COPO: Causal-Oriented Policy Optimization for Hallucinations of MLLMs

Supplementary Material

Appendix

The appendix is organized into several sections:

- **Appendix A** provides details for all notations used in this paper.
- **Appendix B** provides the pseudo-code of our method.
- **Appendix C** provides more discussion about the proposed methodology.
- **Appendix D** provides details for the datasets used in this paper.
- **Appendix E** provides details for the baselines mentioned in the main text.
- **Appendix F** provides details for the benchmarks used in the experiments.
- **Appendix G** contains details for the implementation of the experiment.
- **Appendix H** provides the full results and analyses of the experiment.

A. Notations

In this section, we briefly describe the symbols that we mainly use in this article. In **table 4**, we give the definitions of notation according to their role.

B. Pseudo-code

The pseudo-code of causal-oriented reinforcement learning framework is shown in **Algorithm 1** and **Algorithm 2**, mainly showing the steps on the basis of the implementation of base models. The main code is provided in the supplementary materials.

Algorithm 1 Causal Completeness Reward

Input: Token sequence $\mathbf{o} = \{o_1, o_2, \ldots, o_T\}$, ground-truth answer Y, weights λ_s, λ_n
Output: Token-level causal reward $\mathbf{r}_{\text{causal}} = \{r_{\text{causal}}(o_t)\}_{t=1}^{T}$

for $t = 1$ **to** T **do**
  // Causal sufficiency
  Generate answers $\{\tilde{Y}_{(t)}^k\}_{k=1}^H$ from prefix $\mathbf{o}_{<t}$
  Generate answers $\{\tilde{Y}_{(t-1)}^k\}_{k=1}^H$ from prefix $\mathbf{o}_{<t}$
  if $r(\tilde{Y}_{(t)}^k) > r(\tilde{Y}_{(t-1)}^k)$ **then**
    $S_{\text{suff}}(o_t) \leftarrow \frac{1}{H}\sum_{k=1}^H r(\tilde{Y}_{(t)}^k) - \frac{1}{H}\sum_{k=1}^H r(\tilde{Y}_{(t-1)}^k)$
  else
    $S_{\text{suff}}(o_t) \leftarrow 0$
  end if
  // Causal necessity
  Construct a masked sequence where o_t is replaced by a token $\tilde{o}_t$
  Generate answer $\tilde{Y}_{(t)}^{\text{mask}}$ from the masked sequence
  $S_{\text{nec}}(o_t) \leftarrow r(\tilde{Y}) - r(\tilde{Y}_{(t)}^{\text{mask}})$
  // Causal completeness reward
  $r_{\text{causal}}(o_t) \leftarrow \lambda_s S_{\text{suff}}(o_t) + \lambda_n S_{\text{nec}}(o_t)$
end for
return $\{r_{\text{causal}}(o_t)\}_{t=1}^T$

Algorithm 2 Causal-Oriented Policy Optimization

Input: Trajectories $\{\tau_0^i\}_{i=1}^G$ sampled from current policy π_θ; causal-oriented advantage $\hat{A}_{i,t}$; reference policy π_{ref}; hyperparameters β, ϵ
Output: Updated policy π_θ

for each token o_t^i in the sampled trajectories **do**
  Compute importance weight: $\rho_{i,t} \leftarrow \frac{\pi_\theta(o_t^i|s_t^i)}{\pi_{\text{old}}(o_t^i|s_t^i)}$
  Apply clipped weighting: $\Psi(\hat{A}_{i,t}) \leftarrow \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\cdot\hat{A}_{i,t}$
end for
Compute KL penalty: $\mu(\pi_\theta) \leftarrow D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})$
Compute joint objective $\mathcal{J}(\theta)$ as in Eq. (11)
Update policy π_θ by maximizing $\mathcal{J}_{\text{COPO}}(\theta)$

C. More Discussion

C.1. More Details for Motivation Experiment

To probe how such hallucinations emerge, we conduct a controlled toy experiment that contrasts a MLLM with a text-only LLM under the same GRPO optimization settings and inspects their gradient patterns with respect to the input.

Setup. We adopt GRPO post-training for both models to keep the optimization identical. The MLLM is LLaVA [39], which encodes the image with a ViT, projects the image features into the same embedding space as text, and feeds the concatenated visual–text embeddings into a LLaMA decoder. The LLM is LLaMA [16], a text-only Transformer that receives no pixels. We choose this pair because (i) they share the same decoder backbone and comparable generation pipeline, (ii) the only substantive difference is the presence or absence of visual input, which lets gradient contrasts be attributed to the visual pathway, and (iii) both are strong, widely used baselines with public checkpoints. For each example, LLaVA takes the image-text pair (e.g., "Is there a traffic signal light directing traffic?"), and LLaMA takes the same question plus a faithful textual description of

Table 4. The definitions of notations

Notations	Definitions
Notations of Data	*Definitions of Data*
$I^{(n)} = \{I_v^{(n)}, I_t^{(n)}\}$	The paird input with a image $I_v^{(n)}$ and a text prompt $I_t^{(n)}$
$\mathcal{D} = \{(I_v^{(n)}, I_t^{(n)}, Y^{(n)})\}_{n=1}^{N}$	The dataset of paired inputs and corresponding labels with N samples
Notations of Model	*Definitions of Model*
$S_{\text{suff}}(\cdot)$	Causal sufficiency score
$S_{\text{nec}}(\cdot)$	Causal necessity score
$r_{\text{causal}}(\cdot)$	Causal completeness reward function
$r(\cdot)$	Reward model in GRPO
$\pi_\theta(o_t^i \mid I^{(n)}, o_{<t}^i)$	Probability of generating token o_t^i
$\pi_{\theta_{\text{old}}}(o_t^i \mid I^{(n)}, o_{<t}^i)$	Probability generating token o_t^i under the old policy
$\pi_{\theta_{\text{ref}}}$	Reference policy
Notations of Variables	*Definitions of Variables*
I_v	Image input
I_t	Text prompt
Y	Corresponding ground-truth answer
L_c	Set of causal factors
L_s	Set of non-causal factors
$o = \{o_t\}_{t=1}^{T}$	Token sequence generated from MLLM with T tokens
$\tilde{Y} = \{\tilde{y}_1, \cdots, \tilde{y}_{T_{y^i}}\}$	The final answer within o
$\bar{o} = \{o\} \setminus \{\tilde{Y}\} = \{\bar{o}_1, \cdots, \bar{o}_{T_{\bar{o}}}\}$	Reasoning tokens of reasoning steps
$\{o_{(t)}^k\}_{k=1}^{H} = \{\bar{o}_1^k, \cdots, \bar{o}_t^k, \tilde{o}_{t+1}^k, \cdots, \tilde{o}_{T_k}^k\}_{k=1}^{H}$	Token sequences generated by MLLM with $\bar{o}_{\leq t}$
$\tilde{o}_{(\cdot)}^{(\cdot)}$	Tokens which are newly generated
$\bar{o}_{(t)}^{\text{mask}} = \{\bar{o}_1, \ldots, \bar{o}_t^{\text{mask}}, \cdots, \bar{o}_{T_{\bar{o}}}\}$	Token sequence with token o_t is masked
λ_s, λ_n	Hyperparameters for causal sufficiency score and causal necessity score
$\{o^i = (o_1^i, \ldots, o_{T_i}^i)\}_{i=1}^{G}$	Group of G sampled trajectories
r_t^i	Reward of the t^{th} token in the i^{th} trajectory
$r_{i,t}^{\text{causal}}$	Causal completeness reward of the t^{th} token in the i^{th} trajectory
$A_{i,t}^{\text{orig}},\ A_{i,t}$	The GRPO advantage function
$\hat{A}_{i,t}$	The causal-oriented advantage function
$\rho_{i,t},\ R_{i,j}(\theta)$	Importance weight of the t^{th} token in the i^{th} trajectory
Notations of Learning Objective	*Definitions of Learning Objective*
$\mathcal{J}_{\text{GRPO}}(\theta)$	The optimization objective of GRPO
$\mathcal{J}_{\text{COPO}}(\theta)$	The optimization objective of COPO
$\Psi(\hat{A}_{i,t}) = \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) \cdot \hat{A}_{i,t}$	The clipped surrogate objective
$\mu(\pi_\theta) = D_{\text{KL}}\left(\pi_\theta \parallel \pi_{\text{ref}}\right)$	The KL divergence between π_θ and π_{ref}

