CHiP: Cross-modal Hierarchical Direct Preference Optimization for Multimodal LLMs

Jinlan Fu[1], Shenzhen Huangfu[1,2], Hao Fei[1], Xiaoyu Shen[3],
Bryan Hooi[1], Xipeng Qiu[2], See-Kiong Ng[1]

[1]National University of Singapore, [2]Fudan University,
[3]Eastern Institute of Technology

https://github.com/LVUGAI/CHiP

Code

Paper

Outline

1. Task
2. Background
3. Motivation and Limitation
4. Method
5. Experiment
6. Conclusion and Contribution

1. Task

➢ MLLM Hallucination

- The model’s output is not based on the visual input.

Question

What is the color of the banana?

Image

The figure is taken from [1].

GPT-4o

Answer

Yellow.

Black.

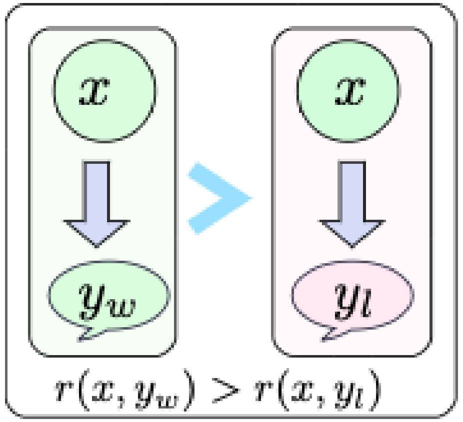
[1] Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding

2. Background

➢ Application of DPO in Different Scenarios

- ➢ Direct Preference Optimization meets LLMs (a)

$$r(x, y_w) > r(x, y_l)$$



(a) DPO

2. Background

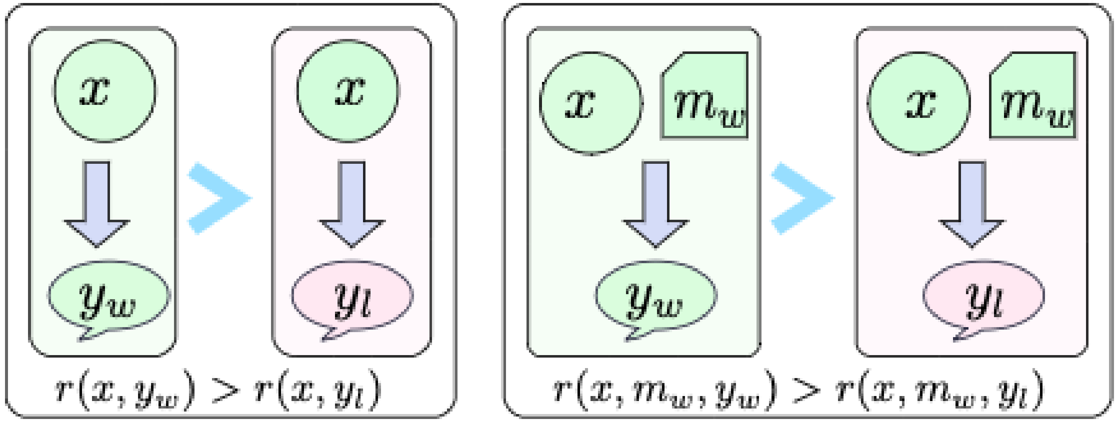
➢ Application of DPO in Different Scenarios

- Direct Preference Optimization meets LLMs (a)

$$r(x, y_w) > r(x, y_l)$$

- Direct Preference Optimization meets MLLMs (b)

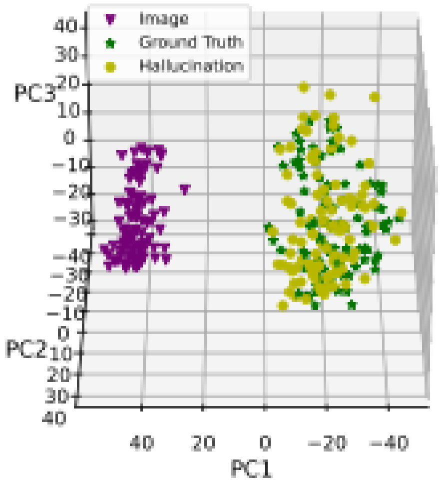
$$r(x, m_w, y_w) > r(x, m_w, y_l)$$



(a) DPO (b) Multimodal DPO

3. Motivation & Limitation

➢ Limitation of Existing Methods:



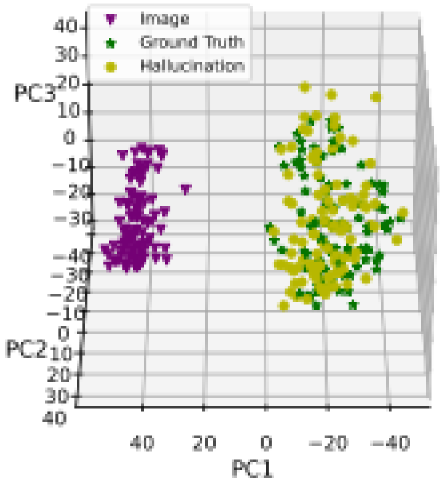
(a) LLaVA

Ideally, in well-aligned MLLMs, image and ground-truth representations should be close, while hallucinated ones should be distant from the ground-truth.

