Seeing Through the Chain: Understanding and Mitigating Hallucinations in Multimodal Large Reasoning Models

Hao Fang[*1], **Jinyu Li**[*1,2], **Jiawei Kong**[1], **Tianqu Zhuang**[*1], **Kuofeng Gao**[1], **Bin Chen**[†2], **Shu-Tao Xia**[1]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University,
[2]Harbin Institute of Technology, Shenzhen,
{fangh25, kjw25, zhuangtq23, gkf21}@mails.tsinghua.edu.cn,
2023311402@stu.hit.edu.cn, chenbin2021@hit.edu.cn
xiast@sz.tsinghua.edu.cn

Abstract

While multimodal large reasoning models (MLRMs) have exhibited impressive capabilities, they remain prone to hallucinations, and effective solutions remain underexplored. In this paper, we investigate the underlying causes of hallucinations in MLRMs and accordingly propose a reasoning-centric training framework for mitigation. Specifically, we find that introducing reasoning mechanisms exacerbates models' reliance on language priors and overlooks visual inputs, leading to CoTs with reduced visual cues but redundant text tokens. Guided by the Information Bottleneck (IB) principle, we propose selectively filtering redundant thinking tokens to obtain a more compact, signal-efficient CoT that preserves task-relevant information while suppressing noise. We also observe that the quality of the reasoning trace largely determines whether hallucination emerges in final answers. To leverage this, we introduce a reasoning-enhanced preference optimization scheme that constructs training pairs using high-quality AI feedback. We further propose generating high-quality negative samples for contrastive optimization via a multimodal hallucination-inducing mechanism that exposes model hallucination behaviors through carefully crafted visual and textual inducers. By modeling CoT chains as intermediate representations between inputs and final answers, we establish an IB-based theoretical justification for our method. Extensive experiments reveal consistent hallucination reduction across diverse MLRMs and benchmarks.

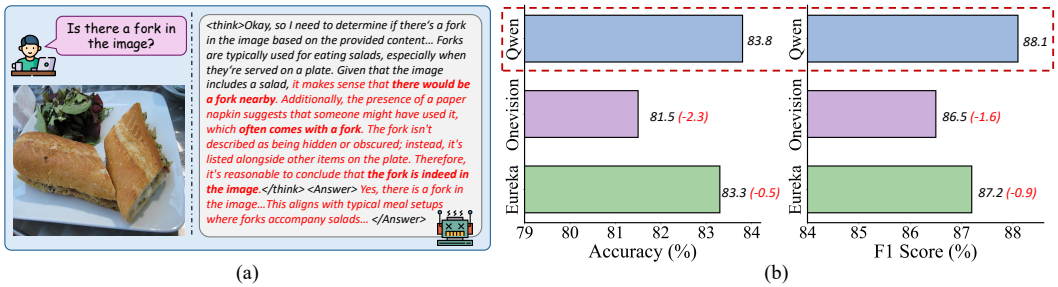
1 Introduction

Large Reasoning Models (LRMs) such as DeepSeek-R1 [1] have recently demonstrated remarkable problem-solving capabilities across a wide range of scenarios. Motivated by these advances, numerous studies have sought to extend such powerful multi-step reasoning mechanisms to multimodal models by applying supervised fine-tuning (SFT) or reinforcement learning (RL) algorithms. Benefiting from meticulous chain-of-thought (CoT) supervision, Multimodal Large Reasoning Models (MLRMs) have achieved impressive performance on various challenging tasks such as visual mathematics [2].

Despite recent progress, the hallucination remains a persistent issue for MLRMs [7], *i.e.*, the model occasionally generates semantically plausible yet factually incorrect or visually irrelevant content, as in Figure 1(a). More strikingly, evaluation of representative MLRMs and the base model reveals that

[*]Equal Contribution
[†]Corresponding Author

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



(a) (b)

Figure 1: (a) A hallucination case from R1-Onevision [3]. (b) Performance of MLRMs R1-Onevision and MM-Eureka [4], and the base model Qwen2.5-VL-7B [5] on AMBER [6], where *high values indicate fewer hallucinations*. The hallucination increases from the base model to reasoning variants.

MLRMs can hallucinate even more severely than their non-reasoning counterparts (Figure 1(b)). This is a counterintuitive phenomenon, as the presence of logical, structured reasoning should, in principle, facilitate a more detailed perception of the image and hence yield more faithful and image-grounded answers. These observations underscore the need for a deeper understanding of hallucinations specific to MLRMs and the development of effective mitigation strategies. However, existing studies primarily focus on constructing benchmarks to better evaluate MLRM's hallucinations and influential factors [7], or exclusively discuss the self-contradictory reasoning within CoT traces [8], leaving approaches for hallucination mitigation in general multimodal tasks largely underexplored.

In this work, we first conduct two analysis tailored to the unique characteristics of MLRMs and observe two key factors relevant to their hallucinations. Based on these insights, we propose C3PO, a two-stage hallucination mitigation framework specifically designed for MLRMs via **C**oT **C**ompression and **C**ontrastive **P**reference **O**ptimization. We begin by analyzing the attention distribution of MLRMs and observe that introducing explicit reasoning mechanisms causes models to allocate even less attention to visual tokens than their non-reasoning counterparts. This issue weakens visual signals in the generated CoTs while introducing redundant low-information text tokens. From the Information Bottleneck (IB) perspective [9, 10], we model the reasoning chain as an intermediate representation between visual inputs and final answers, which should preserve critical visual cues relevant to the answer while discarding redundant or noisy content. Therefore, we propose to selectively filters low-importance reasoning tokens to produce a more compact and signal-efficient representation, which is theoretically justified under the IB framework. Remarkably, this strategy directly reduce hallucinations without introducing any additional prior information.

Next, we investigate the causal relationship of hallucinations from reasoning chains to final answers. Our analysis reveals a key finding that high-quality CoTs consistently lead to hallucination-free responses, whereas hallucinated CoTs substantially amplify the likelihood of hallucinations in final answers. Building on this, we propose a reasoning-enhanced preference learning strategy based on direct preference optimization (DPO). Specifically, we leverage expert feedback from advanced MLLMs to enhance the quality of model-generated CoTs. Data samples with enhanced traces are used as preferred samples, while the original outputs serve as rejected ones. To expose the model's intrinsic hallucination behaviors, we further introduce a multimodal hallucination-inducing mechanism that carefully crafts visual and textual inputs to elicit hallucination patterns as additional negative contrasts. By performing contrastive preference tuning on these crafted pairs, the model is trained to generate more reliable and coherent reasoning chains, substantially reducing hallucinations in final answers.

Contributions. We first uncover the hallucination causes specific to the reasoning mechanism in MLRMs and identify two key challenges that motivate the design of C3PO, a two-stage hallucination mitigation framework for MLRMs. C3PO first leverages SFT to obtain more information-dense CoTs, and then performs the reasoning-enhanced contrastive preference optimization (CPO). In addition, we introduce a multimodal hallucination-inducing technique that exploits their inherent failure behaviors via well-designed inducers to generate contrastive references. Furthermore, we develop a theoretical framework from an IB perspective that provides principled justification for C3PO's superiority.

To validate the effectiveness, we conduct extensive experiments on various MLRMs across multiple benchmarks, which demonstrate that the proposed framework substantially reduces hallucinations.

Related work. Due to page limits, we discuss related studies in Appendix D.

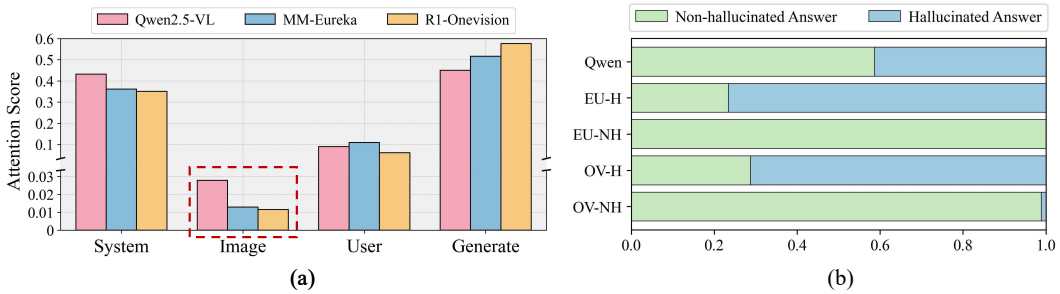


Figure 2: Two important hallucination analyses for MLRMs. (a) Attention distributions of MLRMs and the non-reasoning Qwen2.5-VL on 300 samples from MSCOCO [11]. As observed, MLRMs exhibit a substantial reduction in attention to visual tokens. (b) Proportion of hallucinated answers given hallucinated (-H) and non-hallucinated (-NH) CoTs under CHAIR [12] evaluation. EU denotes MM-Eureka, and OV denotes R1-Onevision. The result of Qwen2.5-VL is reported as a reference.

2 Understanding Hallucination in MLRMs

This section provides preliminary analyses tailored to the unique characteristics of MLRMs to inspire the principled design of our framework. A key distinction between MLRMs and conventional multimodal large language models (MLLMs) is the explicit incorporation of CoT reasoning, based on which we investigate two fundamental questions closely related to hallucination behaviors:

- **Q1:** *How does the introduction of an additional reasoning process influence models' attention allocation? Does reasoning encourage more attention to visual inputs or instead shift reliance toward language priors due to the increased number of textual tokens?*
- **Q2:** *Does hallucination arising in the reasoning chain causally propagate to the final answer due to the autoregressive nature of LLMs? More importantly, does a hallucination-free reasoning chain necessarily guarantee a hallucination-free final answer?*

Question 1. Previous studies [13, 14] have shown that a primary cause of hallucinations is models' tendency to allocate excessive attention to text tokens while ignoring critical visual inputs. As a result, the responses are more influenced by the language priors rather than the actual image input.