the visible content. Both produce short factual answers. For analysis, we visualize their gradient distributions. We back-propagate to the visual input to obtain gradient saliency maps on LLaVA. And, we back-propagate to the token embeddings to gradient saliency on LLaMA. All maps are channel-aggregated and min–max normalized for display.

Observation and analysis. **Figure 2** and **Figure 8** shows that gradient patterns differ markedly across models regardless of answer correctness. After GRPO post-training, the MLLM exhibits broader saliency that covers

both task-relevant regions and visually salient but irrelevant background, whereas the LLM's gradient saliency concentrates on semantic elements directly tied to the question (e.g., "white bird", "traffic light"). Under outcome-based rewards, trajectories that yield correct answers while partially relying on background regularities are still positively reinforced, which may strengthen associations between background features and outcomes over time. Consequently, the MLLM may enhance the dependence on the background cues, which may cause hallucinations.

C.2. Intuitive Causal Analysis

To intuitively understand spurious correlations in MLLMs, we analyze concrete examples from **Figure 2** and **Figure 8** through the lens of the SCMs introduced earlier. In our SCM, the ground-truth answer Y (e.g., "Yes, there is a white bird standing on the rock") is determined by a set of causally relevant factors L_c, such as the presence of the bird, its color, and its position on the rock. These causal factors directly contribute to the accurate generation of the correct answer.

However, as demonstrated in **Figure 2** and **Figure 8**, we observe that MLLMs often exhibit significant gradient saliency over irrelevant background features L_s, in addition to the task-relevant foreground features. For example, in **Figure 8a**, when asked "Is there a couch placed on the grass by the roadside?", the MLLM generates the correct answer "Yes, there is a couch placed on the grass by the roadside". However, the gradient saliency map shows that the model allocates considerable gradient saliency to irrelevant background regions such as the kiddle pool, despite the task-relevant foreground containing the couch.

Similarly, in **Figure 8b**, the question "Is there a rider sitting on the ground leaning on a motorcycle?" results in the correct answer "Yes, there is a rider sitting on the ground leaning on a motorcycle". Yet, the model's gradient saliency map reveals that the model is still heavily influenced by background regions, such as the surrounding trees, instead of focusing purely on the task-relevant object, the rider and motorcycle.

Further examples, such as **Figure 8c** and **Figure 8d**, also reveal similar patterns. In **Figure 8c**, the MLLM answers correctly about the presence of a group of people on a boat, but the gradient saliency map shows unnecessary focus on background water areas rather than the people. In **Figure 8d**, the MLLM does not answers correctly about the number of banana basins, its gradient saliency map suggests undue influence from irrelevant background objects. These observations across multiple scenarios consistently demonstrate that MLLMs may rely on spurious correlations, where irrelevant background features unduly influence the model's reasoning process.

These findings emphasize that the presence of spurious correlations in MLLMs, as captured by gradient saliency maps, may play a critical role in generating hallucinations. The model may rely on irrelevant background cues, even when the final answer may be correct. It is necessary to guide MLLMs to focus on relevant information and reduce the influence of background elements, ultimately reducing the likelihood of hallucinations.

C.3. Comparison and Uniqueness Discussion

"Causal sufficiency and necessity" is a foundational concept in causal theory, formally introduced in the book "Causality" [50]. Recent works have explored this concept in various contexts, including CoT-based reasoning in LLMs [76]. To clarify the scope and positioning of our framework and distinguish it from existing studies, we summarize the key differences in modeling assumptions and optimization objectives.

C.3.1. Different Motivations and Core Ideas.

While both methods are inspired by causal theory, our framework is motivated by hallucinations in MLLMs, which are often caused by spurious correlations between task-irrelevant background cues and the final answer. We formulate spurious correlations from a causal perspective and introduce token-level constraints based on causal sufficiency and necessity to guide generation. In contrast, [76] aim to improve interpretability in LLM reasoning, focusing on analyzing the contribution of intermediate reasoning steps. [76] leverage post-hoc attribution to assess whether specific tokens are causally linked to the model's final decision.

C.3.2. Different Methodologies and Experiments.

Our method integrates causal constraints into the optimization through reinforcement learning. We construct a causal completeness reward that encourages the model to consider both sufficiency and necessity tokens for correct answers. This allows the model to gradually shift attention toward more causally relevant information during optimization. In comparison, [76] adopt a causal probing strategy, evaluating fixed model outputs without modifying the learning procedure. [76] propose a causal attribution method intended to interpret model predictions without altering the training procedure. In experiments, our framework is assessed on multiple MLLM benchmarks with hallucination metrics to examine its practical effectiveness in reducing hallucinations. On the other hand, [76] focus on qualitative and interpretability-driven analysis in text-only settings.

In summary, these differences reflect distinct goals and methodological choices. Our approach targets hallucination mitigation in multimodal tasks through reward-based optimization, whereas [76] pursue causal interpretability for language reasoning.

