BEYOND SAYING LESS: FINE-GRAINED ALIGNMENT FOR INFORMATIVE AND FAITHFUL VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object hallucination remains a major challenge for large vision-language models. While off-policy preference optimization proves to be an effective solution, on-policy reinforcement learning provides a more promising direction as it directly targets a model's current failure modes. However, we find that without fine-grained reward formulation and allocation, on-policy optimization often falls into an easy shortcut: reducing hallucinations merely by saying less—making fewer valid claims. To comprehensively resolve this, we propose a fine-grained alignment framework that couples dense reward signals at the data level with precise credit assignment at the algorithmic level. Specifically, we first construct the **D**ense **O**bject **P**resence and **A**bsence (DOPA) dataset to address sparse annotations that prevent valid object claims from being verified and rewarded. DOPA exhaustively annotates the deterministic presence and absence of every concept across an expanded vocabulary, significantly increasing the density of reliable reward signals during on-policy rollouts. Second, we propose **S**ubsentence-level **C**redit **A**ssignment for on-**P**olicy **O**ptimization (SCAPO) to prevent response-level shared advantages from allowing local hallucinations to compromise all other valid outputs within the same response. By assigning credit to each subsentence independently based on its object claims, SCAPO can precisely reinforce faithful generations and penalize hallucinations. Furthermore, we leverage the resulting faithful image descriptions as auxiliary context to transfer generative gains to discriminative tasks. Experiments demonstrate that our method produces highly informative, faithful descriptions in generative tasks while yielding clear performance gains on discriminative evaluation.

1 INTRODUCTION

Large vision-language models (LVLMs) have achieved substantial progress in image description, visual question answering, and multimodal reasoning. Despite these advances, they frequently generate content that is unsupported by the available visual evidence, a failure commonly referred to as hallucination Liu et al. (2024a); Bai et al. (2024); Lan et al. (2024); Rani et al. (2024); Li et al. (2023). For example, a model can claim that an object is present even though it does not appear in the image. Such errors reduce response reliability, mislead subsequent reasoning, and impede the deployment of LVLMs in settings that require dependable visual understanding. Ensuring that generated content remains consistently grounded in visual evidence has therefore become a central problem in reliable LVLM research.

Existing hallucination-mitigation approaches broadly fall into inference-time and training-based categories. Inference-time methods operate without retraining by employing contrastive decoding Leng et al. (2024); Wang et al. (2024b), representation interventions Huang et al. (2024); Gong et al. (2024); Liu et al. (2025b), or post-hoc self-correction Yin et al. (2024); Zhou et al. (2024b). In contrast, training-based methods intrinsically align LVLM generation preferences. While most existing paradigms rely on off-policy optimization using static datasets Yu et al. (2024a); Zhao et al. (2023); Xiao et al. (2025); Zhou et al. (2024a); Sarkar et al. (2025); Rafailov et al. (2023), recent studies demonstrate that learning directly from annotated on-policy rollouts Yang et al. (2025) is substantially more effective in mitigating policy-data mismatch.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

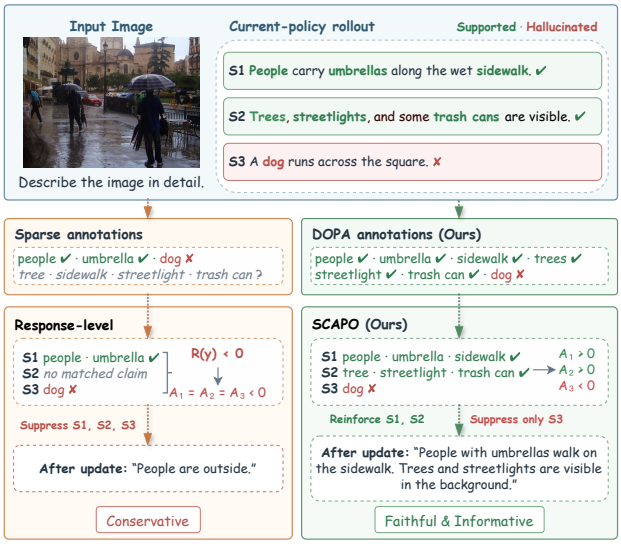

Figure 1: Motivation for our approach. Conventional optimization suffers from sparse annotations leaving valid claims unrewarded, and response-level advantages allowing local hallucinations to suppress correct outputs. By combining the dense rewards of DOPA with the subsentence-level credit assignment of SCAPO, we achieve fine-grained alignment, preventing this suppression and yielding informative and faithful descriptions.