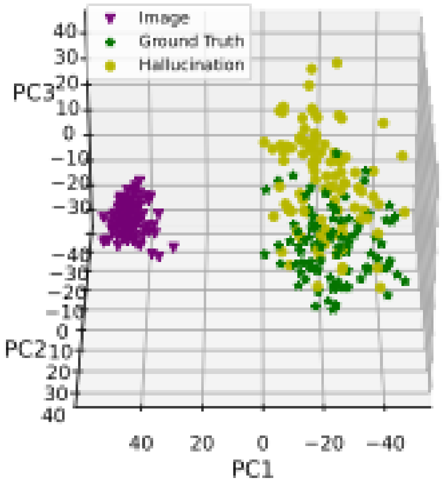
3. Motivation & Limitation

➢ Limitation of Existing Methods:

- ➢ ***DPO** struggles to align image and description representations and to effectively distinguish between hallucinated and non-hallucinated descriptions. (Fig. 1-(b));*



(a) LLaVA



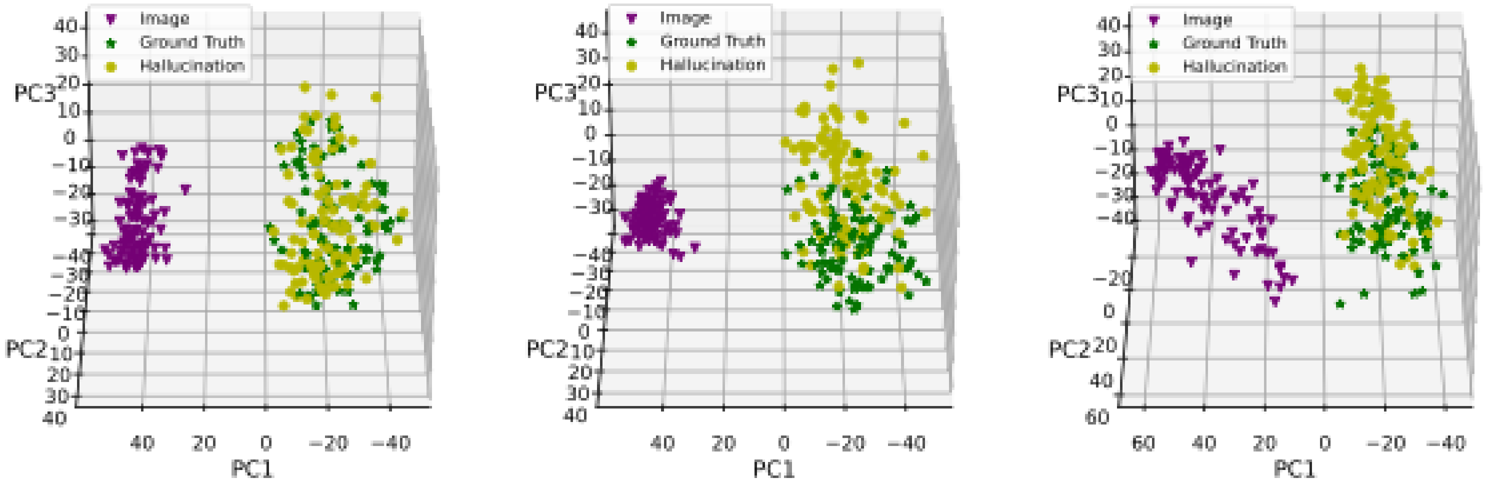
(b) LLaVA+DPO

Ideally, in well-aligned MLLMs, image and ground-truth representations should be close, while hallucinated ones should be distant from the ground-truth.