Figure 8. More motivating results. Both MLLM and LLM are trained via GRPO.

C.4. Intuitive Outcome-based Reward Effect

In this section, we provide a more intuitive explanation of the effect of outcome-based reward of GRPO on gradient saliency maps in MLLMs.

We visualize the gradient saliency maps of the MLLM before and after GRPO post-training in **Figure 9**. In the Vanilla MLLM (w/o GRPO), gradient saliency is predominantly concentrated on task-relevant foreground objects, such as the boat, the rider, or the athlete. However, following post-training with standard GRPO, the model exhibits a marked divergence: the gradient saliency over task-irrelevant background regions increases significantly. This indicates that the model's reliance on background regions is not an inherent architectural artifact but a learned behavior developed during the optimization process.

C.5. Intuitive Causal Completeness Reward Effect

In this section, we provide a more intuitive explanation of the effect of the causal completeness reward on mitigating hallucinations in MLLMs.

Hallucinations in MLLMs often arise when the model generates plausible-sounding responses that are not grounded in the visual evidence, typically due to reliance on non-causal background information. **Figure 6** and **10** illustrate this effect by comparing the behavior of an MLLM with and without our proposed COPO. Without COPO, the gradient saliency of the model is typically spread across both task-relevant regions and irrelevant background features. For example, when asked about the presence of traffic lights in **Figure 10**, the MLLM might generate the incorrect response "No, there are no traffic lights", despite background elements like building façades influencing its decision. After applying COPO, the model shifts its attention to task-relevant regions, such as the traffic light itself, producing the correct response "Yes, there are traffic lights in the image".

Additionally, in cases where the MLLM initially provides the correct answer, COPO can further refine the model's reasoning process. For example, in **Figure 2**, the model answers correctly, "Yes, there is a brown baseball glove", but without COPO, the gradient saliency is distributed over both the glove and background regions. After applying COPO, the gradient saliency becomes significantly more concentrated on the foreground object (the

Figure 9. More illustration of outcome-based reward effect.

glove), reinforcing the model's focus on causally relevant features, which may improve its reasoning process.

These results suggest that COPO can make the reasoning of the model more consistent with the causal relevant elements, thus minimizing the influence of irrelevant background information and reducing the likelihood of hallucinations.

C.6. Quantitative Analysis of Causal Completeness Reward Effect

We randomly sampled 500 images from the MSCOCO validation set. We define the Background Saliency Ratio (BSR) as the proportion of the total saliency mass that falls within background (non-object) regions. A lower BSR indicates a reduced reliance on background cues. The results in **Table 5** indicate that the phenomenon in **Figure 6** is a consistent trend across the dataset.

Table 5. Comparison of gradient saliency.

Method	Background Saliency Ratio
Vanilla MLLM	32.4%
MLLM w/ GRPO	41.7%
MLLM w/ COPO	20.8%

C.7. Qualitative Analysis

To better illustrate the behavior of our causal completeness reward, we conduct a qualitative analysis on an example from the image captioning task. As shown in **Figure 11**, the input image depicts a brown and white cat drinking water from a metal faucet in a bathroom sink.

The MLLM w/o COPO identifies several relevant visual attributes such as brown, white, cat, bathroom, sink, and water. These content words obtain moderately high causal rewards, while many tokens associated with the actual action (e.g., drinking, mouth, faucet interaction) are absent from the reasoning steps. As a consequence, the trajectory contains incomplete causal evidence, and its causal reward remains relatively low. The final answer, though partially correct, fails to state that the cat is drinking from the faucet.

For the MLLM w/ COPO, the reasoning steps become more causally aligned with the true visual factors. Tokens capturing critical relations-—such as metal faucet, thin water stream, cat's mouth, and drinks—-now appear explicitly and receive substantially higher causal rewards. Meanwhile, non-informative connectives continue to obtain low scores, showing that the causal reward selectively emphasizes semantically grounded, evidence-bearing tokens. This produces a reasoning trajectory with a higher overall causal completeness.

These results imply that COPO can reliably elevate the causal completeness reward of tokens that genuinely contribute to a correct description, and sequences with higher causal reward correspond to significantly more accurate, evidence-grounded outputs. This demonstrates the effectiveness of COPO.

C.8. Broader Impacts and Limitations

In this subsection, we briefly illustrate the broader impacts and limitations of this work.

Figure 10. More illustration of causal completeness reward effect.

Broader Impacts. This work investigates hallucination mitigation in MLLMs and proposes a causal-oriented policy optimization (COPO) framework. It aims to reduce hallucinations by encouraging outputs that are both causally sufficient and necessary for correct answers. Our approach offers a conceptual bridge between causal inference and hallucinations of MLLMs, potentially inspiring further research into causal-aware generation strategies and intervention-based learning for foundation models. Extensive experiments across diverse benchmarks validate the effectiveness of our COPO in mitigating hallucinations.

Limitations. Our implementation is based on open-source MLLMs and LLMs, and we have not evaluated models at the scale of 72B or above 100B parameters. Future work may explore applying COPO to larger-scale models to further assess its robustness and scalability.

D. Datasets

In this section, we provide a brief overview of the datasets used in our experiments. We utilize publicly available datasets.

- MSCOCO [37] is a large-scale benchmark designed for image captioning, object detection, and scene understanding. The 2014 version contains 82,783 training images and 40,504 validation images, each annotated with five human-written captions describing the image content. The dataset features diverse everyday scenes with multiple objects in context, making it suitable for evaluating visual grounding, object recognition, and multimodal generation tasks. Its high-quality annotations and broad coverage of object categories have made it a standard resource in vision-language research.
- V^* dataset [69] is constructed based on COCO2017, specifically tailored for evaluating visual search and fine-grained grounding capabilities in multimodal models. It features high-resolution natural images where the target objects are often small, densely distributed, and surrounded by clutter. This setting poses a significant challenge for MLLMs, as models must detect and localize fine-scale objects while integrating visual and linguistic cues. The dataset is particularly useful for assessing models' abilities in precise visual grounding and attention allocation within complex scenes.
- ArxivQA [35] is a multimodal question-answering dataset that focuses on scientific diagrams and structured visual representations extracted from research papers on arXiv. The dataset consists of questions grounded in visual elements such as line charts, bar graphs, tables, and flow diagrams. Answering these questions requires cross-modal understanding of the visual layout, semantic parsing of textual annotations, and precise diagram interpretation. This subset effectively evaluates the abilities of

Figure 11. Qualitative analysis of causal completeness reward.

MLLMs to align textual queries with structured visual information and reason over schematic content.

