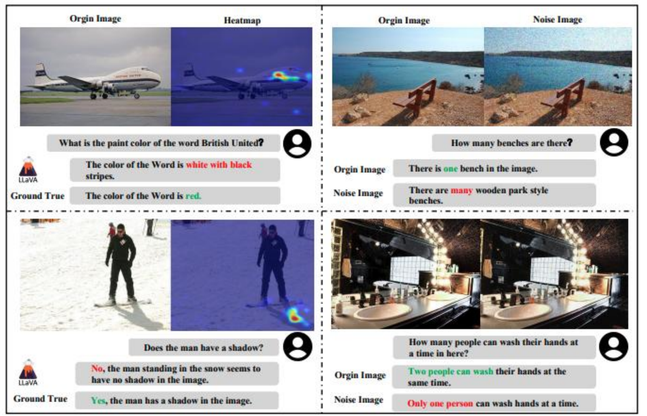


Figure 1: Left: Perceptual Bottleneck. Attention heatmaps visualize that the model correctly focuses on the key entity, yet still hallucinates. Right: Lack of Robustness. The model hallucinates when presented with an image containing noise imperceptible to humans. *All shown answers are generated by LLaVA-1.5-7B.*

Problem discussion & Method

➢ **Vision-Agnostic** Existing pairs are not visually sensitive, relying on textual correction instead of visual intervention.

➢ **Off-policy** Winning responses often come from external sources, causing a policy mismatch that weakens DPO.

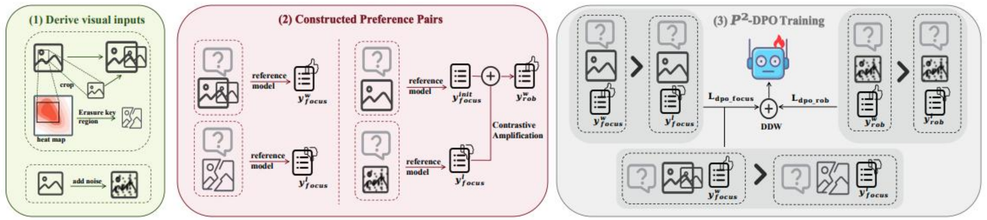


Figure 2: An overview of our proposed P^2-DPO framework. The process flows from left to right: (1) Derive Visual Inputs: Based on an initial forward pass to obtain an attention map ($\mathcal{A}$), we create an enhanced input via cropping (I_{aug}), a degraded input via erasure (I_{deg}), and a noisy input (I_{noise}), alongside the original image (I). (2) Generate Preference Pairs: The reference model generates two orthogonal preference pairs. (3) P^2-DPO Training: Difference losses are dynamically combined.

We propose ***P^2-DPO***, a self-correction framework that synergistically combines the generation of on-policy, vision-aware preference pairs with a purpose-built optimization strategy, including a dynamic weighting mechanism and a novel Calibration loss.

- **Focus-and-Enhance pairs** to correct perceptual bottlenecks by contrasting responses from enhanced and degraded visual evidence.
- **Visual Robustness pairs** to improve stability under noisy or degraded inputs.
- **Calibration Direct Preference Optimization**, which encourages the model not only to prefer better responses, but also to align them more strongly with visual evidence.
- **Dynamic Deficit-Weighting (DDW)** to adaptively rebalance the two objectives for each sample.

$$\mathcal{L}_{\text{Calib}} = -\mathbb{E}_{(y^w_{\text{focus}}, y^l_{\text{focus}}) \sim D_{\text{focus}}} \left[\log \sigma \left(\beta \log \frac{\pi_\theta(y^w_{\text{focus}} | I, I_{\text{crop}}, P)}{\pi_\theta(y^w_{\text{focus}} | I_{\text{deg}}, P)} - \beta \log \frac{\pi_{\text{ref}}(y^l_{\text{focus}} | I, I_{\text{crop}}, P)}{\pi_{\text{ref}}(y^l_{\text{focus}} | I_{\text{deg}}, P)} \right) \right]$$

$$\mathcal{L}_{\text{focus}} = \mathcal{L}_{\text{dpo\_focus}} + \lambda_{\text{calib}} \cdot \mathcal{L}_{\text{Calib}} \quad \mathcal{L}_{\text{total}} = \mathbb{E}\left[w_{\text{focus}} \cdot \mathcal{L}_{\text{focus}} + w_{\text{robust}} \cdot \mathcal{L}_{\text{dpo\_rob}} \right]$$