3. Motivation & Limitation

➢ Limitation of Existing Methods:

- ➢ ***DPO** struggles to align image and description representations and to effectively distinguish between hallucinated and non-hallucinated descriptions. (Fig. 1-(b));*
- ➢ *Although visual preference optimization (**CMDPO**) has alleviated the issue to some extent, there is still substantial room for improvement. (Fig. 1-(c));*

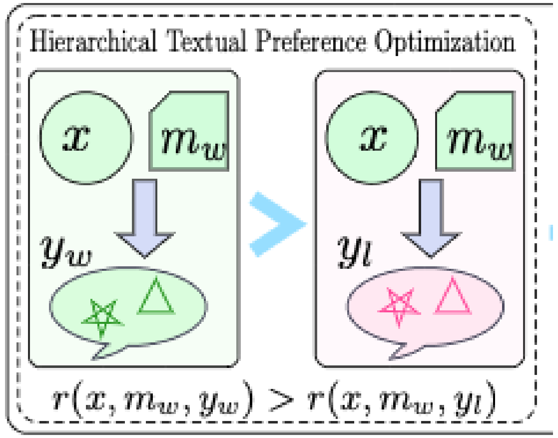


(a) LLaVA (b) LLaVA+DPO (c) LLaVA+CMDPO

Ideally, in well-aligned MLLMs, image and ground-truth representations should be close, while hallucinated ones should be distant from the ground-truth.

4. Methodology: CHiP

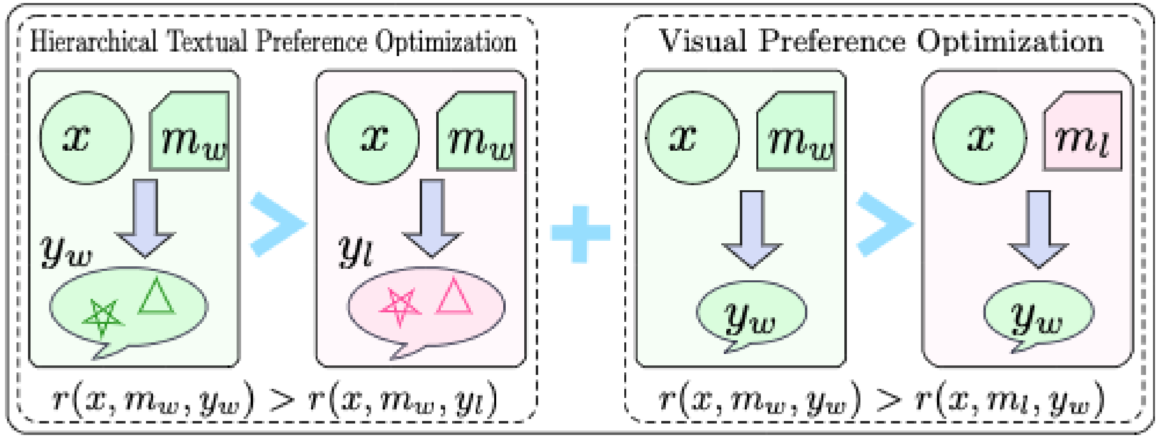
- **CHiP**: **C**ross-modal **Hi**erarchical Direct **P**reference Optimization
  - Hierarchical Textual Preference Optimization
    - Response-level Preference Optimization ($\mathcal{L}_{DPO_r}$)
    - Segment-level Preference Optimization ($\mathcal{L}_{DPO_s}$)
    - Token-level Preference Optimization ($\mathcal{L}_{PO_k}$)



4. Methodology: CHiP

- **CHiP**: **C**ross-modal **Hi**erarchical Direct **P**reference Optimization
  - Hierarchical Textual Preference Optimization
    - Response-level Preference Optimization ($\mathcal{L}_{DPO_r}$)
    - Segment-level Preference Optimization ($\mathcal{L}_{DPO_s}$)
    - Token-level Preference Optimization ($\mathcal{L}_{PO_k}$)
  - Visual Preference Optimization ($\mathcal{L}_{DPO_v}$)

$$\mathcal{L}_{CHiP} = \mathcal{L}_{DPO_v} + \mathcal{L}_{DPO_r} + \lambda \mathcal{L}_{DPO_s} + \gamma \mathcal{L}_{PO_k}\,.$$



5. Experiment

- Benchmarks
  - Object HalBench (ObjHal)
  - MMHal-Bench (MMHal)
  - HallusionBench
  - AMBER
- Training Data:
  - RLHF-V-Dataset (Yu et al.)

5.1 Experiment: Hallucination Mitigation Results

Table 1: The results of hallucination evaluation on the Object HalBench (ObjHal), MMHal-Bench (MMHal), HallusionBench, and AMBER datasets. Values in **bold** indicate the best performance under the same setting. "↑" indicates that a higher value is better for this metric, while "↓" indicates that a lower value is better. The baseline results are reported in (Yu et al., 2024a) for ObjHal and MMHal, in (Guan et al., 2024) for HallucinationBench, and in (Wang et al., 2024) for AMBER.