While optimizing on-policy rollouts directly exposes a model's current errors, we find that without fine-grained reward formulation and allocation, it often falls into an easy shortcut: the model reduces hallucinations not by improving faithfulness, but by reducing informativeness—that is, essentially saying less by making fewer valid claims. As illustrated in Figure 1, we identify two coarse-grained bottlenecks driving this behavior. The first is a data-level bottleneck: sparse annotations leave valid object claims in current rollouts unverified and unrewarded. The second is an algorithmic-level bottleneck: response-level optimization assigns shared advantages across the entire text, allowing local hallucinations to compromise all other valid outputs within the same response. Consequently, the model is incentivized toward global conservatism rather than precise error correction, underscoring the critical need for denser reward coverage and localized credit assignment.

To comprehensively resolve these bottlenecks, we propose a fine-grained alignment framework coupling dense reward signals with precise credit assignment. At the data level, we construct the Dense Object Presence and Absence (DOPA) dataset, which provides dense, reliable training rewards by exhaustively annotating the deterministic presence or absence of concepts over an expanded vocabulary. At the algorithmic level, we introduce Subsentence-level Credit Assignment for on-Policy Optimization (SCAPO). Rather than relying on coarse response-level shared advantages, SCAPO segments responses at subsentence boundaries and independently assigns credit. This localized mechanism ensures that reinforcement and suppression are precisely targeted, preventing isolated hallucinations from penalizing other valid outputs within the same response.

Our contributions are summarized as follows:

- We construct the **D**ense **O**bject **P**resence and **A**bsence (DOPA) dataset to overcome data-level reward sparsity. By exhaustively annotating the deterministic presence and absence of concepts across an expanded vocabulary, DOPA provides dense and reliable reward signals for object claims during on-policy rollouts.
- We propose **S**ubsentence-level **C**redit **A**ssignment for on-**P**olicy **O**ptimization (SCAPO) to resolve the algorithmic bottleneck of response-level shared advantages. By assigning credit independently to each subsentence based on its specific claims, SCAPO prevents local hallucinations from suppressing all other valid outputs.

- Extensive experiments show that our fine-grained alignment framework overcomes the "saying less" shortcut, producing descriptions that are both faithful and informative. As an additional transfer evaluation, we further show that these improved descriptions provide useful auxiliary context for discriminative tasks.

2 Related Work

2.1 RLHF and Preference Optimization

Reinforcement learning from human feedback (RLHF) fundamentally aligns large models using on-policy algorithms such as PPO Schulman et al. (2017), though maintaining separate reward and value models is computationally expensive. Direct Preference Optimization (DPO) Rafailov et al. (2023) provides a simpler offline alternative, while policy-data mismatch has motivated iterative preference-refreshing approaches Pang et al. (2024). More recently, methods such as GRPO Shao et al. (2024) have revitalized on-policy optimization by estimating advantages from relative rewards without requiring a critic model, with subsequent variants addressing issues such as length-biased credit weighting Yu et al. (2025a); Liu et al. (2025d). However, GRPO-style methods typically assign rewards at the response level. This granularity is particularly limiting for hallucination mitigation, where hallucinated content often constitutes only a small portion of an otherwise faithful response. This motivates finer-grained credit assignment that can selectively penalize hallucinations while reinforcing correct content.

2.2 Inference-Time Hallucination Mitigation

Hallucination mitigation in LVLMs has attracted extensive attention Liu et al. (2024a); Bai et al. (2024); Lan et al. (2024); Rani et al. (2024). A prominent line of research focuses on inference-time methods that improve visual faithfulness without updating model parameters. These include contrastive decoding to suppress language priors Leng et al. (2024); Wang et al. (2024b), generation interventions through attention or logit manipulation Huang et al. (2024); Gong et al. (2024); Liu et al. (2025b), post-hoc self-correction Yin et al. (2024); Zhou et al. (2024b), external visual-evidence augmentation using auxiliary vision models Zhao et al. (2025); Li et al. (2025), and staged prompting that verbalizes visual evidence before final prediction Xuan et al. (2025). However, while inference-time approaches can improve visual faithfulness without retraining, they inevitably introduce inference latency and fail to internalize visual faithfulness directly into the model parameters.

2.3 Training-Based Hallucination Mitigation

Training-based methods intrinsically improve visual grounding via preference optimization. Early offline approaches construct preference pairs using expert corrections Yu et al. (2024a); Zhao et al. (2023); Xiao et al. (2025) or synthetic hallucination injection Zhou et al. (2024a); Sarkar et al. (2025). However, recent studies Yu et al. (2025b); Yang et al. (2025) demonstrate that on-policy training is more effective, as it directly targets the evolving policy's failure modes. Yet, existing online methods face critical bottlenecks: accurate credit assignment demands either complex rollout modification Yang et al. (2025) or computationally expensive iterative LLM judge evaluation Yu et al. (2025b). To overcome these challenges, we introduce the DOPA dataset as a reusable, high-density reward oracle providing exhaustive object-existence annotations on an expanded vocabulary. Building on this, we propose SCAPO to leverage these dense signals, shifting from coarse response-level evaluation to precise subsentence-level credit assignment for on-policy rollouts.