- ThinkLite-VL [67] is designed to benchmark complex multimodal reasoning abilities. It includes diverse question-answering tasks that span several categories such as mathematical reasoning, physical and commonsense inference, temporal understanding, and causal analysis. Each question is paired with multimodal context (image and/or text) that requires step-by-step reasoning rather than shallow pattern recognition. The dataset emphasizes robustness and generalization in reasoning, making it particularly suitable for evaluating models' inference chains and causal understanding in multimodal scenarios.

E. Baselines

In this section, we provide a brief overview of the baselines used in our experiments.

- DoLa [10] enhances factual grounding in MLLMs by leveraging the contrast between shallow and deep semantic features across the model layers. During decoding, it compares early-layer and late-layer activations to identify inconsistencies and promote factually grounded token selection. This mechanism helps suppress hallucinations by ensuring that surface-level visual signals are supported by deeper semantic representations.
- OPERA [23] mitigates hallucinations by identifying and calibrating over-trusted visual tokens. It analyzes self- and cross-attention distributions to locate tokens that disproportionately dominate the model's generation process. By down-weighting such tokens through attention rescaling, OPERA reduces hallucinated content and improves factual alignment in generated text.
- VCD [30] improves visual grounding by enforcing contrastive consistency between relevant and irrelevant visual-text pairs at the representation level. It introduces a contrastive training objective that pulls matched image-text features closer while pushing mismatched ones apart. This contrastive signal enhances the model's ability to focus on visually grounded elements and reduces reliance on hallucinated concepts.
- GPT-4o [1] is OpenAI's flagship MLLM, designed to handle both textual and visual inputs with high fidelity. It incorporates proprietary training strategies and massive-scale data to achieve strong cross-modal alignment, reasoning, and generation capabilities. GPT-4o represents one of the best-performing commercial MLLMs, making it a valuable reference point for evaluating model performance on real-world multimodal tasks, including hallucination robustness.
- Qwen-VL-Max [3] is a high-capacity vision-language

model developed by Alibaba, designed for detailed visual understanding and instruction following. It extends the Qwen-VL family by incorporating larger backbone architectures and fine-grained alignment techniques, supporting multi-turn interaction, high-resolution image inputs, and extended reasoning abilities. Qwen-VL-Max is trained on a broad set of image-text pairs and instruction-following data, and optimized for tasks such as image captioning, VQA, and OCR. Due to its closed-source nature, we use its official public API for inference-based comparison in our evaluation.

- InternVL-1.5 [9] an open-source large vision-language model introduced by Shanghai AI Laboratory. It builds upon previous InternVL versions, incorporating enhanced multi-modal alignment, improved image understanding, and more efficient training techniques. The model excels in zero-shot and few-shot settings, making it a strong benchmark for open-source multimodal research.
- LLaVA-OneVision [32] is an advanced open-source vision-language model that unifies training across diverse image resolutions and visual contexts. It improves upon previous LLaVA models by introducing resolution-aware learning and high-resolution instruction tuning, enabling robust performance on complex visual tasks. As a strong open-source baseline, LLaVA-OneVision provides a reliable comparison point for models targeting high-resolution and fine-grained visual understanding.
- Qwen2.5-VL [4] is a powerful multimodal extension of the Qwen language model series, developed by Alibaba. It incorporates vision encoders aligned with large language models and supports instruction tuning for a wide range of vision-language tasks. With both 7B and 32B variants, Qwen2.5-VL demonstrates strong performance in image captioning, visual QA, and reasoning, serving as a representative of scalable, open-source MLLMs.
- DeepEyes [79] is a vision-language model that emphasizes external tool usage and modular vision-language reasoning. It integrates structured reasoning modules and supports external search or computation for improved factuality and interpretability. DeepEyes has shown effectiveness in reducing hallucinations through hybrid symbolic-neural workflows, making it a particularly relevant baseline for hallucination mitigation studies.
- Grounding DINO [43] is a multi-stage system that integrates object detection and open-set visual grounding within a single framework. By combining powerful transformer backbones with structured reasoning workflows, Grounding DINO excels at referring expression comprehension and open-vocabulary detection, offering precise and interpretable grounding results in complex scenes.
- SEAL [69] is a workflow-based system designed for fine-grained visual understanding, especially in high-resolution visual search tasks. It incorporates structured pipelines including proposal generation, visual grounding, and language reasoning to address small-object detection and disambiguation. SEAL's multi-stage reasoning capabilities offer high accuracy at the cost of complexity, and it serves as a strong non-end-to-end baseline.
- DyFo [33] introduces a dynamic fusion framework that adaptively integrates visual and textual features through modular fusion layers and reasoning controllers. It leverages intermediate symbolic representations to guide multimodal inference, enhancing interpretability and performance on challenging tasks. DyFo represents a hybrid architecture combining structured reasoning and deep learning, useful for assessing generalization under complex scenarios.
- ZoomEye [55] is a multi-stage system that targets high-resolution image analysis by progressively zooming in on relevant image regions. It mimics human attention by narrowing focus to informative areas before performing visual-language reasoning, significantly improving performance on cluttered and detail-rich images. ZoomEye's pipeline exemplifies how structured workflows can improve precision and reduce hallucinations in vision-language tasks.
- DeCo [59] is a training-free, model-agnostic decoding framework proposed to mitigate hallucinations in MLLMs. The core insight of DeCo is that MLLMs often correctly recognize visual objects in their preceding (intermediate) layers, but this information is suppressed in deeper layers due to strong language model priors, resulting in hallucinated outputs. DeCo dynamically selects the most informative preceding (anchor) layer, and proportionally integrates its knowledge into the final layer's output logits during inference. This adjustment recalibrates the predicted token probabilities, enhancing the accuracy of object descriptions while reducing hallucination rates. DeCo is compatible with various decoding strategies—such as greedy search, nucleus sampling, and beam search—and can be seamlessly applied to different MLLMs without additional training. Extensive experiments show that DeCo significantly reduces hallucinations on image captioning and visual question answering benchmarks, with only a modest increase in inference latency compared to baseline methods.
- VTI [44] introduces a method called Visual and Textual Intervention (VTI), designed to reduce hallucinations in large vision-language models (LVLMs). VTI works by steering the latent space representations of both the vision encoder and text decoder during inference. This test-time intervention enhances the stability of vision features and improves the alignment between vision and text, without the need for additional training. The technique involves pre-computing the directions of feature shifts in the latent space based on a set of training examples. These

directions are then applied consistently to all queries during inference. The method is task- and dataset-agnostic, making it widely applicable without incurring additional computational costs. VTI has shown significant effectiveness in reducing hallucinations across multiple evaluation benchmarks.