Model	ObjHal		MMHal		HallusionBench			AMBER			
	R.↓	M.↓	Ova.↑	R.↓	qA↑	fA↑	aA↑	CHAIR↓	Cover↑	Hal↓	Cog↓
Referenced Results (Not Directly Comparable)											
LLaVA-1.0 (Liu et al., 2024c)	63.0	29.5	-	70.8	-	-	-	-	-	-	-
Muffin (Yu et al., 2023)	50.5	24.5	-	68.8	-	-	-	-	-	-	-
LRV (Liu et al., 2023a)	32.3	22.3	-	78.1	8.8	13.0	42.8	-	-	-	-
LLaVA-RLHF (Sun et al., 2023)	38.1	18.9	2.5	57.0	-	-	-	7.7	52.1	39.0	4.4
InstructBLIP (Dai et al., 2023)	25.9	14.3	2.1	58.0	9.5	10.1	45.3	8.8	52.2	38.2	4.4
Qwen-VL-Chat (Bai et al., 2023)	43.8	20.0	2.9	43.0	5.9	6.7	39.2	6.6	53.2	31.0	2.9
LLaVA-1.5 (Liu et al., 2023c)	46.3	22.6	2.4	52.1	10.6	24.9	46.9	7.8	51.0	36.4	4.2
RLHF-V (Yu et al., 2024a)	12.2	7.5	2.5	51.0	-	-	-	6.3	46.1	25.1	2.1
HALVA (Sarkar et al., 2024b)	-	-	-	-	13.9	20.1	49.1	-	-	-	-
GPT-4V (OpenAI, 2023)	13.6	7.3	-	31.3	28.8	39.9	65.3	4.6	67.1	30.7	2.6
Muffin (13B)	21.5	11.6	2.4	60.42	16.0	20.8	50.9	8.0	**48.3**	32.1	3.5
+DPO	13.1	6.6	2.5	52.1	17.4	23.4	52.5	6.2	46.9	26.5	2.5
+CHiP	**6.2**	**3.9**	**2.6**	**49.0**	**19.1**	**24.9**	**54.0**	**4.4**	45.3	**17.6**	**1.5**
LLaVA-1.6 (7B)	14.1	7.4	2.8	42.7	15.8	20.8	51.6	8.3	**61.0**	48.6	4.2
+DPO	11.0	6.6	2.7	43.8	22.2	**28.3**	56.6	5.9	**61.0**	38.9	3.0
+CHiP	**4.9**	**3.2**	**2.9**	**39.6**	**23.5**	26.0	**58.5**	**3.7**	57.8	**24.5**	**1.6**

5.1 Experiment: Hallucination Mitigation Results

Table 1: The results of hallucination evaluation on the Object HalBench (ObjHal), MMHal-Bench (MMHal), HallusionBench, and AMBER datasets. Values in **bold** indicate the best performance under the same setting. "↑" indicates that a higher value is better for this metric, while "↓" indicates that a lower value is better. The baseline results are reported in (Yu et al., 2024a) for ObjHal and MMHal, in (Guan et al., 2024) for HallucinationBench, and in (Wang et al., 2024) for AMBER.

Model	ObjHal		MMHal		HallusionBench			AMBER			
	R.↓	M.↓	Ova.↑	R.↓	qA↑	fA↑	aA↑	CHAIR↓	Cover↑	Hal↓	Cog↓
Referenced Results (Not Directly Comparable)											
LLaVA-1.0 (Liu et al., 2024c)	63.0	29.5	-	70.8	-	-	-	-	-	-	-
Muffin (Yu et al., 2023)	50.5	24.5	-	68.8	-	-	-	-	-	-	-
LRV (Liu et al., 2023a)	32.3	22.3	-	78.1	8.8	13.0	42.8	-	-	-	-
LLaVA-RLHF (Sun et al., 2023)	38.1	18.9	2.5	57.0	-	-	-	7.7	52.1	39.0	4.4
InstructBLIP (Dai et al., 2023)	25.9	14.3	2.1	58.0	9.5	10.1	45.3	8.8	52.2	38.2	4.4
Qwen-VL-Chat (Bai et al., 2023)	43.8	20.0	2.9	43.0	5.9	6.7	39.2	6.6	53.2	31.0	2.9
LLaVA-1.5 (Liu et al., 2023c)	46.3	22.6	2.4	52.1	10.6	24.9	46.9	7.8	51.0	36.4	4.2
RLHF-V (Yu et al., 2024a)	12.2	7.5	2.5	51.0	-	-	-	6.3	46.1	25.1	2.1
HALVA (Sarkar et al., 2024b)	-	-	-	-	13.9	20.1	49.1	-	-	-	-
GPT-4V (OpenAI, 2023)	13.6	7.3	-	31.3	28.8	39.9	65.3	4.6	67.1	30.7	2.6
Muffin (13B)	21.5	11.6	2.4	60.42	16.0	20.8	50.9	8.0	**48.3**	32.1	3.5
+DPO	13.1	6.6	2.5	52.1	17.4	23.4	52.5	6.2	46.9	26.5	2.5
+CHiP	**6.2**	**3.9**	**2.6**	**49.0**	**19.1**	**24.9**	**54.0**	**4.4**	45.3	**17.6**	**1.5**
LLaVA-1.6 (7B)	14.1	7.4	2.8	42.7	15.8	20.8	51.6	8.3	**61.0**	48.6	4.2
+DPO	11.0	6.6	2.7	43.8	22.2	**28.3**	56.6	5.9	**61.0**	38.9	3.0
+CHiP	**4.9**	**3.2**	**2.9**	**39.6**	**23.5**	26.0	**58.5**	**3.7**	57.8	**24.5**	**1.6**