3 Construction of DOPA Dataset

To overcome the data-level bottleneck where sparse annotations leave valid object claims unrewarded, we construct the Dense Object Presence and Absence (DOPA) dataset to provide dense and reliable reward signals for on-policy rollouts. We build DOPA on 5,000 images sampled from the MS COCO Lin et al. (2014) training split, disjoint from all evaluation images. As illustrated in Figure 2, its construction comprises vocabulary construction, coarse annotation, edge-case refinement, and manual verification, which are detailed in the following subsections.

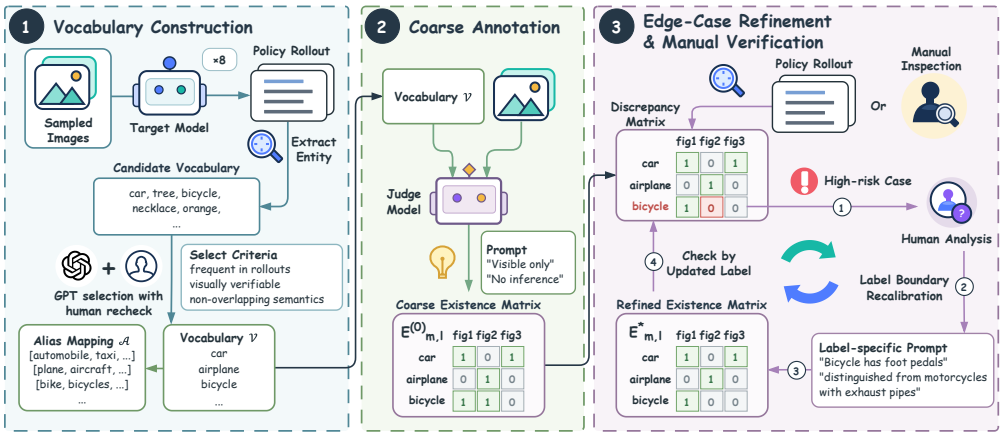

Figure 2: Pipeline for constructing the DOPA dataset. The process begins with rollout-driven vocabulary construction to establish the canonical vocabulary $\mathcal{V}$ and alias mapping $\mathcal{A}$. Based on these, large-scale coarse annotation and edge-case refinement are performed. Ultimately, a rigorous manual verification cycle yields the high-fidelity object-existence matrix E^*.

3.1 Vocabulary Construction

To increase reward coverage for concepts naturally mentioned by the target model, we derive an expanded vocabulary directly from its rollouts on the training data. After extracting and normalizing entity mentions, we curate these candidates into canonical object labels based on three strict criteria: **(i)** sufficient frequency in target-model rollouts; **(ii)** direct visual verifiability without subjective inference; and **(iii)** semantic clarity with minimal overlap. We initially use GPT-5.5 OpenAI (2026) to cluster the candidates into a draft set, and then manually refine this draft against the aforementioned criteria. Finally, we construct a conflict-free alias mapping covering synonyms and morphological variations for each retained label. This process yields the final vocabulary $\mathcal{V}$ comprising **160** canonical labels and a mapping $\mathcal{A}$ of **894** aliases.

3.2 Coarse Annotation

Existing datasets typically provide only positive annotations, leaving it ambiguous whether an unannotated concept is absent or simply omitted. DOPA instead defines a closed-set object-label space over $\mathcal{V}$, in which every image-label pair is explicitly annotated as either present or absent. This exhaustive presence-absence annotation removes the unknown state for all in-vocabulary claims and provides unambiguous reward signals during on-policy training. We employ Qwen3.6-27B Qwen Team (2026) as the initial coarse judge to implement this exhaustive annotation. Prompted with strict criteria for handling occlusion and explicitly prohibiting commonsense inference, the judge determines object presence based solely on visual evidence. This produces a complete initial object-existence matrix $E^{(0)}_{m,l} \in \{0, 1\}$, indicating whether the object corresponding to the m-th canonical label is present in the l-th image.

3.3 Edge-Case Refinement & Manual Verification

To improve annotation fidelity, we implement an iterative refinement and manual verification protocol. First, we evaluate model-generated object claims against the coarse matrix to isolate high-risk edge cases, such as heavy occlusion or unclear semantic boundaries. We resolve these systematic ambiguities by crafting tailored, label-specific prompts with rigorous inclusion/exclusion criteria to re-query the judge model. Subsequently, for each canonical label, we randomly sample 30 image-label pairs for manual evaluation. After each round of prompt refinement and re-annotation, we draw a new random set of 30 pairs and repeat the evaluation until no errors are observed in the newly sampled subset. Once every label passes this sample-based quality-control procedure, we finalize the re-annotated object-existence matrix $E^*_{m,l} \in 0, 1$.