- HA-DPO [78] is a fine-tuning method designed to reduce hallucinations in Large Vision-Language Models (LVLMs). The key idea is to treat hallucination mitigation as a preference learning problem: for each image, the model is provided with a pair of responses—one faithful and one hallucinated—and DPO is used to directly optimize the model to prefer the faithful response. To construct training data, the authors curate paired hallucination/non-hallucination outputs with consistent style, and apply lightweight LoRA fine-tuning on multiple LVLM architectures such as MiniGPT-4, InstructBLIP, and LLaVA-1.5. Experiments on standard hallucination benchmarks (e.g., POPE, SHR) show that HA-DPO effectively suppresses visual hallucinations and improves response reliability.
- CLIP (OHD) [45] presents a focused analysis of object-hallucination phenomena in the commonly used visual encoder CLIP, which serves as the backbone for many large vision-language models (LVLMs). The authors first introduce a dedicated benchmark called OHD-Caps (Object Hallucination Detection-Caps) comprising paired positive captions and counterfactual negatives (captions with nonexistent or removed objects) drawn from datasets such as COCO, Flickr30K and NoCaps. Their empirical findings show that even the CLIP encoder in isolation is prone to mistaking hallucinated objects, indicating that hallucination is not purely a vision-language fusion issue. To mitigate this, they apply a counterfactual data-augmentation strategy to fine-tune the CLIP encoder using a fine-grained object-level contrastive loss on the OHD-Caps dataset, showing large reductions in object-hallucination rates and improvements when the tuned encoder replaces the vanilla CLIP in downstream LVLMs.
- POVID [81] presents a novel approach for reducing hallucination in vision-language models (VLLMs) by explicitly aligning the visual and language modalities through preference-based fine-tuning. The authors propose the method named POVID, which generates preference pairs by first using a large multimodal model (e.g., GPT-4V) to inject plausible hallucinations into correct responses, then distorting input images to trigger inherent hallucination behaviours of the VLLM. These preference pairs (faithful vs hallucinated) are then used to fine-tune the model via Direct Preference Optimization (DPO). Experimental results across multiple image-instruction benchmarks show that this modality alignment strategy reduces hallucination error rates and improves overall generation reliability.
- CSR [82] introduces a novel mechanism for improving the alignment of visual and language modalities in large vision-language models (LVLMs) by using self-rewarding fine-tuning. Specifically, the authors propose to generate model-internal reward signals calibrated to identify hallucination versus faithful outputs, and then use these self-derived rewards to fine-tune the model without reliance on extensive human annotation. Through experiments on image-instruction benchmarks, the paper demonstrates that the self-rewarding framework reduces hallucination rates and enhances output fidelity while maintaining inference efficiency.
- GCPO [17] introduces a novel RL-style fine-tuning method designed for large-scale language models that addresses the limitations of existing group-wise preference learning approaches. The key innovation is to integrate causal structure among candidate responses by (1) applying a causally informed reward adjustment that accounts for interactions (e.g., complementarity or contradiction) among grouped candidates; and (2) incorporating a KL-regularization term aligning the policy with a causally projected reference distribution. Extensive experiments on multiple reasoning benchmarks demonstrate that GCPO consistently outperforms prior methods such as Group Relative Policy Optimization (GRPO) by achieving better calibration and reward alignment.

F. Benchmarks

In this section, we provide a brief overview of the benchmarks used in our experiments.

- CHAIR [53] is a standard metric for measuring object hallucination in image captioning. It compares model-generated captions against ground-truth object annotations from MSCOCO, quantifying hallucination at two levels: sentence-level (CHAIR_S), which computes the fraction of captions containing hallucinated objects, and instance-level (CHAIR_I), which measures the fraction of hallucinated object mentions among all mentioned objects. Following previous works [23], we use 500 randomly sampled images from the MSCOCO 2014 validation set and adopt the fixed prompt "Please help me describe the image in detail."
- Perception-Oriented Perturbation Evaluation (POPE) [36] introduces controlled adversarial and semantic perturbations into visual inputs to assess hallucination sensitivity and robustness. It contains three subsets—Adversarial, Popular, and Random—that evaluate whether a model can avoid producing hallucinated content when facing visually misleading or out-of-distribution inputs. POPE is a standard benchmark for hallucination detection and mitigation evaluation.
- MME [15] is a diagnostic benchmark focusing on fine-

grained multi-modal capabilities, particularly in assessing text generation quality across sub-skills such as attribute grounding, object counting, relation understanding, OCR, and commonsense reasoning. For each model, we extract the averaged score across all text-related sub-skills. Evaluation is based on matching predicted answers against ground-truth labels using the official toolkit.

- GPT-4o Assistance [1] is an open-ended caption assessment benchmark designed to evaluate model performance in factuality, coherence, and informativeness of generated descriptions. It uses GPT-4o to compare captions from different models on a set of 100 randomly selected images from the MSCOCO 2014 validation set. For each image, two model outputs are compared side-by-side, and GPT-4o is prompted to assess them along three axes: Accuracy (A), measuring the factual correctness of the description; Correctness (C), assessing logical consistency; and Detailedness (D), capturing the richness of content. This evaluation follows the same format as [23, 30] using a standardized prompt template.
- V^* Bench [69] is a high-resolution benchmark built on the COCO2017 dataset, specifically curated for fine-grained object attribute and spatial relation evaluation. It contains complex scenes with small, dense visual targets and requires precise object localization and captioning under high-resolution settings. This benchmark is designed to test a model's ability to maintain grounding accuracy and attribute fidelity in visually cluttered scenarios.
- HR-Bench(4K/8K) [66] provides a series of vision-language evaluation tasks based on ultra high-resolution images at 4K and 8K scales. The benchmark includes factual QA, reasoning, and captioning tasks over visually complex environments, often involving minute object details. It challenges models to maintain accurate semantic understanding under extreme resolution and compositional complexity, making it ideal for evaluating resolution scalability and hallucination resistance.
- refCOCO [6] is a benchmark designed for evaluating object grounding via referring expressions in natural images. Each instance provides an image and a natural language expression referring to a specific object, and the task is to localize the correct object. The expressions in refCOCO often include both appearance and spatial relations, making it a strong testbed for assessing multimodal models' grounding fidelity and cross-modal understanding under relatively simple sentence structures.
- refCOCO+ [6] is a variant of refCOCO that restricts referring expressions to exclude spatial terms. This forces models to rely more heavily on visual attributes—such as color, size, and object type—rather than relative positioning. As a result, refCOCO+ is particularly useful for evaluating a model's capacity to attend to fine-grained visual details and object characteristics when grounding textual descriptions.
- refCOCOg [26] extends the previous benchmarks by providing longer and more descriptive referring expressions that often include multiple attributes and complex sentence structures. Unlike refCOCO and refCOCO+, it contains full-image annotations rather than cropped regions, increasing contextual ambiguity. This makes refCOCOg a challenging benchmark for assessing how well multimodal models can handle long-range dependencies and integrate multiple pieces of information for accurate grounding.
- ReasonSeg [28] is a recent benchmark that evaluates visual segmentation tasks requiring multi-hop reasoning. Each instance includes an image, a question, and a target segmentation mask, requiring the model to understand complex dependencies between visual regions and linguistic cues. It is particularly useful for assessing how well a model can integrate reasoning with pixel-level grounding and visual comprehension.
- MathVista [47] is a multimodal benchmark composed of visual math problems that require understanding diagrams, plots, or geometric figures. It tests spatial reasoning and diagram interpretation in mathematical contexts, making it particularly suitable for evaluating visual grounding in structured mathematical domains. MathVista highlights the challenge of aligning symbolic and visual information for accurate problem-solving.
- MathVerse [77] provides a broad spectrum of math-related visual-language problems ranging from elementary arithmetic to higher-order logic. It combines equation understanding with diagram-based inputs, evaluating the model's ability to perform compositional reasoning across modalities. MathVerse is a robust benchmark for assessing general-purpose multimodal mathematical reasoning.
- MathVision [65] focuses on assessing visual mathematical reasoning through tasks such as geometry, bar charts, and algebraic diagram interpretation. It contains highly structured problems where success depends on identifying visual evidence and integrating it with symbolic operations. MathVision is particularly valuable for evaluating reasoning robustness and multi-hop inference over structured inputs.
- WeMath [51] introduces a diverse set of visual math challenges involving open-ended questions that mimic real-world math education scenarios. The tasks test a model's capacity for high-level reasoning, multi-step solution planning, and causal inference based on visual stimuli. WeMath promotes evaluation under pedagogical constraints and human-like reasoning processes.
- DynaMath [84] emphasizes dynamic visual reasoning over animated or multi-frame visual content. It includes