The attention analysis in Figure 2(a) shows that reasoning models further exacerbate attention bias. Compared to their non-reasoning base models, MLRMs significantly decrease attention to visual tokens while allocating even more attention to textual tokens. This shift amplifies the imbalance between visual grounding and language priors during generation, increasing the risk of hallucinations.

Summary 1. Introducing an additional reasoning mechanism decreases focus on visual tokens and amplifies reliance on textual priors, which can produce CoTs with diluted visual information and redundant low-information text tokens. Building on the IB principle and empirical findings that many CoT tokens are redundant and irrespective of model decisions [15, 16], we propose a selective CoT token pruning strategy to enhance information flow and encourage reliable answers.

Question 2. The CoT chain plays a critical role in shaping the final answer. To investigate the potential causal relationship between hallucinations in reasoning chains and final answers, we decouple these two sources of hallucination under the CHAIR benchmark and conduct a causal analysis.

Figure 2(b) shows that when the CoT is hallucinated, most generated answers also contain hallucinated content, with a higher hallucination rate compared with the non-reasoning model. However, a non-negligible portion of hallucination-free answers remains, suggesting that hallucinations in CoTs mostly, but not always, propagate to the final answers. In contrast, we find that non-hallucinated reasoning chains almost always yield hallucination-free answers. This provides preliminary evidence that hallucinations can be effectively mitigated by improving the quality of reasoning chains.

Summary 2. Hallucinated CoTs increase the likelihood of hallucinations in final answers, whereas reliable reasoning chains strongly suppress such errors. Accordingly, we design a novel preference optimization approach with a reasoning-centric objective for enhanced CoT chains.

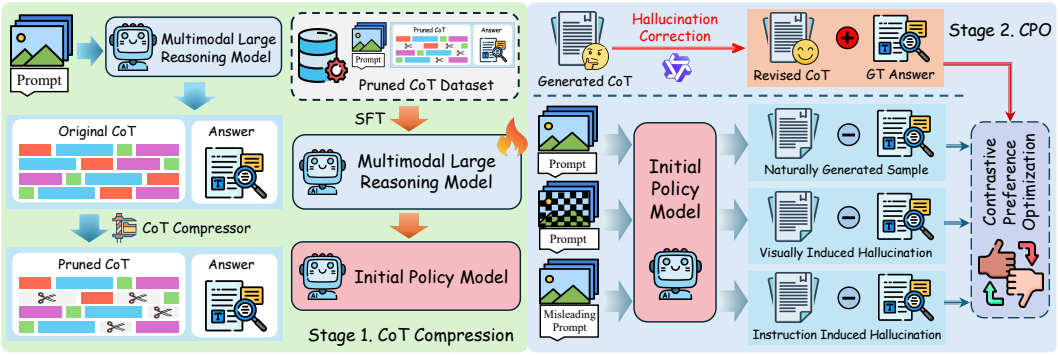


Figure 3: Overview of the proposed C3PO framework. We first construct SFT datasets by removing redundant low-importance tokens in generated CoTs. Next, we perform hallucination-aware preference optimization, where preferred samples contain CoTs enhanced by an advanced open-source MLLM, while the original unmodified outputs serve as rejected ones. To further enhance the optimization, we propose a multimodal hallucination-inducing technique that crafts degraded visual inputs and misleading prompts to elicit MLRM's inherent hallucination patterns as negative contrasts.

3 Method

3.1 Chain-of-Thought Compression

Based on our preliminary analysis, we aim to remove low-information reasoning tokens from the reasoning chain to obtain a more compact and informative intermediate representation, which aligns with the IB principle [9] and is also motivated by the over-thinking issue in reasoning models [17, 18]. Inspired by recent findings that LLMs can perform reliably with pruned CoTs [15], we implement compression by training the model to reason and predict based on trimmed CoT traces.

Data Construction. Given a multimodal reasoning model f_θ, we query f_θ with image–question pairs (v, x) to generate reasoning chain z and final answer y. We then identify reasoning tokens with limited impact on model predictions based on LLMlingua-2 [19], a token-level importance scoring model trained with GPT-4 annotations that has been widely adopted in compression studies [19, 15]. Specifically, LLMlingua-2 is capable of assigning higher importance scores to semantically informative tokens (*e.g.*, object attributes and scene descriptions in image captioning), while lower scores to function words and transitional tokens with limited semantics and minimal influence on model decisions (See examples in Appendix G.1). Tokens are then ranked by importance, and the top γ percentile is retained to form the pruned chain z'. We adopt a threshold of $\gamma = 90\%$ to reduce reasoning redundancy while preserving sufficient information to maintain clean performance. The resulting dataset of N samples is denoted as $\mathcal{D}_s = \{(x_{(i)}, v_{(i)}, z'_{(i)}, y_{(i)})\}_{i=1}^{N}$.

Model Training. We then use the constructed dataset $\mathcal{D}_s$ to perform LoRA-based SFT on the MLRM, which trains the model to reason with compressed CoTs and generate faithful answers. Let $l_{z'}$ and l_y denote the sequence lengths of z' and y respectively, the training objective can be expressed as:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(v,x,z',y)\sim\mathcal{D}_s} \sum_{t=1}^{l_{z'}} \log p_\theta(z'_t \mid v, x, z'_{<t}) + \sum_{t=1}^{l_y} \log p_\theta(y_t \mid v, x, z', y_{<t}). \tag{1}$$

Notably, this SFT process does not introduce any additional supervision information beyond the model's own outputs. We simply prune the generated reasoning chains and pair them with the original, unmodified answers for training. As shown in Appendix C, this strategy already brings a notable reduction in hallucinations, confirming the rationality of our analysis and design.

3.2 Contrastive Preference Optimization

Motivated by the preceding causal analysis, we propose a reasoning-oriented preference optimization strategy that constructs high-quality preference data pairs to effectively strengthen the model's reasoning quality, which then facilitates the generation of hallucination-free answers.

To obtain reliable CoTs as preferred samples, we adopt the well-established RLAIF paradigm that constructs data by incorporating high-quality feedback from advanced MLLMs. We decompose the refinement into sentence-level hallucination detection and correction [20], which simplifies the task and enables open-source MLLMs to provide more reliable guidance. Specifically, we condition the advanced Qwen3-VL [21] on the input image v, the question x, and the ground-truth answer y_{GT}, and instruct it to identify and correct hallucinated content in the reasoning chain z_{Gen} generated by the MLRM. Accordingly, the refined reasoning chain z_{Rev} is then used as the preferred sample, and the original unmodified reasoning chains z_{Gen} that may contain hallucinations serve as inferior ones.

Multimodal Hallucination-Inducing Contrast. Beyond crafting high-quality positive samples, we further conduct a reverse-thinking analysis. While it is challenging to consistently derive hallucination-free responses from a given MLRM, eliciting the opposite hallucinated predictions is considerably easier. Such hallucinated samples are also highly informative as they explicitly expose the model's intrinsic hallucination patterns, which can serve as negative references for contrastive correction.

Therefore, we propose a multimodal hallucination-inducing mechanism that carefully crafts both visual and textual inputs to elicit outputs with rich hallucinated content. As shown in Fig. 3, we induce visual hallucinations by applying random masks to input images, which diminishes substantial ground-truth visual evidence and implicitly forces the model to rely on language priors, resulting in semantically coherent yet hallucinated CoT chains and answers $(z_{\mathrm{img}}, y_{\mathrm{img}})$. For text modality, we design a misleading prompt (see Appendix B.1) that directly instructs the model to prioritize its language priors while partially ignoring the visual input to create a plausible-sounding hallucinated reasoning trace and answer $(z_{\mathrm{ins}}, y_{\mathrm{ins}})$. These two mechanisms effectively simulate MLRM's inherent hallucination behaviors, encouraging the model to perceive and correct such failure modes.

Reasoning-Enhanced Optimization. Given the preference dataset $\mathcal{D}_p$ constructed by the above operations, we strengthen the reasoning process by introducing a reasoning-centric preference loss, which encourages the model to favor reliable reasoning trajectories over inferior ones:

$$\mathcal{L}_{\mathrm{RE}} = -\mathbb{E}_{(v,x,z_{\mathrm{Rev}},\mathcal{N}_z)\sim\mathcal{D}_p} \sum_{z_l\in\mathcal{N}_z} \log\sigma\Big[\beta\log\frac{p_\theta(z_{\mathrm{Rev}}|v,x)}{p_{\theta_{\mathrm{ref}}}(z_{\mathrm{Rev}}|v,x)} - \beta\log\frac{p_\theta(z_l|v,x)}{p_{\theta_{\mathrm{ref}}}(z_l|v,x)}\Big], \quad \text{where } \mathcal{N}_z = \{z_{\mathrm{Gen}}, z_{\mathrm{img}}, z_{\mathrm{ins}}\}. \tag{2}$$

Here, the θ_{ref} denotes the parameters of the reference model. This design provides direct preference supervision on the reasoning process, which effectively stabilizes and facilitates the generation of faithful CoT chains, hence significantly reducing hallucinations in final answers.