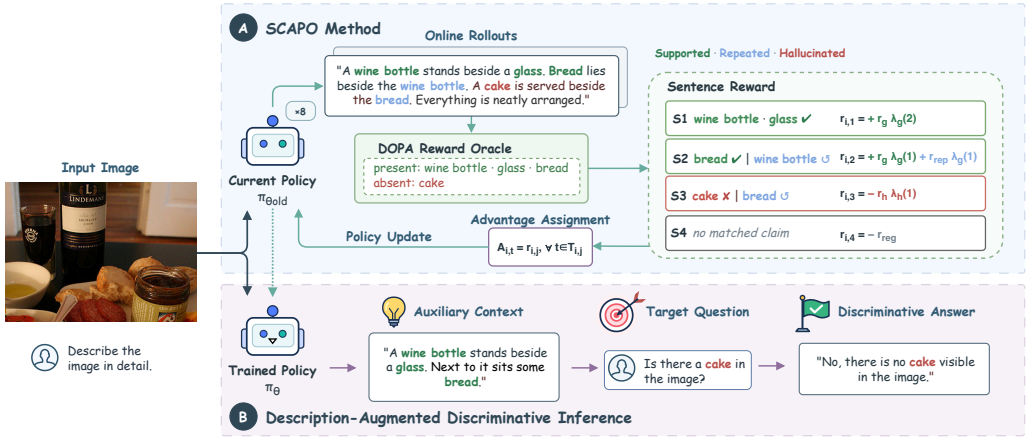

Figure 3: Overview of the SCAPO method and the adopted inference paradigm. Current-policy rollouts are segmented at subsentence boundaries and evaluated against the DOPA reward oracle. The resulting subsentence-level rewards are independently assigned to their corresponding token spans for the on-policy update. Furthermore, the optimized model leverages its generated faithful descriptions as auxiliary context for description-augmented discriminative inference.

Ultimately, this exhaustive process provides dense and reliable reward signals for the subsentence-level credit assignment, enabling DOPA to serve as a high-quality, reusable reward oracle during training that requires only deterministic lookups rather than repeated judge-model calls.

4 SCAPO METHOD

To prevent coarse response-level shared advantages from allowing local hallucinations to suppress valid outputs, we build on the DOPA dataset and propose Subsentence-level Credit Assignment for on-Policy Optimization (SCAPO). As illustrated in Figure 3, SCAPO segments each response at subsentence boundaries and evaluates the object claims within each subsentence against the DOPA annotations, effectively encouraging highly informative and faithful generations rather than leading to global conservatism. We provide formal definitions and the pseudocode of our SCAPO method in Appendix B.

4.1 SUBSENTENCE-LEVEL REWARD

To evaluate the object mentions within the j-th subsentence of the i-th response, we categorize them according to the DOPA annotations and define three explicit counts: $n_{i,j}^{\text{new}}$ represents the number of newly covered supported objects (valid claims appearing for the first time in the response), $n_{i,j}^{\text{rep}}$ denotes the repeated supported objects (valid claims already mentioned in preceding subsentences), and $n_{i,j}^{\text{hal}}$ counts the hallucinated objects. The subsentence-level reward is then defined as:

$$r_{i,j} = \begin{cases} -r_h \, \lambda_h\big(n_{i,j}^{\text{hal}}\big), & n_{i,j}^{\text{hal}} > 0, \\ -r_{\text{reg}}, & \text{no objects}, \\ +r_g \, \lambda_g\big(n_{i,j}^{\text{new}}\big) + r_{\text{rep}} \, \lambda_g\big(n_{i,j}^{\text{rep}}\big), & \text{otherwise}. \end{cases} \tag{1}$$

Here, $r_h, r_g, r_{\text{rep}}, r_{\text{reg}} \geq 0$ are base coefficients, scaled by monotonically increasing functions λ_h and λ_g. This formulation explicitly enforces three design principles: **(1) Zero-tolerance for hallucinations:** any hallucination triggers a penalty, strictly preventing correct information from offsetting explicit errors. **(2) Grounded informativeness:** hallucination-free subsentences are rewarded for novel claims and, marginally, for repetitions. **(3) Anti-conservatism:** subsentences lacking verifiable objects incur a regularization penalty ($-r_{\text{reg}}$) to discourage the model from generating generic, uninformative content.