Findings

- CHiP significantly reduces hallucinations of base models Muffin and LLaVA.
- CHiP outperforms DPO on the four benchmarks.
- LLaVA and Muffin with CHiP achieve fewer hallucinations compared to GPT-4 on the ObjHal and AMBER datasets.

5.2 General Capability Analysis

- Preference learning may compromise a model's general understanding capabilities. Here, we evaluate and analyze the general capability performance of an MLLM enhanced by our CHiP.

Table 4: The general capability evaluation results. Values in black indicate the best performance, in red show improvement with CHiP, and in green indicate a decline. Values with * are reproduced results. For LLaVA-Wild, we used *gpt-4o-2024-05-13* as evaluator due to *GPT-4-0314* was outdated; for MMMU-test, there was a lack of official LLaVA-1.6 reports.

	MMMU(val)	MMMU(test)	MMB-ENG	MMB-CN	ScienceQA	LLaVA-Wild
Num Samples	900	10500	6666	6666	4241	90
LLaVA	35.80	31.70*	**67.40**	60.60	70.10	74.90
LLaVA+CHiP	**36.8**$^{+1.0}$	**32.1**$^{+0.4}$	66.6$^{-0.8}$	**60.82**$^{+0.22}$	**70.15**$^{+0.05}$	**76.2**$^{+1.3}$

5.2 General Capability Analysis

➢ Preference learning may compromise a model's general understanding capabilities. Here, we evaluate and analyze the general capability performance of an MLLM enhanced by our CHiP.

Table 4: The general capability evaluation results. Values in black indicate the best performance, in red show improvement with CHiP, and in green indicate a decline. Values with * are reproduced results. For LLaVA-Wild, we used *gpt-4o-2024-05-13* as evaluator due to *GPT-4-0314* was outdated; for MMMU-test, there was a lack of official LLaVA-1.6 reports.

	MMMU(val)	MMMU(test)	MMB-ENG	MMB-CN	ScienceQA	LLaVA-Wild
Num Samples	900	10500	6666	6666	4241	90
LLaVA	35.80	31.70*	**67.40**	60.60	70.10	74.90
LLaVA+CHiP	**36.8**$^{+1.0}$	**32.1**$^{+0.4}$	66.6$^{-0.8}$	**60.82**$^{+0.22}$	**70.15**$^{+0.05}$	**76.2**$^{+1.3}$

Findings

LLaVA+CHiP outperforms LLaVA on five out of the six datasets. This indicates that CHiP can mitigate the hallucination of MLLMs and slightly improve general capability.

5.3 Human Evaluation

- Due to incomplete text annotations on the MMHal, GPT-4 couldn't reliably detect hallucinations. To make the results more reliable, we invited experts to conduct the human evaluation.

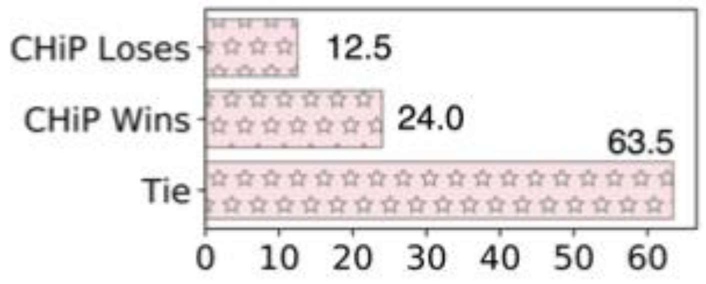


Figure 3: Human evaluation results on MMHal-Bench (MMHal).

5.3 Human Evaluation

- Due to incomplete text annotations on the MMHal, GPT-4 couldn't reliably detect hallucinations. To make the results more reliable, we invited experts to conduct the human evaluation.