However, optimizing reasoning preferences alone does not guarantee that the outputs consistently follow the desired *(reasoning, answer)* format, and may even degrade the coherence and quality of the final answer. To stabilize holistic generation behavior and explicitly improve answer quality, we further adopt a standard DPO objective that models preferences over the joint output (z, y):

$$\mathcal{L}_{\mathrm{DPO}} = -\mathbb{E}_{(v,x,z_{\mathrm{Rev}},y_{\mathrm{GT}},\mathcal{N})\sim\mathcal{D}_p} \sum_{(z_l,y_l)\in\mathcal{N}} \log\sigma\Big[\beta\log\frac{p_\theta(z_{\mathrm{Rev}}, y_{\mathrm{GT}} \mid v,x)}{p_{\theta_{\mathrm{ref}}}(z_{\mathrm{Rev}}, y_{\mathrm{GT}} \mid v,x)} - \beta\log\frac{p_\theta(z_l, y_l \mid v,x)}{p_{\theta_{\mathrm{ref}}}(z_l, y_l \mid v,x)}\Big], \quad \text{where } \mathcal{N} = \{(z_{\mathrm{Gen}}, y_{\mathrm{Gen}}), (z_{\mathrm{img}}, y_{\mathrm{img}}), (z_{\mathrm{ins}}, y_{\mathrm{ins}})\}. \tag{3}$$

By additionally conducting preference optimization over the holistic output distribution, we guide the model to generate well-formed and faithful responses with both robust CoTs and final answers. Furthermore, existing studies [22] have observed that the probability of preferred responses may decrease because DPO essentially focuses on relative preferences. Similarly, we introduce the anchor regularization to constrain the positive reward above a predefined threshold δ:

$$\begin{aligned}\mathcal{L}_{\mathrm{Anc}} = -\mathbb{E}_{(v,x,z_{\mathrm{Rev}},y_{\mathrm{GT}})\sim\mathcal{D}_p}\Big[&\log\sigma\Big(\beta r_\theta(z_{\mathrm{Rev}} \mid v,x) - \delta\Big) \\ &+ \lambda_{\mathrm{DPO}}\log\sigma\Big(\beta r_\theta(z_{\mathrm{Rev}}, y_{\mathrm{GT}} \mid v,x) - \delta\Big)\Big],\end{aligned} \tag{4}$$

where λ_{DPO} balances the contributions of two loss terms. With all the above loss terms, the overall optimization objective can be formulated as:

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{RE}} + \lambda_{\mathrm{DPO}}\mathcal{L}_{\mathrm{DPO}} + \lambda_{\mathrm{Anc}}\mathcal{L}_{\mathrm{Anc}}, \tag{5}$$

where λ_{Anc} controls the influence of anchor loss term $\mathcal{L}_{\mathrm{Anc}}$.

3.3 Theoretical Analysis

In addition to empirical analysis, we build a theoretical framework to validate the effectiveness of C3PO. Let X denote the input, Y the ground truth response, and Z the reasoning trace. Note that $Y \to X \to Z$ is a Markov chain, where Z serves as the intermediate representation between input and output. Based on the Information Bottleneck theory [9, 10, 23], a well-trained model should:

- Minimize $\mathcal{I}(X; Z)$, *i.e.*, the mutual information between the model input X and intermediate representation Z. This urges the model to compress redundant information from the input, thereby improving the information density of the reasoning chain Z.
- Maximize $\mathcal{I}(Y; Z)$, *i.e.*, the mutual information between the ground truth Y and intermediate representation Z. This encourages the model to retain task-relevant information in the CoTs, such as critical visual cues necessary for producing the correct answer.

Overall, the Information Bottleneck objective is defined as

$$\mathcal{L}_{\text{IB}}(Z) = \mathcal{I}(X; Z) - \lambda_{\text{IB}}\mathcal{I}(Y; Z), \tag{6}$$

where $\lambda_{\text{IB}} > 0$ is a hyperparameter that controls the trade-off between compression of redundant information from the input and preservation of predictive information about the ground truth. A smaller $\mathcal{L}_{\text{IB}}(Z)$ corresponds to a better trade-off, yielding an intermediate representation that is both compact and highly informative for prediction. In the considered scenario, we focus on the regime $\lambda_{\text{IB}} > 1$ to prioritize accuracy over compression, as suggested in a previous study [24].

Formally, we justify the effectiveness of our proposed two key strategies in reducing the IB objective by the following two theorems, respectively:

Theorem 1 (Compression reduces IB)**.** *Let the reasoning chain Z be decomposed into Z_{ret} and Z_{trim}, where Z_{ret} denotes the retaining task-critical part, Z_{trim} represents the trimmed redundant part, $Z_{\text{ret}} \cup Z_{\text{trim}} = Z$ and $Z_{\text{ret}} \cap Z_{\text{trim}} = \emptyset$. Then, the IB objective satisfies*

$$\mathcal{L}_{\text{IB}}(Z) \geq \mathcal{L}_{\text{IB}}(Z_{\text{ret}}). \tag{7}$$

Theorem 2 (Enhancement reduces IB)**.** *Let the reasoning chain Z be refined to Z' via our reasoning-enhanced optimization, which supplements additional information about the ground truth Y. Then, the IB objective satisfies*

$$\mathcal{L}_{\text{IB}}(Z) > \mathcal{L}_{\text{IB}}(Z'). \tag{8}$$

Detailed proofs are provided in Appendix A. Theorems 1 and 2 reveal that both compressing and enhancing the CoT reduce the model's information bottleneck objective, which facilitates a more compact and faithful intermediate reasoning, thereby leading to more reliable final answers.

4 Experiments

To reveal the superiority of our method, this section provides comprehensive experiments on various MLRMs across different evaluation benchmarks. Due to page limits, more experimental results and visualization comparisons are provided in Appendix C and G, respectively. The code is available at: https://github.com/ffhibnese/MLRM_Understanding_Mitigating_Hallucination.

4.1 Experimental Setup

Models and Datasets. To verify the effectiveness of our method, we conduct experiments on five representative MLRMs: Orsta-R1-7B [25], ThinkLite-7B [2], MM-Eureka-7B [4], MM-R1-7B [26], and R1-Onevision-7B [3]. To construct training data, we randomly sample 20k data pairs from the RLAIF-V dataset [27] as seed data. The CoT compression ratio is 90% by default. We apply a random masking ratio of 0.3 to the visual input for hallucination-inducing responses. More details, such as the designed *misleading prompt*, are in Appendix B.1.

Evaluation Benchmarks. We evaluate the performance across a wide range of benchmarks. For hallucination evaluation, we consider (1) the CHAIR metric [12], which evaluates object hallucinations in open-ended caption generation; (2) POPE [28], a discriminative benchmark that evaluates object existence consistency via *yes/no* questions; (3) AMBER [6], a systematic benchmark designed to

Table 1: Comparison of the proposed C3PO with competitive baselines on the CHAIR benchmark. We evaluate on images from MSCOCO. C_S and C_I denotes CHAIR_S and CHAIR_I respectively.

Method	R1-Onevision		Orsta-R1		MM-Eureka		MM-R1		ThinkLite	
	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$	$C_S \downarrow$	$C_I \downarrow$
Vanilla	36.7	8.2	40.7	8.5	44.0	8.6	37.0	7.4	41.7	7.8
RLAIF-V	39.0	8.3	39.7	7.6	41.3	8.5	38.7	7.5	40.3	7.5
OPA-DPO	36.7	7.7	41.7	8.2	41.3	7.7	36.3	7.2	40.0	7.9
C3PO	**35.0**	**7.4**	**35.3**	**6.9**	**40.0**	**7.7**	**35.0**	**7.2**	**38.0**	**7.3**

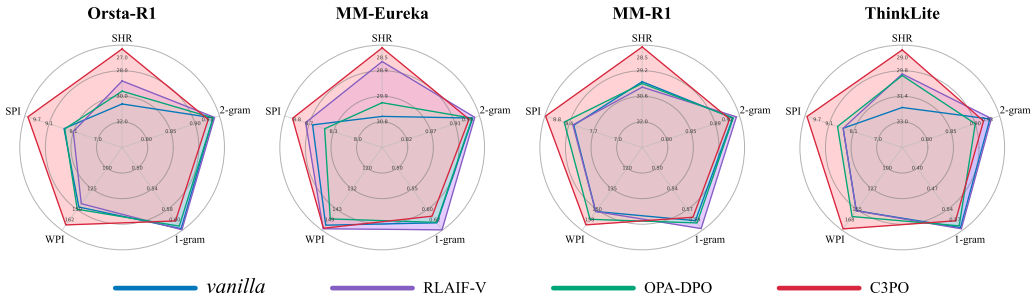


Figure 4: GPT-4 assisted benchmark. Sentence-level Hallucination Ratio (SHR) measures the hallucination degree. We also provide 1&2-gram, the number of sentences per image (SPI), and words per image (WPI). A larger radar area indicates better performance.

assess fine-grained hallucinations; (4) GPT-4 assisted Evaluation [29] that uses GPT-4 series models to compute Sentence-level Hallucination Ratio (SHR) to detect fine-grained hallucinations. In addition, we evaluate the models on two widely used general-purpose benchmarks to assess their overall multimodal capabilities: (1) MME [30], a comprehensive benchmark measuring multiple dimensions of multimodal proficiency; and (2) MMBench [31], a multiple-choice benchmark analyzing complex multimodal comprehension and reasoning, covering a wide range of perception and cognition tasks.