4.2 SUBSENTENCE-LEVEL ON-POLICY RL TRAINING

During each optimization step, we sample image-prompt pairs from the current policy snapshot $\pi_{\theta_{\text{old}}}$. For each response y_i, we compute subsentence-level rewards $r_{i,j}$ via Eq. (1). To bypass coarse response-level aggregation, we directly broadcast these local rewards as token advantages: $A_{i,t} = r_{i,j}$ for all tokens $t \in T_{i,j}$, where $T_{i,j}$ is the token span of subsentence $s_{i,j}$. Using these token advantages, we optimize the standard clipped surrogate objective:

$$\mathcal{L}_{\text{SCAPO}} = -\frac{1}{N_{\text{tok}}} \sum_{i,j} \sum_{t \in T_{i,j}} \min\Big(\rho_{i,t} A_{i,t}, \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon) A_{i,t}\Big) + \beta_{\text{KL}} D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}) \,, \quad (2)$$

where N_{tok} is the total number of response tokens in the batch, $\rho_{i,t}$ is the standard policy importance ratio, ϵ is the clipping coefficient, and β_{KL} controls the KL regularization toward the fixed reference policy π_{ref}. Thus, while policy optimization remains token-wise, credit is localized at the subsentence level, with KL regularization preventing excessive policy deviation.

Unlike conventional PPO-based RLHF, SCAPO does not require training or maintaining an auxiliary value model. PPO-based RLHF typically estimates token-level advantages through a learned critic and GAE, whereas SCAPO assigns each subsentence an independently verified factual reward that is shared only among its own tokens. Advantage signals are not propagated across subsentence boundaries, preventing a later hallucination from altering the credit assigned to an earlier faithful statement. This design replaces general-purpose return estimation with explicit semantic locality, which is well suited to sparse and localized hallucination errors.

4.3 DESCRIPTION-AUGMENTED DISCRIMINATIVE INFERENCE

Following SCAPO optimization, the policy exhibits a marked improvement in generating faithful and informative visual descriptions. To transfer these generative grounding capabilities to downstream discriminative tasks, we adopt the description-augmented inference paradigm. Specifically, given an image and a discriminative query, the optimized model first generates a grounded description. The model then conditions its final discriminative answer on both the original image and this generated description jointly.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate our approach across both generative and discriminative tasks. For generative tasks, we assess hallucination rate and informativeness using MS COCO Lin et al. (2014) and AMBER Wang et al. (2023). For discriminative tasks, we utilize PhD Liu et al. (2025a) and the discriminative tasks of AMBER.

Comparison Methods. All methods use Qwen2.5-VL-7B Bai et al. (2025b) as the backbone. We organize the evaluated methods into three groups. The inference-time group includes the unmodified backbone, VCD Leng et al. (2024), OPERA Huang et al. (2024), and VTI Liu et al. (2025b). The off-policy group includes standard DPO Rafailov et al. (2023), multimodal DPO (mDPO) Wang et al. (2024a), SymMPO Liu et al. (2025c), and OPA-DPO Yang et al. (2025). The preference data for these methods are constructed from Qwen2.5-VL-7B rollouts and annotated by Qwen3.6-27B Qwen Team (2026). The on-policy group includes GRPO Shao et al. (2024), DAPO Yu et al. (2025a), Dr. GRPO Liu et al. (2025d), and our SCAPO. All the on-policy methods are trained using our DOPA annotations. Complete training configurations are reported in Appendix C.1.

Evaluation Metrics. For discriminative evaluation, we report standard metrics including accuracy, precision, recall, and F1, as well as the benchmark-specific PhD-index. For generative evaluation, to enable the metric to simultaneously reflect the faithfulness and informativeness of the model output, we introduce a new metric **Caption Score (Cap. Score)** built on the harmonic mean of **1-Hal. Rate** and **Cover Rate**, where Hal. Rate corresponds to CHAIR_i. The formal definitions are given in Appendix C.2.

Table 1: Evaluation on generative and discriminative hallucination benchmarks. **Cap. Score = HarMean**(1-Hal. Rate, Cover Rate) jointly measures caption faithfulness and informativeness. (Cap. Score is computed with the original object-existence annotation of the dataset. Cap. Score† is computed with our expanded DOPA annotation.)