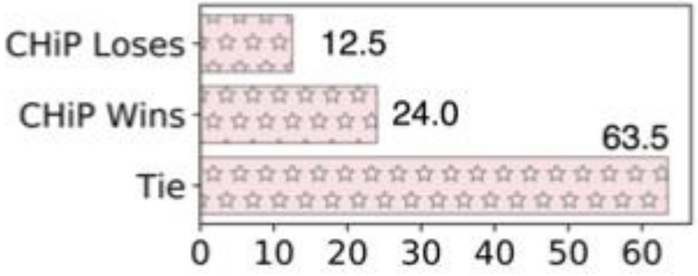


Figure 3: Human evaluation results on MMHal-Bench (MMHal).

Findings

- CHiP and DPO performed equally on 63.5% of samples, with CHiP winning 24%.
- In 36.5% of samples where a distinction was possible, CHiP outperformed DPO in 31.6%.

5.4 Effect of Component Combination

➢ To evaluate the contribution of each component in CHiP and the effect of their combinations, we conducted a comprehensive ablation study on CHiP based on LLaVA.

Table 2: The ablation results of CHiP based on LLaVA. Values in **bold** denote the best performance.

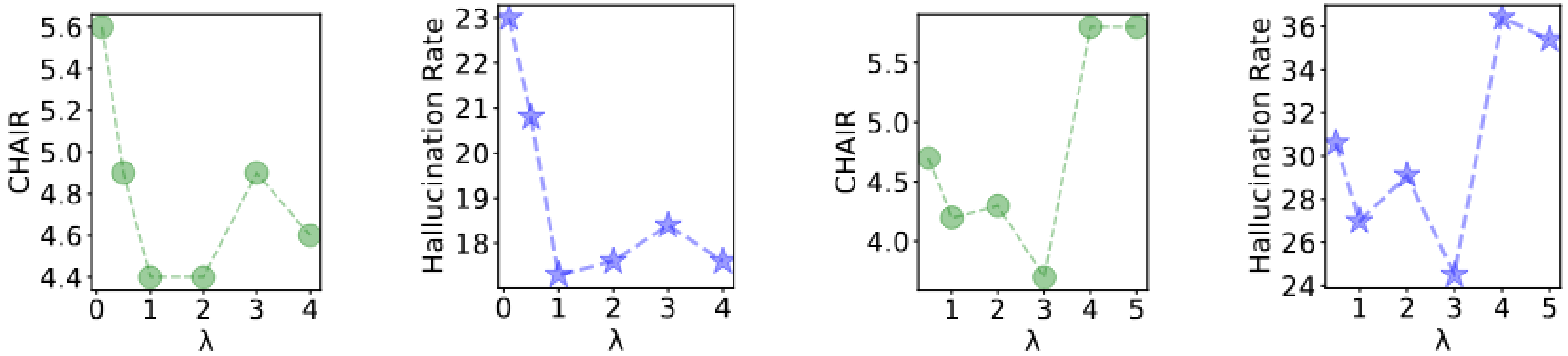
Model	ObjHal		MMHal	
	R.↓	M.↓	Ova.↑	R.↓
DPO	11.03	6.61	2.73	43.75
CHiP	**4.92**	**3.21**	**2.89**	**39.63**
-$\mathcal{L}_{DPO_v}$	9.19	5.77	2.70	42.40
-$\mathcal{L}_{DPO_s}$	8.55	5.16	2.69	40.63
-$\mathcal{L}_{PO_t}$	6.08	3.77	2.71	40.75
-$\mathcal{L}_{DPO_s}$-$\mathcal{L}_{PO_t}$	9.76	5.47	2.78	41.71

Findings

- Both hierarchical textual preference optimization (HDPO) and visual preference optimization (CMDPO) are effective.

5.5 Strength of HDPO

- Hierarchical text preference optimization (HDPO) includes preference optimization at the response, segment, and token levels. Here, we discuss the impact of their weights.



(a) Muffin+CHAIR (b) Muffin+Hal Rate (c) LLaVA+CHAIR (d) LLaVA+Hal Rate

Figure 4: Results of Muffin+CHiP and LLaVA+CHiP evaluated on the AMBER dataset with different choices of weight λ to control the strength of segment-level preference optimization.

Findings

- When $\lambda = 1$ $(\lambda = 3)$, the best performance of the CHAIR and Hallucination Rate metric is achieved on AMBER based on Muffin (LLaVA-1.6).

5.6 Impact of Training Paradigm

- Most previous approaches freeze the visual encoder and train only the connector and LLM during preference optimization.
- Question: Can full-model training during MLLM preference optimization reduce hallucinations?

Table 3: Results of training or freezing the visual encoder (VE) in LLaVA during preference optimization. × and ✓ denote the visual encoder states of training and freezing, respectively.