motion-based problems that require temporal reasoning, causal tracking, and spatial transformation understanding. DynaMath serves as a comprehensive benchmark for temporal multimodal reasoning beyond static-image settings.

- LogicVista [71] targets visual logic reasoning through a collection of puzzles and structured visual problems. It focuses on logic rule inference, pattern completion, and counterfactual reasoning, challenging models to deduce correct answers through symbolic and visual clues. LogicVista evaluates higher-order causal and logical reasoning in visually constrained contexts.

G. Implementation Details

Our implementation builds upon open-source reinforcement learning frameworks, modified to incorporate a causal completeness reward for hallucination mitigation. We conduct training on four representative open-source vision-language models: InstructBLIP [11], MiniGPT-4 [83], LLaVA-1.5 [40], and Qwen-VL [3]. All models are initialized from publicly released Hugging Face checkpoints. Training is conducted on A100 GPU clusters using FSDP with parameter and optimizer offloading enabled to ensure memory efficiency. We use vLLM for rollout generation, with GPU memory utilization capped at 60% and support for long-sequence prefill enabled.

The training follows the GRPO setup, with each batch consisting of 512 sampled prompts and a mini-batch size of 256. We use a causal completeness reward function that balances causal sufficiency and necessity, with weights $\lambda_s = 0.35$ and $\lambda_n = 0.35$. The maximum prompt length is set to 1,024, and the maximum response length is extended to 8,192 tokens to support complex reasoning. The KL divergence regularization is enabled with a low-variance KL loss, and the reference model shares the same architecture as the actor. Learning rate is fixed at $1e^{-6}$. COPO is applied in an offline manner [31]. Rollouts and their corresponding causal rewards are collected before optimization, effectively decoupling this overhead from the training loop. During optimization, the model efficiently accesses these precomputed values. The computational complexity per training step remains acceptable. The training phase matches GRPO, as COPO merely incorporates a modified advantage term using fixed batches. The preprocessing phase incurs a one-time labeling cost which is parallelizable and scales linearly with the dataset size, rendering it independent of the number of training epochs. All experiments are executed with mixed precision. Additionally, we enforce a two-stage generation strategy via specific prompts. The model is instructed to generate reasoning tokens first—comprising a comprehensive description or analysis of the visual input—which serve as the causal basis for the subsequent answer tokens. This structural distinction between intermediate reasoning and the final answer is widely adopted in multimodal tasks to ensure logical consistency.

For evaluation, we follow the exact protocol and data splits from prior works. For CHAIR [53], we use 500 images from the MSCOCO 2014 validation set with the fixed prompt "Please help me describe the image in detail", reporting both $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_I}$ metrics. For POPE [36], we follow the VQA-style evaluation on 500 MSCOCO images, using six structured object-centric questions per image across three splits (random, popular, adversarial), and report averaged F1 scores. The base reward $r \in \{0, 1\}$ is a binary score: (i) Captioning tasks: we utilize string matching against the standard 80 MSCOCO object categories and their synonyms. Both the generated text and ground-truth annotations undergo normalization (including lowercasing and lemmatization). If a generated object keyword matches an object in the ground truth, we assign $r = 1$; otherwise, if it constitutes a hallucinated object, we assign $r = 0$. (ii) VQA tasks: if the normalized generated answer is identical to the ground-truth label, then $r = 1$; otherwise, $r = 0$. For text quality, we evaluate on MME [15]. MME results are based on the averaged scores of 12 text-related sub-skills. For GPT-4o assisted evaluation, we sample 100 images from the COCO validation set and adopt the annotation protocol introduced by Huang et al. [23], Leng et al. [30], where GPT-4o evaluates two model outputs per image on three axes: Accuracy (A), Correctness (C), and Detailedness (D). Both text quality and GPT-4o assisted evaluation are conducted base on LLaVA-1.5.

H. Additional Experiments and Full results

H.1. More Evaluation Results

To further evaluate the capability of our method, we design a series of experiments targeting distinct but complementary aspects of multi-modal model performance. High-resolution visual understanding focuses on the model's ability to accurately perceive and ground objects in ultra-high-resolution images. This evaluates fine-grained visual recognition and spatial grounding in dense scenes where hallucination is prone to occur due to the small size of objects. Grounding fidelity examines whether the model can align textual references with visual regions across diverse referential expression tasks. It tests the model's precision in resolving visual-language correspondence under different linguistic contexts. Reasoning and mathematical capability assesses the model's competence in multi-step reasoning and math-related vision-language understanding, including symbolic computation, logical inference, and numerical accuracy in visual settings. Together, these evaluations provide a broad assessment of our method's effectiveness in improving visual grounding, factual consistency, and reasoning depth under challenging multimodal scenar-

Table 6. Results on high-resolution visual understanding. The best results are highlighted in **bold**.