	AMBER		MS COCO		PhD				
Method	Cap. Score ↑	Dis. Acc. ↑	Cap. Score ↑	Cap. Score† ↑	PhD-base ↑	PhD-sec ↑	PhD-icc ↑	PhD-ccs ↑	PhD-all ↑
(i) Inference-time Methods									
Qwen2.5-VL	76.5	83.9	73.2	64.2	76.8	68.2	62.1	59.5	66.6
VCD	76.5	83.8	72.7	65.9	77.0	71.1	63.6	65.2	69.2
OPERA	76.3	84.5	73.7	65.3	76.4	70.9	64.6	62.0	68.5
VTI	47.7	85.9	50.2	33.6	79.0	57.3	58.3	66.8	65.4
(ii) Off-policy Methods									
DPO	77.7	71.8	74.8	70.0	70.8	52.9	52.4	67.9	61.0
mDPO	76.3	82.9	74.2	66.5	77.0	70.0	66.7	67.4	70.3
SymMPO	75.5	85.4	74.2	65.4	80.2	62.8	55.6	**73.3**	68.0
OPA-DPO	67.8	85.4	72.2	53.9	76.1	70.9	52.7	69.5	67.3
(iii) On-policy Methods									
GRPO	69.9	83.8	66.5	51.4	77.2	63.9	63.5	61.8	66.6
DAPO	61.1	83.1	51.8	33.3	77.1	57.9	59.8	63.3	64.5
Dr. GRPO	68.3	84.1	65.1	47.4	76.8	56.2	62.5	62.7	64.5
SCAPO (Ours)	**82.8**	**87.4**	**79.6**	**78.7**	**80.3**	**74.5**	**67.5**	68.2	**72.6**

Table 2: Effect of the DOPA dataset used for on-policy training. Both variants below utilize the same SCAPO objective but derive rewards from different annotations.

	MS COCO Label		DOPA Label		AMBER		MMHal-Bench	
Method	Hal. Rate ↓	Cover Rate ↑	Hal. Rate ↓	Cover Rate ↑	Hal. Rate ↓	Cover Rate ↑	Score ↑	Hal. Rate ↓
Qwen2.5-VL (Baseline)	15.1	64.3	11.8	50.5	4.9	64.0	3.26	39.6
SCAPO Training								
w/ MS COCO annotations	13.9	**78.8**	11.6	52.0	**4.6**	66.1	3.45	34.4
w/ DOPA annotations (**Ours**)	**13.4**	73.7	**9.4**	**69.6**	4.8	**73.2**	**3.52**	**33.3**

5.2 Main Results

Table 1 presents the main results across three hallucination benchmarks. When simultaneously considering faithfulness and informativeness, most hallucination mitigation methods struggle to yield visible improvements in the Cap. Score metric. Aside from our SCAPO, only DPO achieves consistent gains across the three generative settings. Notably, despite being trained on the same DOPA data as SCAPO, other GRPO-style on-policy methods yield Cap. Scores below the Qwen2.5-VL baseline. This degradation stems from an optimization shortcut toward global conservatism, which we analyze in detail in Section 5.4. In contrast, SCAPO effectively overcomes this bottleneck, achieving strong Cap. Scores of 82.8% on AMBER and 79.6% on MS COCO.

Importantly, these generative improvements also transfer to discriminative tasks. SCAPO achieves an 87.4% discriminative accuracy on AMBER and the best results among the compared methods on PhD-base, PhD-sec, PhD-icc, and PhD-all. Overall, this consistent improvement across both generative and discriminative settings distinguishes SCAPO from alternatives that inflate an isolated metric at the expense of cross-benchmark performance.

5.3 Experiments on DOPA

Generalization via DOPA Annotations. Table 2 isolates the effect of the annotation source while keeping the SCAPO objective unchanged. We additionally incorporate the LLM-judge-based MMHal-Bench Sun et al. (2024) for rigorous cross-benchmark assessment. It can be observed that while training with sparse MS COCO annotations leads to substantial coverage improvements on its specific label set, the gains on other benchmarks are notably smaller, as claims outside this space remain unverifiable and unrewarded. In contrast, replacing these sparse labels with the dense DOPA dataset drives generalized improvements. It significantly boosts in-domain performance under the DOPA evaluation, reducing the Hal. Rate from 11.8% to 9.4% and raising the Cover Rate from 50.5% to 69.6%. Furthermore, it exhibits superior cross-dataset generalization, maintaining an overall advantage over the sparsely annotated counterpart on AMBER and MMHal-Bench.

Table 3: Annotation density on training-set rollouts. For each annotation setting, we sample eight responses per training example and report the per-response averages. "Sup." and "Hal." denote ground-truth-supported and hallucinated content, respectively.