Baselines. To the best of our knowledge, no prior work has specifically addressed hallucination mitigation for MLRMs. Therefore, we adapt two state-of-the-art (SOTA) training-based methods originally designed for MLLMs as baselines, including RLAIF-V [27] with iterative DPO and inference-time scaling and the powerful OPA-DPO [20] that employs on-policy DPO to ensure the alignment of the reference model with the preference data distribution

Implementation Details. In the first SFT stage, we employ Low-Rank Adaptation (LoRA) with a rank of $r = 8$ and a scaling factor of $\alpha = 16$ for 2 epochs with a learning rate of 5e^{-5} and a global batch size $B = 256$. For the subsequent DPO stage, we adopt a learning rate of 1e^{-6} and $B = 32$ for 2 epochs of training, with the KL penalty coefficient $\beta = 0.1$. As suggested in [22, 20], we set $\delta = 0$ in Eq. (4) and $\lambda_{\text{Anc}} = 1.0$ for the anchor loss $\mathcal{L}_{\text{Anc}}$ in Eq. (5). For the hyperparameter λ_{DPO}, we conduct experiments to obtain the optimal value for each model in Appendix C. Following [20, 27], we adopt the greedy decoding for evaluation reproducibility. More details are in Appendix B.

4.2 Performance Evaluation

CHAIR Evaluation. Following [27, 20], we instruct the MLRMs to generate captions on 300 images from the MSCOCO validation set [11] and report results in Table 1. By effectively enhancing the model's reasoning quality, the proposed framework consistently reduces hallucination across all MLRMs, significantly outperforming competitive baselines. *E.g.*, C3PO achieves a 13.27% and 18.82% relative reduction of C_S and C_I on Orsta-R1, respectively.

Note that in contrast to consistent hallucination mitigation achieved by our method, baseline methods may even exacerbate hallucinations in certain cases, *e.g.*, RLAIF-V on R1-Onevision and OPA-DPO on Orsta-R1. This phenomenon reveals the instability of transferring methods for MLLMs to the MLRM setting, further highlighting the necessity and effectiveness of the proposed framework, which explicitly accounts for their unique reasoning process.

Table 2: Comparison of the proposed C3PO with competitive baselines on the POPE benchmark (%). We report the accuracy and F1 score for evaluation.

Model	Method	Random		Popular		Adversarial	
		Acc.	F1 score	Acc.	F1 score	Acc.	F1 score
MM-Eureka	*Vanilla*	84.12	81.90	83.90	81.23	83.47	80.85
	RLAIF-V	83.81	81.46	83.70	80.89	83.27	80.47
	OPA-DPO	85.29	83.46	84.93	82.70	84.37	82.16
	C3PO	**86.22**	**84.68**	**85.87**	**83.94**	**85.23**	**83.34**
MM-R1	*Vanilla*	84.85	82.83	84.47	82.04	83.97	81.56
	RLAIF-V	85.12	83.22	84.70	82.39	84.23	81.95
	OPA-DPO	85.33	83.52	85.00	82.79	84.50	82.31
	C3PO	**86.91**	**85.57**	**86.37**	**84.68**	**85.57**	**83.91**
ThinkLite	*Vanilla*	83.23	80.63	83.03	79.97	82.77	79.72
	RLAIF-V	83.71	81.29	83.37	80.50	83.00	80.14
	OPA-DPO	85.02	83.07	84.70	82.34	84.10	81.77
	C3PO	**85.98**	**84.37**	**85.63**	**83.63**	**84.97**	**83.00**

Table 3: Comparison of the proposed C3PO with competitive baselines on the AMBER benchmark (%). We report the accuracy and F1 score for evaluation.

Method	**R1-Onevision**		**Orsta-R1**		**MM-Eureka**		**MM-R1**		**ThinkLite**	
	Acc.	F1 score	Acc.	F1 score	Acc.	F1 score	Acc.	F1 score	Acc.	F1 score
Vanilla	81.52	86.54	82.94	87.31	83.27	87.15	84.38	88.20	82.25	86.96
RLAIF-V	80.99	84.97	83.14	87.50	83.46	87.49	84.17	88.11	82.31	87.04
OPA-DPO	80.71	85.30	83.58	87.67	83.72	87.59	84.76	88.51	83.02	87.36
C3PO	**81.90**	**86.73**	**84.57**	**88.50**	**85.01**	**88.80**	**85.61**	**89.25**	**83.96**	**88.26**

POPE Evaluation. POPE is another popular evaluation framework for object hallucination, which queries models with "`Is there a <object> in the image?`" to answer a *yes/no* question. These questions are further categorized into Random, Popular, and Adversarial based on the object type (see Appendix B.2 for details). We report the accuracy and F1 scores over 3,000 samples in Table 2. The results indicate that the proposed method facilitates more effective reasoning and enables better perception of the images, substantially reducing hallucination rates across all three question types. Due to space limits, results for R1-Onevision and Orsta-R1 are provided in Appendix C.

GPT-4 Assisted Evaluation. While CHAIR and POPE are two widely used and reliable hallucination benchmarks, they focus exclusively on object hallucination and do not consider more complex positional, relational, or attribute hallucinations. To evaluate these aspects, we adopt the GPT-Assisted Benchmark [29], an LLM-as-judge framework that comprehensively evaluates fine-grained hallucinations using the Visual Genome dataset [32]. Specifically, the MLRM first generates a description for a given image, and the GPT 4-series model is instructed to assess the hallucination level of the generated content based on fine-grained human annotations. We additionally report 1&2-gram metrics for word diversity, as well as SPI and WPI to measure the level of detail.

As shown in Figure 4, C3PO achieves strong overall performance across different MLRMs, significantly reducing fine-grained hallucinations. For instance, C3PO reduces the SHR by 14.05% for Orsta-R1 compared to the *Vanilla* model. Moreover, it substantially improves response detail (higher WPI and SPI), possibly benefiting from reasoning-enhanced optimization that allows a more comprehensive and detailed perception of the visual input.

AMBER Evaluation. AMBER [6] is another challenging benchmark designed to measure models' multi-dimensional hallucinations, including existence, attribute, and relation hallucinations. Following [27], we focus on the discriminative part for an objective and reproducible evaluation and report accuracy and F1 scores on more than 15,000 carefully designed yes/no questions in Table 3. While baselines provide only limited improvements or even degrade performance, the proposed C3PO consistently achieves improvements in hallucination mitigation, again demonstrating the necessity and effectiveness of the proposed framework.

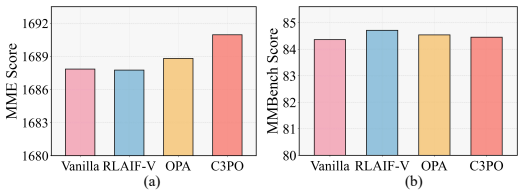


Figure 5: Performance of different methods on two general-purpose benchmarks.

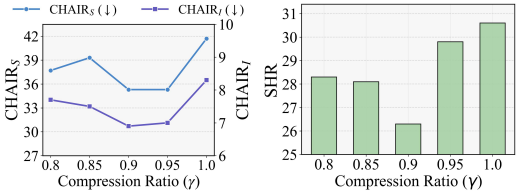


Figure 6: Ablation study of the compression ratio γ on CHAIR and SHR.

MME and MMBench Evaluations. In addition to hallucination benchmarks, we additionally consider two popular general-purpose benchmarks that measure general multimodal capabilities. MME provides a suite of fine-grained and multiple-choice questions across various categories. We align with [33] and report the overall perception score covering 10 multimodal sub-tasks. MMBench is another large-scale benchmark consisting of over 3,000 curated multiple-choice questions. We compute the average score across 20 complex reasoning tasks. We provide results on the representative Orsta-R1 model in Figure 5. As observed, MRLMs trained with the proposed reasoning-oriented strategy generally preserve their multimodal capabilities and even improve over the base model.

4.3 Ablation Study

In this section, we adopt Orsta-R1 as a representative model to investigate the impact of key hyperparameters and quantify the contribution of each proposed technique. Note that we provide ablation using more MLRMs in Appendix C due to page limits.

Ablation on the token compression ratio. During the construction of the SFT data, γ is a critical hyperparameter that controls the number of remaining tokens in the reasoning chain. A larger γ may fail to sufficiently remove noise and redundancy, while a smaller γ may unnecessarily discard informative tokens and thus impair model performance. We provide an ablation study of γ in Figure 6. As shown, $\gamma = 0.9$ achieves optimal performance on both CHAIR and GPT-4 assisted evaluations. Moreover, we observe that pruning rates smaller than 1.0 generally outperform the no pruning ($\gamma = 1$), which again validates the effectiveness of our CoT compression strategy.

Table 4: Ablation study of the proposed techniques in C3PO under CHAIR and GPT-4 assisted evaluation benchmarks.

Method	CHAIR Eval $C_S \downarrow$	CHAIR Eval $C_I \downarrow$	GPT-4 Eval SHR $\downarrow$
C3PO	**35.3**	**6.9**	**26.3**
w/o loss $\mathcal{L}_{\text{RE}}$	37.3	7.2	28.8
w/o loss $\mathcal{L}_{\text{DPO}}$	39.7	8.1	29.4
w/o loss $\mathcal{L}_{\text{Anc}}$	37.7	7.5	27.2
w/o MHI strategy	37.0	7.9	28.4

Ablation on the designed techniques. In addition to CoT compression, we then verify the effectiveness of the other proposed techniques. Specifically, we evaluate the designed loss functions $\mathcal{L}_{\text{RE}}$, $\mathcal{L}_{\text{DPO}}$, and $\mathcal{L}_{\text{Anc}}$, as well as the proposed Multimodal Hallucination-Inducing (MHI) strategy.