Model	Param Size	V^* Bench			HR-Bench 4K			HR-Bench 8K		
		Attr	Spatial	Overall	FSP	FCP	Overall	FSP	FCP	Overall
GPT-4o [1]	-	-	-	66.0	70.0	48.0	59.0	62.0	49.0	55.5
Qwen-VL-max [3]	-	-	-	-	65.0	52.0	58.5	54.0	51.0	52.5
SEAL [69]	7B	74.8	76.3	75.4	-	-	-	-	-	-
DyFo [33]	7B	80.0	82.9	81.2	-	-	-	-	-	-
ZoomEye [55]	7B	93.9	85.5	90.6	84.3	55.0	69.6	88.5	50.0	69.3
InternVL-1.5 [9]	26B	-	-	-	69.5	51.8	60.6	69.3	48.5	57.9
LLaVA-OneVision [32]	7B	75.7	75.0	75.4	72.0	54.0	63.0	67.3	52.3	59.8
Qwen2.5-VL [4]	7B	73.9	67.1	71.2	85.2	52.2	68.8	78.8	51.8	65.3
Qwen2.5-VL [4]	32B	87.8	88.1	87.9	89.8	58.0	73.9	84.5	56.3	70.4
DeepEyes [79]	7B	91.3	88.2	90.1	91.3	59.0	75.1	86.8	58.5	72.6
Ours	7B	**95.0**	**91.1**	**93.4**	**92.7**	**61.1**	**76.7**	**90.1**	**62.8**	**77.2**
Δ (*vs* SOTA)	-	+1.1	+2.9	+2.8	+1.4	+2.1	+1.6	+1.6	+4.3	+4.6

Table 7. Performance on grounding fidelity. The best results are highlighted in **bold**.

Model	Param Size	refCOCO	refCOCO+	refCOCOg	ReasonSeg
Grounding DINO [43]	7B	88.2	75.9	87.0	
Qwen2.5-VL [4]	7B	89.1	82.6	86.1	68.3
DeepEyes [79]	7B	89.8	83.6	86.7	68.6
Ours	7B	**91.2**	**85.6**	**88.0**	**70.2**
Δ (*vs*SOTA)	-	+1.4	+2.0	+1.0	+1.1

ios. These experiments are conducted based on Qwen2.5-VL-7B-Instruct with COPO on A100 GPU clusters.

High-Resolution Visual Understanding. High-resolution visual understanding benchmarks, e.g. V^* Bench [69] and HR-Bench (4K/8K) [66], present a significant challenge for MLLMs as they contain ultra-high-resolution images (2K-8K) where the target objects referenced in the prompts are often minuscule, sometimes occupying fewer than 200 pixels. This scenario requires both fine-grained visual recognition and precise cross-modal grounding. However, existing MLLMs frequently fail to accurately localize or reason about such small-scale targets, resulting in hallucinated outputs. **Table 6** reports the results on these benchmarks. Our approach achieves the best results among open-source MLLMs, outperforming the state-of-the-art baselines. These results reflect our method's ability to handle fine-grained visual details and complex spatial grounding in high-resolution settings.

Grounding Fidelity. We adopt several standard benchmarks to evaluate our approach for grounding. We report results on refCOCO [6], refCOCO+ [6], refCOCOg [26], and ReasonSeg [28]. As shown in **Table 7**, our method outperforms the base model (Qwen2.5-VL-7B) and other open-source baselines across all benchmarks. These findings suggest that our method maintains stable and accurate alignment between visual input and referential language across varied grounding benchmarks.

Reasoning and Mathematical Capability. To further evaluate the ability in complex multimodal reasoning and mathematical understanding, we conduct experiments on several benchmark datasets, including MathVista [47], MathVerse [77], MathVision [65], WeMath [51], DynaMath [84], and LogicVista [71]. These benchmarks collectively test a wide range of capabilities, from symbolic manipulation and numerical reasoning to logical inference and visual comprehension of math-related images. As shown in **Table 8**, our method achieves strong results across all benchmarks. Notably, compared to baseline models such as Qwen2.5-VL, LLaVA-OneVision and DeepEyes, our model shows considerable gains in both general reasoning and mathematically grounded vision-language understanding. The results

Table 8. Results on reasoning and mathematical capability. The best results are highlighted in **bold**.

Model	Param Size	MathVista	MathVerse	MathVision	WeMath	DynaMath	LogicVista
GPT-4o [32]	-	63.8	50.2	-	-	-	-
LLaVA-OneVision [32]	7B	63.2	26.2	18.3	20.9	-	33.3
Qwen2.5-VL [4]	3B	62.3	47.6	21.2	-	-	-
Qwen2.5-VL [4]	7B	68.3	49.2	25.6	34.6	53.3	45.9
DeepEyes [79]	7B	70.1	47.3	26.6	38.9	55.0	47.7
Ours	7B	**72.8**	**50.6**	**28.9**	**43.5**	**57.5**	**51.0**
Δ (*vs* SOTA)	-	+2.7	+0.4	+2.3	+4.6	+2.5	+3.3

Table 9. Results of different decoding modes. The best results are highlighted in **bold**.

Method	LLaVA-1.5		
	$CHAIR_S$	$CHAIR_I$	POPE
DeCo	37.8	11.1	86.7
DeCo + Beam Search	33.0 ($\downarrow$ 4.8)	9.7 ($\downarrow$ 1.4)	86.7 ($\uparrow$ 0.0)
HA-DPO	38.2	11.0	82.7
HA-DPO + Beam Search	31.7 ($\downarrow$ 6.5)	8.9 ($\downarrow$ 2.1)	85.2 ($\uparrow$ 2.5)
GCPO	21.5	5.8	87.2
GCPO + Beam Search	21.1 ($\downarrow$ 0.4)	5.5 ($\downarrow$ 0.3)	87.5 ($\uparrow$ 0.3)
COPO	19.8	5.3	88.0
COPO + Beam Search	**19.7** ($\downarrow$ 0.1)	**5.3** ($\downarrow$ 0.0)	**88.1** ($\uparrow$ 0.1)

indicate that our method is capable of supporting structured multimodal reasoning and mathematical understanding under diverse task conditions.

H.2. Full Results of Comparison

H.2.1. Different Decoding Modes

To further investigate the inference capability of our method, we conduct an experiment comparing DeepEyes and our method under two decoding modes: standard decoding and beam search. The evaluation is carried out on the CHAIR and POPE based on the LLaVA-1.5. We report results across $CHAIR_S$, $CHAIR_I$, and POPE F1 score.