	Annotated Objects per Response			Subsentences per Response		
Annotation	Avg.	Sup.	Hal.	Avg.	Sup.	Hal.
MS COCO	2.73	2.32	0.41	6.33	2.62	0.45
DOPA (**Ours**)	5.13	4.37	0.76	6.37	3.46	0.81

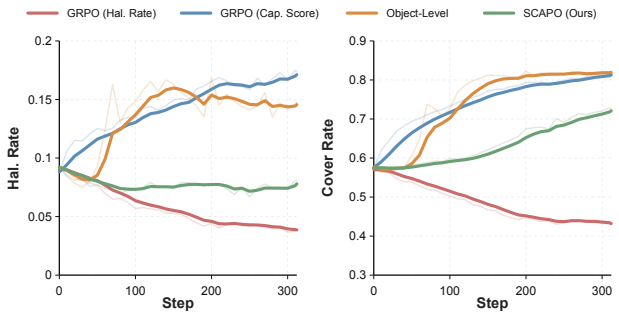

Figure 4: Training dynamics under different credit-assignment strategies. Response-level objectives face a trade-off between limited coverage (when optimizing Hal. Rate) and increased hallucinations (when optimizing Cap. Score). Object-level assignment fragments semantic coherence, causing unstable increases in both metrics. In contrast, SCAPO's subsentence-level credit assignment resolves these issues, steadily expanding valid coverage while keeping the hallucination rate controlled.

Reward Density Analysis. The gains from DOPA can be largely explained by its higher reward density, as demonstrated in Table 3. Compared to MS COCO annotations, DOPA nearly doubles the average number of annotated objects (5.13 vs. 2.73) and significantly increases the total number of reward-bearing subsentence labels. This richer provision of verifiable feedback effectively grounds the on-policy optimization, preventing valid object claims from going unrewarded.

5.4 Analysis of SCAPO

Training Dynamics. Figure 4 compares SCAPO against alternative credit assignment strategies to illustrate their distinct failure modes. Optimizing solely for a response-level Hal. Rate penalty triggers the conservative "saying less" shortcut, causing Cover Rate to plummet despite dropping Hal. Rate to 4%. Conversely, optimizing for a response-level Cap. Score fails to effectively suppress hallucinations, raising Cover Rate but driving Hal. Rate up to 17%. Furthermore, an object-level variant fragments semantic coherence, yielding a highly unstable curve where both metrics increase. SCAPO overcomes these pitfalls through subsentence-level credit assignment. By evaluating claims within their natural semantic boundaries, SCAPO steadily improves Cover Rate while keeping Hal. Rate firmly controlled, successfully avoiding both global conservatism and excessive hallucinations.

Output Factual Richness. Because Cover Rate is bounded by a predefined vocabulary, we further evaluate open-vocabulary performance using the FaithScore benchmark Jing et al. (2024). This benchmark employs an LLM judge to decompose responses into atomic facts and reports two key metrics: the FaithScore metric (which measures overall faithfulness, where higher indicates fewer hallucinations) and Avg. Facts (which measures informativeness by counting the average number of atomic facts per response). As illustrated in Figure 5, response-level GRPO-style methods achieve high FaithScores by exploiting a conservative shortcut: they severely curtail their outputs across all four evaluation settings. For instance, under the detailed setting, the Avg. Facts for GRPO and DAPO drop to 31.65 and 24.31, respectively, compared to 41.63 for the Qwen2.5-VL backbone. In contrast, SCAPO secures a superior FaithScore while simultaneously preserving the Avg. Facts of

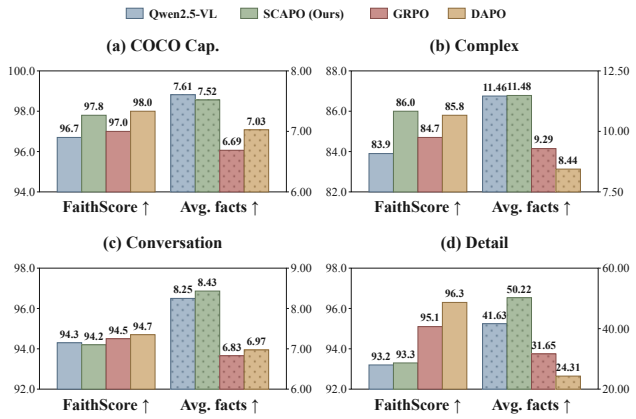

Figure 5: FaithScore and Avg. facts across four generation settings. GRPO-style methods all improve faithfulness at the expense of the number of atomic facts in the output, while SCAPO preserves factual richness and maintains high faithfulness.

Table 4: Progressive ablation on the AMBER discriminative task. We first augment the Qwen2.5-VL baseline with a self-generated image description, and then upgrade the underlying backbone to the SCAPO-trained policy to evaluate the impact of generation quality.

Configuration	Acc. ↑	Pre. ↑	Rec. ↑	F1 ↑
Qwen2.5-VL	83.9	84.5	92.6	88.4
+ Desc. Aug.	86.9 (+3.0)	87.3 (+2.8)	93.8 (+1.2)	90.4 (+2.0)
+ SCAPO	**87.4** (+3.5)	**87.6** (+3.1)	**94.4** (+1.8)	**90.9** (+2.5)

the backbone across all settings. This provides compelling evidence that SCAPO's improvements stem from precise hallucination suppression rather than global output impoverishment.