Table 4 confirms the contribution of each component, validating the rationale behind our designs. Notably, removing $\mathcal{L}_{\text{DPO}}$ leads to a significant performance drop, showing the necessity of modeling the entire output distribution as a regularization.

5 Conclusion

This work investigates the critical issue of hallucinations in MLRMs. We begin by analyzing two key factors contributing to hallucination emergence, based on which we propose C3PO, a two-stage training-based mitigation framework that first adopts SFT to remove noise and redundancy in generated reasoning trajectories, and then explicitly enhances reasoning via contrastive preference optimization with carefully constructed preference pairs. We theoretically justify the superiority of the proposed techniques from the information bottleneck principle. Extensive experiments across diverse MLRMs and benchmarks validate its effectiveness and generality. We highlight that this work makes a pioneering step in mitigating hallucinations in MLRMs, deepening the understanding of the underlying mechanisms, and inspiring future studies on mitigation strategies.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under grant 62301189, 62576122,62571298, Guangdong Basic and Applied Basic Research Foundation under grant 2026A1515011139.

References

[1] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[2] Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*, 2025.

[3] Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*, 2025.

[4] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, et al. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025.

[5] 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.

[6] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, 2023.

[7] Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *Advances in Neural Information Processing Systems*, 2025.

[8] Bowen Dong, Minheng Ni, Zitong Huang, Guanglei Yang, Wangmeng Zuo, and Lei Zhang. Mirage: Assessing hallucination in multimodal reasoning chains of mllm. *arXiv preprint arXiv:2505.24238*, 2025.

[9] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *ArXiv*, 2000.

[10] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. *Information Theory Workshop (ITW)*, 2015.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[13] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312, 2024.

[14] Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *Advances in Neural Information Processing Systems*, 2025.

[15] Heming Xia, Chak Tou Leong, Wenjie Wang, Yongqi Li, and Wenjie Li. TokenSkip: Controllable chain-of-thought compression in LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3351–3363, Suzhou, China, November 2025. Association for Computational Linguistics.

[16] Sicheng Feng, Gongfan Fang, Xinyin Ma, and Xinchao Wang. Efficient reasoning models: A survey. *arXiv preprint arXiv:2504.10903*, 2025.

[17] Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. Do not think that much for 2+ 3=? on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*, 2024.

[18] Gengyang Li, Yifeng Gao, Yuming Li, and Yunfang Wu. Thinkless: A training-free inference-efficient method for reducing reasoning redundancy. *arXiv preprint arXiv:2505.15684*, 2025.

[19] Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 963–981, 2024.

[20] 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 Computer Vision and Pattern Recognition Conference*, pages 10610–10620, 2025.

[21] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025.

[22] 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.

[23] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *ArXiv*, 2017.

[24] Alexander A. Alemi, Ian Fischer, and Joshua V. Dillon. Deep variational information bottleneck. *International Conference on Learning Representations (ICLR)*, 2017.

[25] Yan Ma, Linge Du, Xuyang Shen, Shaoxiang Chen, Pengfei Li, Qibing Ren, Lizhuang Ma, Yuchao Dai, Pengfei Liu, and Junjie Yan. One rl to see them all: Visual triple unified reinforcement learning. *arXiv preprint arXiv:2505.18129*, 2025.

[26] Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, et al. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources. *arXiv preprint arXiv:2509.21268*, 2025.

[27] 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.

[28] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[29] 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.

[30] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 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*, 2025.

[31] 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.

[32] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[33] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427, 2024.

[34] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[35] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[36] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[37] Hao Fang, Jiawei Kong, Wenbo Yu, Bin Chen, Jiawei Li, Hao Wu, Shu-Tao Xia, and Ke Xu. One perturbation is enough: On generating universal adversarial perturbations against vision-language pre-training models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4090–4100. IEEE, 2025.

[38] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.

[39] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[40] 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.

[41] Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. Improve vision language model chain-of-thought reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1662, 2025.

[42] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9062–9072, 2025.

[43] Hao Fang, Jiawei Kong, Tianqu Zhuang, Yixiang Qiu, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8613, 2025.

[44] Yun Xing, Yiheng Li, Ivan Laptev, and Shijian Lu. Mitigating object hallucination via concentric causal attention. *Advances in neural information processing systems*, 37:92012–92035, 2024.

[45] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

[46] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.

[47] Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023.

[48] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

[49] Jiawei Kong, Hao Fang, Shunxiang Liao, Jinyu Li, Bin Chen, Hao Wu, Shu-Tao Xia, and Min Zhang. Reasoning matters: Mitigate hallucination in multimodal large reasoning models via reasoning-conditioned preference optimization. *arXiv preprint arXiv:2605.27906*, 2026.

[50] Xiaochen Yang, Hao Fang, Jiawei Kong, Yaoxin Mao, Bin Chen, and Shu-Tao Xia. Looking back and forth: Cross-image attention calibration and attentive preference learning for multi-image hallucination mitigation. In *European Conference on Computer Vision*, pages 259–277. Springer, 2026.

A Proofs for Theorems

Here we prove our theorems in Section 3.3.

Theorem 1 (Compression reduces IB)**.** *Let the CoT Z be decomposed into Z_{ret} and Z_{trim}, where Z_{ret} denotes the retaining task-critical part, Z_{trim} represents the trimmed redundant part, $Z_{\mathrm{ret}} \cup Z_{\mathrm{trim}} = Z$ and $Z_{\mathrm{ret}} \cap Z_{\mathrm{trim}} = \emptyset$. Then, the IB objective satisfies*

$$\mathcal{L}_{\mathrm{IB}}(Z) \geq \mathcal{L}_{\mathrm{IB}}(Z_{\mathrm{ret}}). \tag{9}$$

Proof. According to $Z = Z_{\mathrm{ret}} \cup Z_{\mathrm{trim}}$ and the chain rule of mutual information, we have

$$\begin{aligned} \mathcal{I}(X;Z) &= \mathcal{I}(X;Z_{\text{ret}}) + \mathcal{I}(X;Z_{\text{trim}}|Z_{\text{ret}}), \\ \mathcal{I}(Y;Z) &= \mathcal{I}(Y;Z_{\text{ret}}) + \mathcal{I}(Y;Z_{\text{trim}}|Z_{\text{ret}}). \end{aligned} \tag{10}$$

When Z_{ret} is given, Z_{trim} can be relevant to the input X but irrelevant to the ground truth Y, because our compressor only trims low-importance words without contributions to final predictions (see Appendix G.1). Therefore

$$\mathcal{I}(X;Z_{\text{trim}}|Z_{\text{ret}}) \geq 0, \quad \mathcal{I}(Y;Z_{\text{trim}}|Z_{\text{ret}}) = 0. \tag{11}$$

(10) and (11) yield

$$\mathcal{I}(X;Z) \geq \mathcal{I}(X;Z_{\text{ret}}), \quad \mathcal{I}(Y;Z) = \mathcal{I}(Y;Z_{\text{ret}}). \tag{12}$$

Then we conclude that

$$\begin{aligned} \mathcal{L}_{\text{IB}}(Z) &= \mathcal{I}(X;Z) - \lambda_{\text{IB}}\mathcal{I}(Y;Z), \\ &\geq \mathcal{I}(X;Z_{\text{ret}}) - \lambda_{\text{IB}}\mathcal{I}(Y;Z_{\text{ret}}), \\ &= \mathcal{L}_{\text{IB}}(Z_{\text{ret}}). \end{aligned} \tag{13}$$

□

Theorem 2 (Enhancement reduces IB)**.** *Let the CoT Z be refined to Z' via our reasoning-enhanced optimization, which supplements information about the ground truth Y. Then, the IB objective satisfies*

$$\mathcal{L}_{\mathrm{IB}}(Z) > \mathcal{L}_{\mathrm{IB}}(Z'). \tag{14}$$

Proof. Note that the correct answer Y is contained within the input X, which implies

$$\mathcal{I}(X;Z) = \mathcal{I}(X,Y;Z), \quad \mathcal{I}(X;Z') = \mathcal{I}(X,Y;Z'). \tag{15}$$

In addition, according to the chain rule of mutual information, we have

$$\begin{aligned} \mathcal{I}(X;Z) &= \mathcal{I}(Y;Z) + \mathcal{I}(X;Z|Y), \\ \mathcal{I}(X;Z') &= \mathcal{I}(Y;Z') + \mathcal{I}(X;Z'|Y). \end{aligned} \tag{16}$$

Our reasoning-enhanced optimization guides the CoT towards the correct answer, thereby reducing the uncertainty of Y when Z' is given. Formally, it makes

$$\mathcal{H}(Y|Z) > \mathcal{H}(Y|Z'), \tag{17}$$

where $\mathcal{H}$ is Shannon entropy to quantify uncertainty. Then

$$\begin{aligned} \mathcal{I}(Y;Z) &= \mathcal{H}(Y) - \mathcal{H}(Y|Z), \\ &< \mathcal{H}(Y) - \mathcal{H}(Y|Z'), \\ &= \mathcal{I}(Y;Z'). \end{aligned} \tag{18}$$

The information supplemented to Z' by the reasoning-enhanced optimization can essentially be contained within the ground truth Y, which implies that

$$\mathcal{H}(Z'|Z,Y) = 0. \tag{19}$$

Then we have

$$\begin{aligned}\mathcal{I}(X;Z'|Y) &\leq \mathcal{I}(X;Z,Z'|Y),\\ &= \mathcal{I}(X;Z|Y)+\mathcal{I}(X;Z'|Z,Y),\\ &\leq \mathcal{I}(X;Z|Y)+\mathcal{H}(Z'|Z,Y),\\ &= \mathcal{I}(X;Z|Y).\end{aligned} \tag{20}$$