As shown in **Table 9**, DeCo, HA-DPO and GCPO demonstrates modest gains when switching from standard decoding to beam search, suggesting that more elaborate inference may help improve its performance. In contrast, our method achieves strong results without requiring additional decoding strategies, with only minor variation observed when beam search is applied. These results indicate that our method is less dependent on inference-time enhancements and may reflect stronger inference ability embedded in the model itself.

H.2.2. Different GRPO Variants

To further validate the generality of our causal-oriented policy optimization (COPO), we apply it on three representative GRPO variants, including DAPO [42], Dr GRPO [46], and GCPO [17]. We report results across $CHAIR_S$, $CHAIR_I$, and POPE F1 score based on the LLaVA-1.5.

As reported in **Table 10**, incorporating COPO consistently improves the hallucination-related metrics on LLaVA-1.5, yielding lower $CHAIR_S$ and $CHAIR_I$ scores and higher POPE values. For example, DAPO + COPO reduces $CHAIR_S$ from 21.3 to 19.3 and $CHAIR_I$ from 5.7 to 5.1, while GCPO + COPO achieves the best overall performance. These results demonstrate that COPO is a plug and play component for existing GRPO-based post-training schemes, which effectively constrains each decoding step to generate causally sufficient and necessary tokens, thereby mitigating hallucinations.

H.2.3. Different Implementation Paths

We investigate two alternative implementations of COPO: (i) full Sequence, where the causal completeness reward is uniformly applied to all output tokens; and (ii) reasoning step, where the reward is only added to reasoning tokens. Both settings are consistent with the core formulation of COPO and effectively integrate causal signals into policy

Table 10. Results of different GRPO variants. The best results are highlighted in **bold**.

Method	LLaVA-1.5 CHAIR$_S$	CHAIR$_I$	POPE
DAPO [42]	21.3	5.7	87.1
DAPO + COPO	19.3 (↓ 2.0)	5.1 (↓ 0.6)	88.2 (↑ 1.1)
Dr. GRPO [46]	22.1	6.1	86.9
Dr. GRPO + COPO	19.6 (↓ 2.5)	5.2 (↓ 0.9)	88.2 (↑ 1.3)
GCPO [17]	21.5	5.8	87.2
GCPO + COPO	19.2 (↓ 2.3)	5.1 (↓ 0.7)	88.4 (↑ 1.2)

Table 11. Results of different implementation paths. The best results are highlighted in **bold**.

Method	LLaVA-1.5 CHAIR$_S$	CHAIR$_I$	POPE
COPO (Full Sequence)	19.8	5.4	87.9
COPO (Reasoning Step)	19.8 (↓ 0.0)	5.3 (↓ 0.1)	88.0 (↑ 0.1)

optimization. We report results across CHAIR$_S$, CHAIR$_I$, and POPE F1 score based on the LLaVA-1.5.

As shown in **Table 11**, their results are similar, indicating that either implementation can realize our COPO. Nevertheless, applying the reward only to reasoning steps yields slightly better CHAIR and POPE scores. Therefore, we adopt this implementation in our experiments.

H.2.4. Different Masking Methods

zero-masking is used to calculate the necessity score S_{nec}. We compared zero-masking against mean ablation and Gaussian noise baselines. The high Pearson correlation coefficients (**Table 12**) suggest that the model reacts primarily to the absence of information rather than the specific masking method.

Table 12. Comparison of masking method.

Method	Pearson Correlation
Zero-Masking v.s. Mean Embedding	0.995
Zero-Masking v.s. Gaussian Noise	0.989

H.3. More Parameter Sensitivity

We conduct experiments on the hyperparameters H of Eq.6. We search H over $[2, 4, 8, 16, 32]$. **Figure 12** shows that the POPE F1 score exhibits a mild upward trend as H increases from 2 to 8, reaching a good performance at $H = 8$. Beyond this point, further increasing H yields only negligible gains while introducing additional computational overhead. This indicates that excessively large H offers limited practical benefit. Based on this trade-off between performance and efficiency, $H = 8$ is selected as the final setting in our experiments.

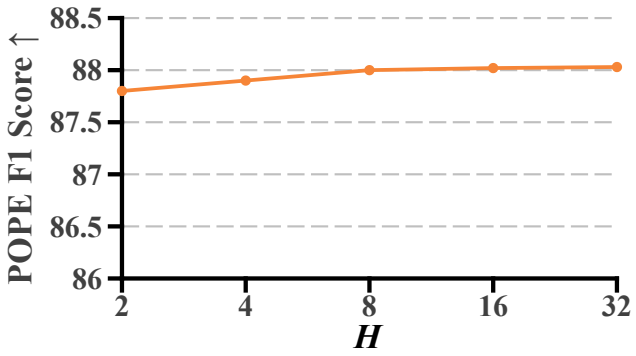

Figure 12. More parameter sensitivity.

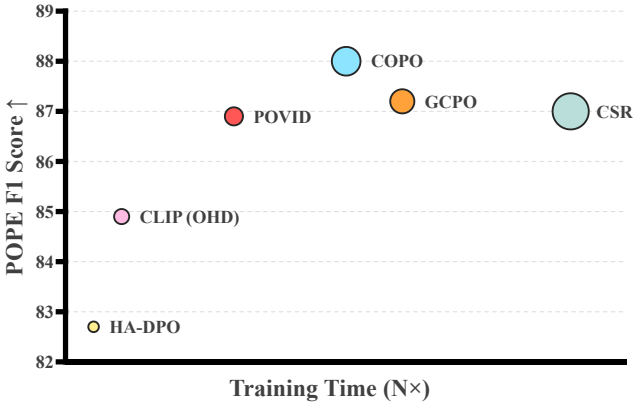

Figure 13. Trade-off performance of various methods.

Table 13. Comparison of computational efficiency.

Method	Wall-Clock Time	GPU-Hours	Peak GPU Memory
GRPO	7h 45min	62.0	74 GB
COPO	9h 12min	73.6	76 GB

H.4. Trade-off Performance

We further analyze the trade-off between hallucination suppression performance and training efficiency. Following a consistent experimental setup, we compare our method with representative baselines, including HA-DPO, CLIP (OHD), POVID, CSR, and GCPO, under the same model backbone.

As shown in **Figure 13**, the vertical axis indicates POPE F1 score, while the horizontal axis reflects the relative training cost (normalized training steps). Each method is marked with a distinct symbol, and our method is denoted by a blue star. The results show that our approach yields the highest F1 score with acceptable training time, indicating a balanced trade-off between performance and computational efficiency. More results shown in **Table 13** further validate that the time cost is acceptable compared to the performance gains.