$$r = \frac{\text{CLIPScore}(P, I_{\text{crop}})}{\text{CLIPScore}(P, I)}$$

$$\alpha = \alpha_{\max} \tanh\left(\frac{r-1}{\tau} \right)$$

$$w_{\text{focus/robust}} = w_{\text{base}} \pm \alpha$$

Partial Experimental Results

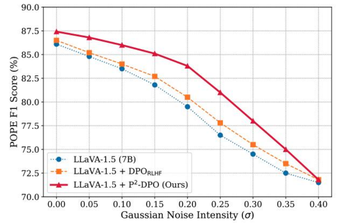
Model	Pairs Source	POPE (F1) ↑ Adv.	Pop.	Rand.	Avg.	MMHal-Bench Hal↓	Score↑	AMBER CHAIRi↓	Hal↓	$F1_E$ ↑	$F1_A$ ↑	$F1_R$ ↑
LLaVA-1.5-7B	Base	81.80	84.36	89.12	85.10	0.62	1.97	7.8	36.4	83.2	64.1	62.4
+ VCD	None	81.33	85.06	87.16	84.51	0.58	2.12	—	—	—	—	—
+ HA-DPO	AI	82.54	87.81	**90.25**	86.86	0.60	1.97	6.7	30.9	88.1	66.1	68.8
+ V-DPO$_{\text{RLHF-V}}$	Human	84.05	87.91	89.90	87.28	0.56	2.16	**5.6**	27.3	91.5	73.7	64.1
+ V-DPO$_{\text{SAD}}$	AI	83.77	87.62	89.57	86.98	**0.53**	2.36	6.6	30.5	**95.2**	**76.1**	61.1
+ DPO$_{\text{RLHF-V}}$	Human	84.03	85.94	89.12	86.38	0.60	2.08	6.4	34.5	90.7	72.6	64.6
+ P^2-DPO (Ours)	Self	**84.53** ↑2.73	**87.91** ↑3.55	89.87 ↑0.75	**87.44** ↑2.34	0.56 ↓0.06	**2.43** ↑0.46	5.9 ↓1.9	**26.7** ↓9.7	92.6 ↑9.4	71.7 ↑7.6	**70.9** ↑8.5
Qwen2.5-VL-3B	Base	86.26	88.05	89.79	88.03	0.42	3.38	6.8	39.1	**93.2**	83.7	77.9
+ DPO$_{\text{RLHF-V}}$	Human	86.41	**88.59**	90.40	88.46	0.40	3.41	6.8	39.5	93.0	83.5	78.1
+ P^2-DPO (Ours)	Self	**86.88** ↑0.62	88.48 ↑0.43	**91.50** ↑1.71	**88.95** ↑0.92	**0.39** ↓0.03	**3.49** ↑0.11	**6.6** ↓0.2	**38.2** ↓0.9	92.7 ↓0.5	**84.2** ↑0.5	**80.9** ↑3.0
Qwen2.5-VL-7B	Base	84.92	86.78	89.01	86.90	0.349	3.47	5.4	29.0	97.3	**83.8**	72.6
+ DPO$_{\text{RLHF-V}}$	Human	84.99	86.81	89.72	87.17	0.345	3.63	4.2	29.3	**98.5**	82.5	72.2
+ P^2-DPO (Ours)	Self	**86.79** ↑1.87	**87.79** ↑1.01	**90.05** ↑1.04	**88.21** ↑1.31	**0.32** ↓0.03	**3.94** ↑0.47	**3.9** ↓1.5	**24.9** ↓4.1	98.5 ↑1.2	83.4 ↓0.4	**72.9** ↑0.3

Table 4: Ablation study of our method's components. We report performance on key hallucination and reasoning benchmarks. Each component contributes positively to the final performance.

Configuration	POPE F1 ↑	AMBER Hal ↓	MMHal Score ↑	HallusionBench qAcc ↑	fAcc ↑	aAcc ↑
P^2-DPO	**87.42**	**29.27**	**2.43**	**26.37**	**30.05**	**55.62**
Ablating Pairs:						
w/o FEPs	85.84	34.70	2.30	23.56	26.62	51.10
w/o VRPs	85.27	33.60	2.26	21.02	25.71	52.10
Ablating Strategy:						
w/o $\mathcal{L}_{\text{Calib}}$	86.17	33.10	2.33	23.56	27.75	53.41
w/o DDW	86.68	31.20	2.39	23.96	28.61	52.70

Table 1: Evaluation on AFR and Processing Accuracy (AFR > 14.0).

Model	AFR (Avg. ↑)	P-Acc. (% ↑)
LLaVA-1.5-7B	14.73	66.29
+ DPO$_{\text{RLHF}}$	15.57 ↑0.84	65.71 ↓0.58
+ P^2-DPO	**18.71** ↑3.98	**70.10** ↑3.81



Paper

Code

ruipengzhang.zrp@gmail.com

202421044797@mail.scut.edu.cn

E-mail