Recall that we focus on $\lambda_{\text{IB}} > 1$ to prioritize accuracy over compression. Then we conclude that

$$\begin{aligned}\mathcal{L}_{\text{IB}}(Z) &= \mathcal{I}(X;Z)-\lambda_{\text{IB}}\mathcal{I}(Y;Z),\\ &= \mathcal{I}(Y;Z)+\mathcal{I}(X;Z|Y)-\lambda_{\text{IB}}\mathcal{I}(Y;Z),\\ &= \mathcal{I}(X;Z|Y)-(\lambda_{\text{IB}}-1)\mathcal{I}(Y;Z),\\ &> \mathcal{I}(X;Z'|Y)-(\lambda_{\text{IB}}-1)\mathcal{I}(Y;Z'),\\ &= \mathcal{I}(Y;Z')+\mathcal{I}(X;Z'|Y)-\lambda_{\text{IB}}\mathcal{I}(Y;Z'),\\ &= \mathcal{I}(X;Z')-\lambda_{\text{IB}}\mathcal{I}(Y;Z'),\\ &= \mathcal{L}_{\text{IB}}(Z').\end{aligned} \tag{21}$$

□

B Experimental Details

This part provides more experimental details regarding data construction and evaluation benchmarks. Note that all the experiments are performed with four 48G NVIDIA A6000 GPUs.

B.1 Data Construction

We randomly sample 20k image-question pairs from the RLAIF-V [27] dataset. Each instance provides an image v, a question x, and a reference answer y_{GT}.

SFT Dataset Construction. For each instance (v, x), we query the target model to generate a structured response that contains an explicit CoT and a final answer. The output is parsed into an original reasoning trace z_{Gen} and an answer y_{Gen} based on the `<think>` and `<answer>` tags. We apply LLMlingua-2 to filter redundant tokens in reasoning traces. By setting the compression ratio to $\gamma = 0.9$, we obtain the compressed reasoning chain $z' = \mathcal{C}(z;\gamma)$. Our SFT data is constructed directly from the compressed reasoning trace paired with the model's original answer. This design reduces redundant textual tokens in CoT, increasing the relative salience of visually relevant evidence during training. We train a reference model using this dataset via SFT, which subsequently serves as the generator for all negative samples in the preference learning stage.

Preference Dataset Construction. To construct the preference dataset, we employ a stronger evaluator Qwen3-VL-30B-Instruct to minimally refine the original reasoning trace and subsequently apply LLMlingua-2 to compress it, yielding the final corrected reasoning chain z_{Rev}. The goal of this step is to correct factual inconsistencies with image content while remaining faithful to the model's stylistic characteristics and reasoning structure. We follow the revision prompt designed in [20], the key refinement instruction is presented as follows:

Refinement Prompt:

Your role is as a discerning assistant tasked with evaluating and refining responses for multimodal tasks. Upon being presented with a question that requires the interpretation of both text and images, you will receive two distinct responses. The first is crafted by our sophisticated multimodal model, while the second represents an approximate ideal answer—it may be incomplete or incorrect. You will also be provided with the images pertinent to the question. Your objective is to meticulously assess these responses. You are to enhance the model-generated response by making precise, minimal modifications that bring it into closer alignment with both the image and the approximate ideal answer. Your revisions should preserve the integrity of the original response as much as possible.

We then construct the positive sample by concatenating this refined chain with the ground truth answer: $z_{\mathrm{Rev}} \,\|\, y_{\mathrm{GT}}$. This ensures that the positive sample acts as a gold standard, being both factually accurate regarding the visual content and structurally optimized for concise and efficient reasoning. For each training instance (v, x), we pair the positive sample with three distinct negative samples derived from the reference model. These negatives comprise: (i) a *natural negative*, generated directly by the reference model given the original image and question, capturing the model's intrinsic tendency to generate non-factual content; (ii) a *visually induced negative*, elicited by applying a random masking ratio of 0.3 to the visual tokens, which forces the model to hallucinate based on language priors due to incomplete visual information; (iii) an *instruction-induced negative*, crafted by appending a specific system prompt that explicitly steers the model towards fabrication:

Misleading Prompt:

Generate a response to the question about the image. Focus on using your language prior to creating a plausible-sounding reasoning trace and answer, but intentionally introduce hallucinations—details that seem related to the image but are not actually supported by it. Ignore the image content when necessary to make the response imaginative yet coherent.

B.2 Evaluation Benchmarks

Object HalBench Evaluations. To evaluate object-level hallucinations in open-ended image description, the CHAIR (Caption Hallucination Assessment with Image Relevance) metric is introduced. This rigorous metric quantifies the discrepancy between objects generated in captions and those present in the ground-truth annotations. Both instance-level (CHAIR_I) and sentence-level (CHAIR_S) scores are reported, where lower values indicate a higher degree of factual consistency. The mathematical formulations for these two variants are defined as follows:

$$C_I = \frac{|\text{hallucinated objects}|}{|\text{all mentioned objects}|}$$

$$C_S = \frac{|\text{captions with hallucinated objects}|}{|\text{all captions}|}$$

AMBER Evaluations. AMBER establishes a versatile framework for evaluating hallucinations in generative and discriminative scenarios. Following [27], we consider the discriminative tasks to assess fine-grained visual grounding for objective and reproducible evaluation. This benchmark formulates hallucination assessment as a rigorous binary classification problem, probing the model with specific "Yes/No" queries to verify visual facts. It comprehensively covers three dimensions of hallucination: (i) *Existence*, verifying the presence of specific objects; (ii) *Attribute*, evaluating detailed properties including object states, counts, and actions; and (iii) *Relation*, assessing spatial relationships and interactions between objects. For R1-Onevision responses that do not begin with Yes/No as instructed, we use Qwen3-VL-30B-Instruct to extract the binary answer. We report both Accuracy and F1 scores to quantify the model's ability to distinguish truthful details from fabrications.

GPT-4 Assisted Evaluations. To quantify the degree of hallucination in detailed image descriptions, the Sentence-level Hallucination Ratio (SHR) is adopted. Following established protocols, models are prompted to describe images from the Visual Genome dataset [32] in detail. The mini version of GPT-4o is utilized as a cost-effective and capable evaluator to identify contradictory statements in the generated text compared to the ground-truth annotations provided by the Visual Genome.

POPE Evaluations. To eliminate the ambiguity often found in free-form generation metrics, the Polling-based Object Probing Evaluation (POPE) reformulates hallucination assessment as a strict binary verification task. An equal distribution of positive and negative samples is enforced to eliminate the influence of prior label frequency. The evaluation includes three distinct difficulty levels: (i) *random* sampling serves as a standard baseline; (ii) *popular* sampling utilizes frequent categories to test resistance against popularity bias; and (iii) *adversarial* sampling employs co-occurring objects to challenge fine-grained visual discrimination. We report Accuracy and F1 scores to evaluate the reliability of object verification across these diverse settings.

Table 5: Additional performance comparison of the proposed C3PO and strong baselines on the POPE benchmark. We supplement results on MLRMs Orsta-R1 and R1-Onevision.

Model	Method	Random		Popular		Adversarial	
		Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
R1-Onevision	*Vanilla*	83.64	81.82	**82.80**	80.59	81.93	79.81
	RLAIF-V	82.85	81.03	81.60	79.45	81.57	79.42
	OPA-DPO	83.85	82.28	82.57	80.67	82.03	80.19
	C3PO	**83.88**	**82.66**	82.53	**81.01**	**82.37**	**80.87**
Orsta-R1	*Vanilla*	85.15	83.29	84.70	82.43	84.13	81.90
	RLAIF-V	85.95	84.32	85.37	83.36	84.57	82.61
	OPA-DPO	86.36	84.88	85.80	83.95	84.90	83.10
	C3PO	**86.98**	**85.67**	**86.27**	**84.62**	**85.03**	**83.45**

Table 6: Ablation study of C3PO on the preference optimization and the CoT compression.

Metric	R1-Onevision			Orsta-R1			MM-Eureka			MM-R1			ThinkLite		
	C3PO	w/o DPO	*Vanilla*	C3PO	w/o DPO	*Vanilla*	C3PO	w/o DPO	*Vanilla*	C3PO	w/o DPO	*Vanilla*	C3PO	w/o DPO	*Vanilla*
$C_S \downarrow$	**35.0**	40.0	36.7	**35.3**	39.7	40.7	**40.0**	40.7	44.0	**35.0**	35.3	37.0	**38.0**	41.3	41.7
$C_I \downarrow$	**7.4**	7.7	8.2	**6.9**	7.7	8.5	**7.7**	8.3	8.6	7.2	**6.9**	7.4	**7.3**	7.2	7.8
SHR ↓	37.9	**36.7**	40.7	**26.3**	29.2	30.6	**28.1**	29.3	30.6	**27.9**	31.0	29.8	**28.5**	29.9	32.1

MME and MMBench Evaluations. To ensure that the hallucination mitigation strategy does not compromise general multimodal capabilities, we conduct evaluations on two comprehensive benchmarks. MME is employed to assess a wide spectrum of perception and cognition tasks. By employing a suite of binary-choice questions with a consistent answer format, MME facilitates a precise measurement of a model's ability to remain faithful to visual content. MMBench is utilized to evaluate complex multimodal comprehension across 20 distinct ability dimensions ranging from basic perception to high-level reasoning. To ensure robustness, it introduces the CircularEval strategy, which requires consistent predictions across multiple permutations of choices to eliminate the influence of random guessing. Together, these evaluations provide an effective validation of the model's ability to preserve general multimodal understanding and instruction-following proficiency.