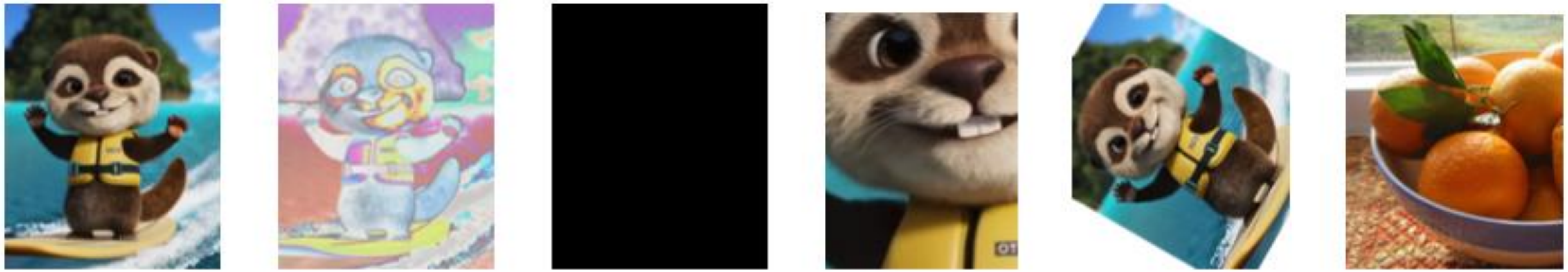
Model	VE	MMHal		AMBER		ObjHal		
		Ova.↑	R.↓	CHAIR↓	Cover↑	Hal↓	R.↓	M.↓
LLaVA	-	2.75	42.7	8.30	61.0	48.6	14.1	7.4
+DPO	×	**2.73**	**43.8**	5.94	61.0	38.9	11.0	6.6
+DPO	✓	2.71	44.8	**5.88**	**61.6**	**38.3**	**10.1**	**5.7**
+CHiP	×	**2.89**	**39.6**	**3.72**	**57.8**	24.5	**4.9**	**3.2**
+CHiP	✓	2.68	43.8	3.74	54.9	**22.1**	5.3	3.3

Findings

DPO achieves a lower hallucination rate when the visual encoder is trained, whereas CHiP does not achieve the expected reduction in hallucination rate.

5.7 Rejection Image Construction Strategy

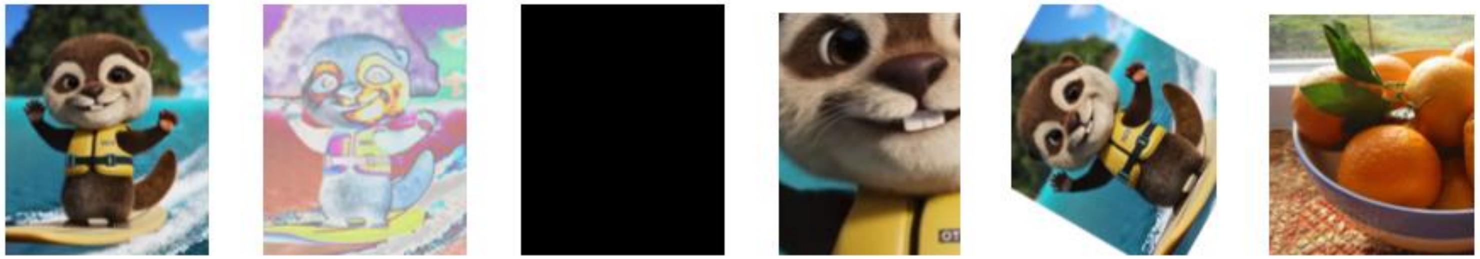
- **Strategies.**
  - **Diffusion**: Following the forward diffusion process in image generation, small amounts of Gaussian noise are gradually added to the chosen image for T=500 steps.
  - **Blackness**: set all the RGB values of the chosen image to 0.
  - **Crop**: A random cropping strategy is utilized on the chosen image.
  - **Rotation**: randomly rotate the chosen image by 10 to 80 degrees.
  - **Randomness**: select an image from the training set randomly.



(a) Chosen (b) Diffusion (c) Blackness (d) Crop (e) Rotation (f) Randomness

Figure 5: Examples of rejection images constructed by different strategies. (a) is the chosen image.

5.7 Rejection Image Construction Strategy



(a) Chosen (b) Diffusion (c) Blackness (d) Crop (e) Rotation (f) Randomness

Figure 5: Examples of rejection images constructed by different strategies. (a) is the chosen image.

Table 5: Results of CHiP under different rejection image construction strategies. The **bold** values indicate the best performance. Observation: CHiP achieves the best performance with the diffusion strategy constructed rejection images.