5.5 Description-Augmented Discriminative Inference

Table 4 presents a progressive ablation of the description-augmented discriminative inference. First, adding a generated description to the direct-answering baseline improves the Qwen2.5-VL backbone's accuracy from 83.9% to 86.9% and its F1 score from 88.4% to 90.4%, confirming the general utility of an explicit auxiliary visual representation. Next, substituting the generative backbone with our SCAPO-aligned model further elevates accuracy to 87.4% and F1 to 90.9%. These progressive gains indicate that the effectiveness of description augmentation depends on the quality of the generated visual context. By supplying denser valid evidence and strictly fewer misleading priors, SCAPO transfers its generative grounding improvements to downstream discrimination, requiring zero task-specific fine-tuning.

6 Conclusion

In this paper, we investigate the conservative "saying less" shortcut that plagues on-policy hallucination mitigation, attributing this degradation to sparse reward supervision and coarse response-level credit assignment. To comprehensively resolve these bottlenecks, we introduce a dual-level fine-grained alignment framework. At the data level, we construct the DOPA dataset as a dense and reliable reward oracle; at the algorithmic level, we propose SCAPO for precise subsentence-level credit assignment. Extensive experiments across multiple hallucination benchmarks demonstrate that our approach effectively resolves the faithful-informative trade-off, substantially reducing unsupported content while expanding the coverage of valid object claims. Furthermore, we show that these faithful and informative descriptions serve as highly effective auxiliary context, successfully transferring our generative gains to downstream discriminative tasks.

AI USE STATEMENT

In this work, we used generative AI tools for dataset construction, which we disclose as part of the required AI-use disclosure. Specifically, GPT-5.5 (OpenAI, 2026) was used to draft a canonical object vocabulary, and Qwen3.6-27B (Qwen Team, 2026) was used to generate the initial and refined object-existence annotations for DOPA. We did not use generative AI to generate research hypotheses, design the proposed method, or derive scientific conclusions. In addition, we used generative AI tools for recommended-disclosure tasks, including language polishing of author-written text and assistance with routine coding. All AI-assisted text was reviewed for consistency with the authors' intended meaning, and AI-assisted code was inspected and tested by the authors. We take responsibility for the final content of this work, including all text, claims, code, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We take several steps to facilitate reproducibility. The appendix provides the complete formalization of our reward construction and optimization procedure, including the claim-set definitions and pseudocode for SCAPO (Appendix B). We report the full experimental configuration, including training hyperparameters and environment, in Appendix C.1. All prompts used for data annotation, training, and inference are provided verbatim in Appendix D.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

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. *Advances in Neural Information Processing Systems*, 38, 2025.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. DAMRO: Dive into the attention mechanism of LVLM to reduce object hallucination. In *Conference on Empirical Methods in Natural Language Processing*, pp. 7696–7712. Association for Computational Linguistics, 2024.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP*, pp. 5042–5063, 2024.

Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. A survey of hallucination in large visual language models. *arXiv preprint arXiv:2410.15359*, 2024.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080. Association for Computational Linguistics, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023.

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*, pp. 740–755. Springer, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pp. 34892–34916, 2023.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. PhD: A ChatGPT-prompted visual hallucination evaluation dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19857–19866, 2025a.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, pp. 72402–72419, 2025b.

Wenqi Liu, Xuemeng Song, Jiaxi Li, Yinwei Wei, Na Zheng, Jianhua Yin, and Liqiang Nie. Mitigating hallucination through theory-consistent symmetric multimodal preference optimization. In *Advances in Neural Information Processing Systems*, volume 38, pp. 111259–111284, 2025c.

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*, pp. 216–233, 2024b.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025d.

OpenAI. GPT-5.5 model. https://developers.openai.com/api/docs/models/gpt-5.5, 2026. Accessed: 2026-07-29.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116617–116637, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model. https://qwen.ai/blog?id=qwen3.6-27b, April 2026.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Anku Rani, Vipula Rawte, Harshad Sharma, Neeraj Anand, Krishnav Rajbangshi, Amit Sheth, and Amitava Das. Visual hallucination: Definition, quantification, and prescriptive remediations. *arXiv preprint arXiv:2403.17306*, 2024.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan O. Arik, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, 2025.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL*, pp. 13088–13110, 2024.

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 *Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, 2024a.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL*, pp. 15840–15853. Association for Computational Linguistics, 2024b.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 25543–25551, 2025.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, pp. 1408–1450, 2025.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025.

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.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, pp. 113222–113244, 2025a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024a.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19985–19995, 2025b.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-Vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning*, 2024b.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *International Conference on Machine Learning*, pp. 77461–77486, 2025.

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.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024a.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, pp. 56969–56998, 2024b.