C Additional Results

POPE evaluation on more MLRMs. As a supplement, we report POPE results on additional MLRMs, including R1-Onevision and Orsta-R1. The quantitative results in Table 5 further corroborate the effectiveness of our method in mitigating object hallucinations, outperforming baseline methods. This validates the necessity and efficacy of our CoT compression and preference optimization.

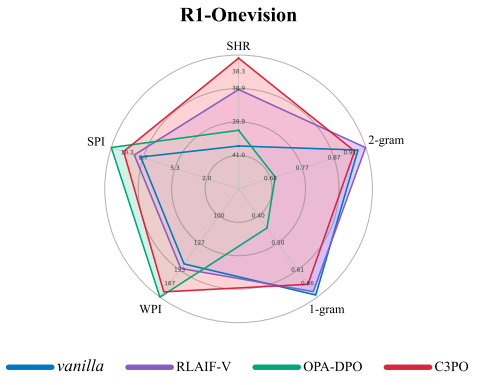


Figure 7: GPT-4 assisted benchmark results for R1-Onevision. A larger radar area indicates better general performance.

GPT-4 assisted evaluation on R1-Onevision. We further supplement the GPT-4 assisted benchmark results on R1-OneVision. As illustrated in Figure 7, the proposed C3PO significantly outperforms existing SOTA baselines. For instance, in comparison with competitive baseline OPA-DPO, C3PO attains a relative improvement of 5.7% on the SHR hallucination metric while maintaining comparable text fluency and details. The results again indicate that the proposed redundant CoT token compression, combined with reasoning-enhanced contrastive preference optimization, synergistically suppresses hallucinations without compromising generation quality.

Ablation on preference optimization and SFT. To evaluate the effectiveness of the preference optimization, we design a variant that undergoes only SFT on compressed text, without subsequent preference optimization. As demonstrated in Table 6, supervised

fine-tuning on compressed text already leads to notable improvements across both evaluation benchmarks. Building upon this foundation, C3PO further achieves reductions of 4.4 and 2.9 in C_S and SHR when equipped with the proposed powerful preference optimization. Overall, these results indicate that the synergy between CoT token compression and DPO allows C3PO to achieve stronger performance gains.

Ablation on the hyperparameter λ_{DPO}. The value of λ_{DPO} is a critical hyperparameter as it controls the contribution of $\mathcal{L}_{\text{DPO}}$ during optimization. We evaluate the performance under various values of λ_{DPO} in Figure 8 to select the optimal value for each MRLM. The results reveal that different models exhibit distinct optimal choices of λ_{DPO}. In particular, we comprehensively consider the performance on both benchmarks to select the optimal values $[0.5, 2.0, 1.5, 2.0, 1.0]$ for R1-Onevision, Orsta-R1, MM-Eureka, MM-R1, ThinkLite, respectively.

D Related Work

D.1 Multimodal Large Reasoning Models

Reasoning has become a prevalent paradigm in LLMs, as the thinking process can significantly enhance models' capabilities of complex problem solving with structured intermediate inference steps [34, 35, 36]. Recent studies extend this paradigm to multimodal models [37] and introduce Multimodal Large Reasoning Models (MLRMs) that conduct an extra reasoning step [38, 39]. To acquire such capabilities, prevalent approaches typically incorporate CoT supervision via supervised fine-tuning (SFT) or reinforcement learning (RL). Early works, including RLHF-V [40], LLaVA-Reasoner [41], and Insight-V [42], primarily leveraged large-scale CoT-style datasets combined with preference optimization to align model behaviors with reliable reasoning patterns. Recent MLRMs typically follow two training paradigms: (i) two-stage SFT + RL strategy (e.g., R1-OneVision [3]). (ii) direct large-scale RL approach (e.g., ThinkLite-VL [2], MM-Eureka [4]). During the RL stage, the powerful Group Relative Policy Optimization (GRPO) [1] is commonly adopted as an effective optimization algorithm for establishing models' reasoning capability. Notably, while models are encouraged to generate reasoning traces to tackle complex tasks, excessively long CoT chains often introduce substantial redundancy. This redundancy not only increases computational overhead but also amplifies error accumulation within the reasoning process, which instead impairs model performance [17, 18].

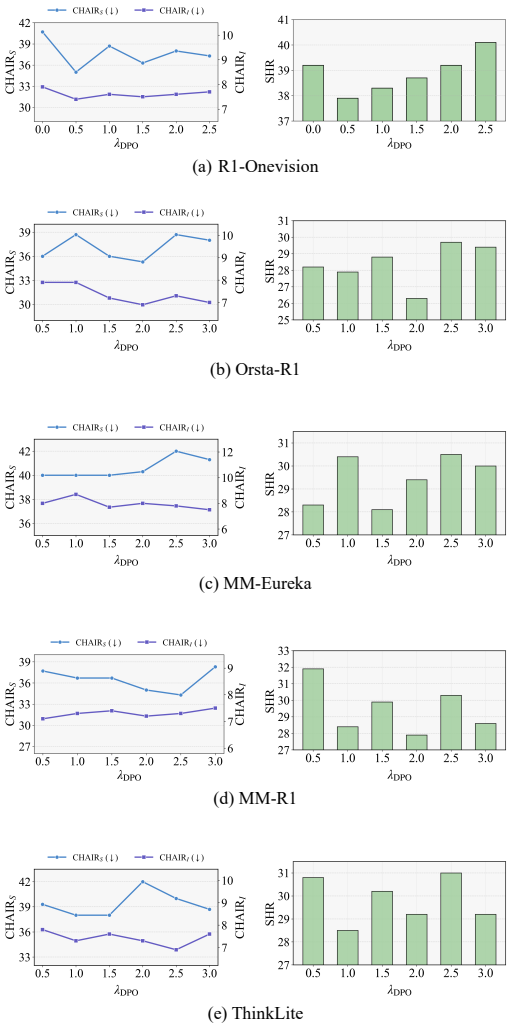


Figure 8: Ablation study on the weight factor λ_{DPO} of the $\mathcal{L}_{\text{DPO}}$ terms.

D.2 Hallucination in MLRMs

Multimodal hallucination, which indicates that generated responses are semantically coherent but contradict the visual input, has been a notorious challenge for multimodal models. The hallucination issue becomes even more prominent in MLRMs as longer reasoning traces shift attention away from image-grounded evidence, amplifying MLRMs' reliance on language priors during generation. [7] designs a systematic benchmark to evaluate the hallucination behaviors of MLRMs and investigate several influential factors. [8] focuses

on self-contradictory hallucinations in reasoning chains and introduces a dedicated benchmark that targets reasoning-stage failures. These findings suggest that hallucinations in MLRMs are tightly coupled with the quality and structure of intermediate reasoning traces. Existing mitigation strategies for regular MLLMs can be broadly grouped into training-free and training-based approaches. Training-free methods generally utilize prompting strategies, post-hoc correction, and improved decoding strategies (*e.g.*, contrastive decoding [43]) to mitigate visual inconsistencies during inference [44, 45, 33, 14]. Training-based approaches enhance output faithfulness by modifying the training pipeline via data curation, SFT, and reinforcement learning [46, 47, 48, 27, 20, 49, 50].

Despite the progress, mitigation strategies for MLRMs in general tasks remain largely unexplored. Existing algorithms are primarily designed for conventional MLLMs and do not account for the distinctive reasoning properties of MLRMs. Consequently, how to utilize the unique reasoning mechanism to improve visual faithfulness in MLRMs remains an open problem. In this paper, we investigate the underlying hallucination causes and present the first training-based framework tailored to MLRMs for effective hallucination mitigation.

E Limitations and Future Work

The CoT compression strategy in our framework is relatively simple and intuitive, which directly removes tokens deemed unimportant by a pretrained LLMlingua-2. Although this approach already yields significant improvements without introducing any additional information, a potentially more effective alternative would involve incorporating reinforcement learning (RL) to encourage the model to generate shorter yet complete CoTs or even adaptively penalizing the length of the reasoning chain specific to the input question and a predefined compression ratio.

Moreover, positive samples used in our optimization are refined by Qwen3-VL-30B-Instruct, indicating that the quality of training data heavily depends on this MLLM's correction capability. While we observe that Qwen3-VL successfully corrects hallucinations in most cases, it still fails to provide satisfactory revisions for a certain subset of samples. Therefore, future work could focus on further improving the quality of preference data, *e.g.*, by introducing a dedicated hallucination-correction module or by leveraging a multi-agent framework for collaborative refinement.

F Impact Statement

This paper makes an early step toward mitigating the serious hallucination issue in MLRMs. The goal is to advance the field of large language models by improving the reliability of MLRM reasoning and reducing hallucinated outputs. There are many potential societal consequences of this line of research, most of which are consistent with those commonly associated with improving the trustworthiness of multimodal language models, and none that we believe require specific discussion here.

G Visualization Results

G.1 Visualization of pruned CoTs.

We visualize the pruned reasoning traces at a mask ratio of 10% in Figure 9. The text marked in red indicates tokens identified as low-importance and are removed during the compression phase. As observed, the removed tokens are predominantly redundant function words (*e.g.*, "that", "is", "the") that contribute little semantic content and are irrespective of the final response. By stripping away these low-information redundancies, our approach forms a more compact and signal-efficient CoT representation that retains task-relevant information while suppressing noise.

G.2 Visualization of Hallucination Mitigation.