Strategy	ObjHal		MMHal	
	R.↓	M.↓	Ova.↑	R.↓
Diffusion	**4.9**	**3.2**	**2.9**	**39.6**
Black	9.4	5.0	2.4	43.8
Cropping	5.8	3.6	2.8	40.6
Random	10.9	5.9	2.9	41.7
Rotate	7.8	4.4	2.8	43.8

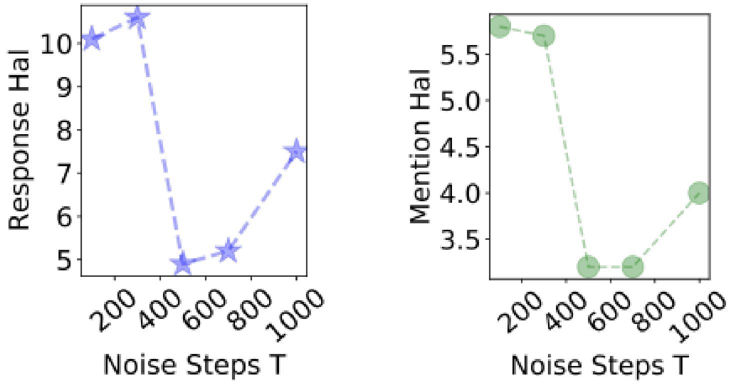
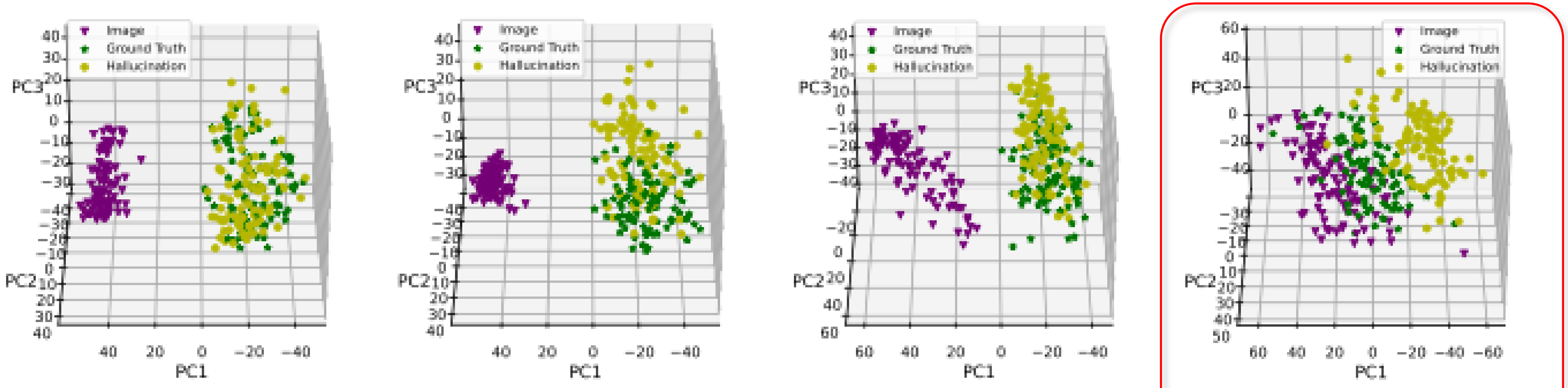


Figure 6: Results of LLaVA+CHiP evaluated on the ObjHal dataset with different values of noise step T. "Response" represents the response-level hallucination rate, while "Mention" represents the mention-level hallucination rate.

Observation:

- Diffusion Strategy achieves better performance.
- CHiP performs best at T=500.

5.8 Representation Visualization



(a) LLaVA (b) LLaVA+DPO (c) LLaVA+CMDPO (d) LLaVA+CHiP

Findings:

- ***DPO*** *struggles to align image and description representations and to effectively distinguish between hallucinated and non-hallucinated descriptions. (Fig. 1-(b));*
- *Although visual preference optimization (****CMDPO****) has alleviated the issue to some extent, there is still substantial room for improvement. (Fig. 1-(c));*
- Diffusion With the introduction of more fine-grained text and image preference optimization, namely CHiP, the alignment between the image and ground-truth descriptions becomes even closer, while maintaining the ability to distinguish between hallucinated and non-hallucinated texts.

6. Conclusion and Contribution

- We analyze the limitations of multimodal DPO through image and text representation distributions, emphasizing its failure to achieve cross-modal semantic alignment and distinguish between hallucinated and non-hallucinated descriptions.
- We propose CHiP to address these limitations. CHiP includes a hierarchical textual preference optimization module to capture fine-grained (i.e., response, segment, and token) preferences and a visual preference optimization module to extract cross-modal preferences.
- We equipped CHiP with various MLLMs, and the results of multiple datasets demonstrate that CHiP reduces hallucinations and enhances cross-modal semantic alignment.

Resource

Paper *Github* *Home* *Twitter/X*