The qualitative comparisons in Figure 10 and Figure 11 reveal that baseline methods tend to exhibit hallucinations due to unreliable and low-quality reasoning. In contrast, C3PO conducts compact and reliable thinking and successfully provides the correct answer, effectively suppressing hallucinations.



Instruction

Explain the image's various components in depth.

Pruned CoTs: ~~The~~ image shows a natural outdoor setting with several components ~~that~~ contribute to the overall scenery: 1. Pathway: ~~There is~~ a central pathway made of ~~a~~ natural material, possibly packed earth or stone, ~~which~~ winds through ~~the~~ landscape. It suggests a walkway designed for exploration or recreational purposes. 2. Vegetation: ~~The~~ area is rich with various types of plants. On ~~the~~ left side, ~~there are~~ tall, slender grasses ~~that~~ add texture and movement to ~~the~~ scene. On ~~the~~ right, ~~there~~ are taller shrubs or small trees with golden-yellow flowers, contributing to the seasonal aesthetics. 3. Trees: ~~There are~~ large trees in the background ~~that~~ provide a sense of depth and scale. These trees are leafy and appear to be mature, suggesting ~~that~~ this is a well-established area. 4. Signpost: In the mid-right of the image, ~~there is~~ a signpost ~~which~~ might indicate directions, information about ~~the~~ plants or ~~the~~ route, or it could be a boundary marker. 5. Lighting: ~~The~~ lighting in ~~the~~ image indicates ~~that the~~ photograph was taken during the day under clear weather conditions. ~~The~~ sunlight seems to be illuminating the scene from the upper left, creating a play of light and shadow. 6. Sky: ~~The~~ sky is visible at the top of the image and appears clear with a light blue tone, suggesting a sunny day without visible clouds. All these elements together create a serene and inviting natural environment, possibly a park, botanical garden, or a nature reserve.

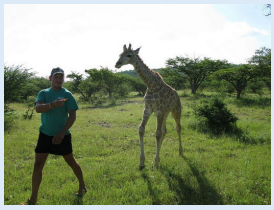


Instruction

Detail the image's components with particular focus on each entity.

Pruned CoTs : ~~The~~ image depicts a rural scene with two horses in the foreground, a fence separating them from a field, and a cluster of houses and trees in ~~the~~ background. ~~Here's~~ a breakdown of ~~the~~ components: 1. Foreground (Left Horse): This is ~~the~~ horse on ~~the~~ left side of ~~the~~ image. It appears to be dark in color. It ~~is~~ standing in a field with a fence in front of it. 2. Foreground (Right Horse): This is ~~the~~ horse on the right side of ~~the~~ image. It appears to be dark in color as well. It ~~is~~ standing in the same field with a fence in front of it. 3. Field: ~~The~~ ground where ~~the~~ horses are standing. It looks grassy with some patches of dirt. 4. Fence: ~~A~~ line of posts and wires runs horizontally across ~~the~~ image, separating the foreground from ~~the~~ background. 5. Background (Houses): ~~There are~~ multiple houses visible, partially obscured by trees. ~~They~~ appear to be two-story buildings with sloped roofs. 6. Background (Trees): Numerous trees ~~are~~ visible in the background, some with bare branches, suggesting ~~it~~ might be winter or early spring. ~~There is~~ also a taller tree to ~~the~~ far right. 7. Sky: ~~The~~ sky is not fully visible but appears to be clear and bright, suggesting daytime.

Figure 9: Visualization of the pruned CoTs.



Question

Is there a sun in this image?

Vanilla:<think>The image shows a person and a giraffe in an outdoor setting with a bright sky. The lighting suggests it is daytime, and the shadows cast by the person and the giraffe indicate the presence of a light source, which is likely the sun. The sky appears clear, which is typical for a sunny day.</think><answer>Yes</answer>

RLAIF-V:<think>The image shows a person standing next to a giraffe in a grassy area with trees in the background. The lighting in the image appears to be natural daylight, which suggests that the sun is likely present, even though it is not directly visible in the image. The shadows cast by the person and the giraffe indicate that the sun is shining from a direction that is not directly visible in the frame.</think><answer>Yes</answer>

OPA-DPO:<think>The image shows a man and a giraffe in a grassy area, but the sun is not visible in the image. However, the presence of shadows and the brightness of the scene suggest that the sun is likely shining. Therefore, it is possible that the sun is in the image. </think><answer>Yes, there is a sun in the image.</answer>

C3PO:<think>image shows a bright, clear sky with no visible sun. lighting suggests it is daytime, but sun itself is not directly visible in the image.</think><answer>No</answer>



Question

Is there a slippers in this image?

Vanilla:<think>The image shows a person sleeping on a bed with two dogs. The person appears to be wearing slippers, which are visible at the bottom of the image. The slippers are gray and are placed near the person's feet.</think><answer>Yes</answer>

RLAIF-V:<think>Upon analyzing the image, there is a person lying on a bed with two dogs. The person appears to be wearing slippers, which are visible at the bottom of the image. The slippers are gray and are placed near the person's feet.</think><answer>Yes</answer>

OPA-DPO:<think>The image shows a person sleeping on a bed with two dogs. The person appears to be wearing slippers, which are visible at the bottom of the image. The slippers are gray and are placed near the person's feet. The dogs are lying on the bed, one of them is looking directly at the camera, and the other is looking to the side. The setting appears to be a bedroom with a bed and curtains in the background. </think><answer>Yes, there is a pair of slippers visible in the image.</answer>

C3PO:<think>image shows a person sleeping on a bed with two dogs. bed has white sheets and a beige blanket. no visible slippers in the image. focus is on person, dogs, and bedding.</think><answer>No</answer>

Figure 10: Visualization results comparing our C3PO and other methods. Hallucinations are marked in red.



Question

Is there a cloud in this image?

Vanilla:<think> The image shows a train on a railway track surrounded by greenery. The sky is visible in the background and appears to be overcast, suggesting the presence of clouds. There are no distinct clouds visible, but the overall appearance of the sky indicates cloud cover. Therefore, it is reasonable to conclude that there are clouds in the image. </think><answer> Yes</answer>

RLAIF-V:<think>The image shows a train on a railway track surrounded by greenery. The sky is visible in the background, and it appears to be overcast, which suggests the presence of clouds. However, the image does not provide a clear view of individual clouds, only the general appearance of the sky. Given the overcast condition, it is reasonable to infer that there are clouds present, even if they are not distinctly visible.</think> <answer> Yes</answer>

OPA-DPO:<think>The image shows a train traveling through a lush, green landscape. The sky is visible in the background and appears to be overcast, suggesting the presence of clouds. There are no distinct shapes or features that would indicate the absence of clouds. </think> <answer>Yes, there is a cloud in the image.</answer>

C3PO:<think>image shows a train on a railway track surrounded by greenery. sky is visible in the background, appears to be overcast with no distinct clouds. lighting suggests a cloudy day, but no specific cloud formations can be identified. Therefore, presence of cloud cannot be confirmed with certainty from the image alone.</think><answer>No</answer>



Question

Is there a glacier in this image?

Vanilla:<think>The image shows two polar bears standing on what appears to be a large piece of ice. The ice looks expansive and is situated in a body of water, which is characteristic of a glacier or a large ice floe. The presence of the ice in a body of water suggests it could be part of a glacier or an ice floe from a larger glacier.</think><answer>Yes</answer>

RLAIF-V:RLAIF-V:<think>The image shows two polar bears standing on a large piece of ice. Glaciers are typically much larger and more permanent features of the landscape, often found in mountainous regions or at higher altitudes. The ice in the image looks more like a chunk of sea ice, which is common in polar regions where polar bears are found. Therefore, it is reasonable to conclude that the image does not depict a glacier.</think><answer> Yes</answer>

OPA-DPO:OPA-DPO:<think>The image shows two polar bears standing on what appears to be a large piece of ice. The ice looks like it could be a glacier or a large ice floe, which is a piece of floating ice that has broken off from a glacier. The presence of the polar bears and the icy environment suggest that this is a cold, Arctic-like setting where glaciers are common. </think> <answer>Yes, there is a glacier in the image.</answer>

C3PO:<think>image shows two polar bears standing on a large expanse of ice. ice appears to be floating on water, is characteristic of a glacier or ice floe. ice is not a glacier itself but rather a large piece of ice floating in water, is often referred to as an ice floe. presence of water and the way ice is floating suggest it is not a glacier.</think><answer>No</answer>

Figure 11: Visualization results comparing our C3PO and other methods. Hallucinations are marked in red.

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: We state our contributions and research scope in the claims presented in the abstract and introduction, all of which are supported by theoretical foundation or extensive experiments.

   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 present the limitations of our paper in section E of the Appendix.

   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: We have provided full proof for the theorems in our 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: All the information of experiments are provided in Section 4 and Appendix B.

   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: We have submitted our code in the supplementary material and will open-source the code and models.

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: We provide exhaustive training and test details in our main paper and code.

   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: [Yes]

   Justification: We provide result analysis across varying values of hyperparameters.

   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: We present these details in the Appendix.

   Guidelines:

   - 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 research conducted in the paper conform every respect of 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 provide the discussion about societal impacts in Appendix 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: The paper poses no such risks.

    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: The license and terms of use are properly respected.

    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: [N/A]

    Justification: No new assets are introduced except code.

    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: The paper does not involve crowdsourcing nor research with human subjects.

    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: This paper does not involve crowdsourcing nor research with human subjects.

    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: In this paper, we does not utilize LLMs for core method development nor as any important, original, or non-standard components.

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.