MARVEL: MULTI-AGENT REINFORCED VISUAL EXAMINATION FOR HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination remains a critical challenge for vision-language models (VLMs), especially in real-world scenarios where models generate image-inconsistent content. Existing mitigation methods typically rely on human annotations or external supervision, limiting their scalability to large-scale unlabeled image corpora. We propose **MARVEL**, a multi-agent framework that mitigates hallucination in VLMs via reinforced visual examination, without any ground-truth labels. MARVEL casts factual verification as a multi-agent examination process: model responses are decomposed into atomic claims, reformulated as visual questions, and examined against the image to produce consistency signals that serve as reinforcement learning rewards. To improve reward reliability, we further introduce a Visual Claim Regulation (VCR) mechanism that encourages informative, image-grounded descriptions while suppressing trivial or redundant claims. We evaluate MARVEL on 12 benchmarks spanning hallucination evaluation (+3.30), real-world VQA (+1.74), and general multimodal understanding benchmarks (+4.07), where it consistently outperforms existing methods without human-annotated factual supervision. Moreover, MARVEL generalizes effectively across model scales, demonstrating its potential as a scalable hallucination mitigation solution on large-scale real-world unlabeled image corpora.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable capabilities across a wide range of multimodal tasks, from image captioning to visual question answering (Achiam et al., 2023; Touvron et al., 2023; Liu et al., 2023b; Chen et al., 2024b; Bai et al., 2025a; Comanici et al., 2025). However, hallucination remains a fundamental bottleneck: models frequently generate descriptions that are fluent and coherent yet factually inconsistent with the visual content of the image (Liu et al., 2024a; Rawte et al., 2025; Li et al., 2025; Yang et al., 2025a). Such visual hallucinations manifest as fabricated objects, incorrect attributes, or erroneous spatial relationships, and represent a critical barrier to the reliable deployment of VLMs in real-world applications (Dahl et al., 2024; Kim et al., 2025). Reinforcement Learning from Verifiable Rewards (RLVR) has emerged as a promising paradigm for improving model factuality by optimizing directly against reward signals derived from output correctness (Shao et al., 2024; Yu et al., 2026; Wen et al., 2025; Yue et al., 2026). However, unlike domains such as mathematics or code generation where correctness can be programmatically verified, factual correctness in VLM generation is inherently difficult to verify from images. Existing approaches to mitigating VLM hallucination largely fall into two categories. The first relies on human annotations or external supervision, such as curated factual datasets, preference annotations, or

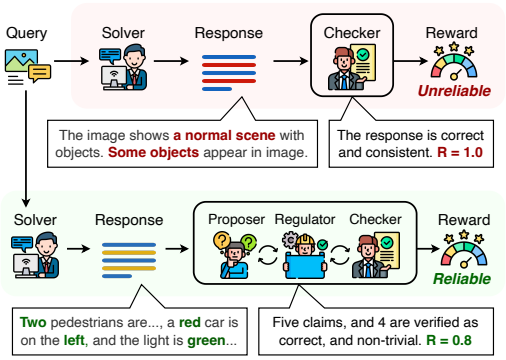


Figure 1: MARVEL achieves reliable visual self-verification rewards through a multi-agent self-check mechanism, encouraging VLMs to generate more informative, image-supported, and diverse responses without annotations.

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.

learned reward models (Ouyang et al., 2022; Bai et al., 2022). While effective, these methods are expensive to scale and difficult to generalize to large-scale unlabeled image corpora. The second category leverages self-checking or self-reflection mechanisms to improve factual consistency through iterative verification or critique (Madaan et al., 2023; Shinn et al., 2023; Li et al., 2026). However, these mechanisms are designed for text, where claims are checked against textual evidence. In VLMs, the only evidence is the image as perceived by the verifier itself, so a consistency-only reward becomes exploitable: image-agnostic claims are verified from language priors alone and near-duplicate claims inflate the reward, leading to reward hacking and entropy collapse (Figure 2). A key open question remains: ***How can VLMs obtain reliable and scalable image-grounded verification signals without relying on annotations or external supervision?***

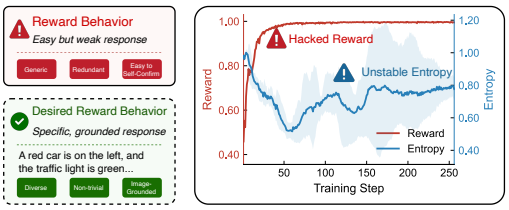


Figure 2: Motivation. Naive self-verification rewards vague but safe descriptions, causing rapid reward growth that reflects reward hacking rather than genuine visual grounding. Reliable visual self-verification should instead favor diverse, non-trivial, and image-grounded claims while maintaining stable RL training.

To address this, we propose **MARVEL**, a visual multi-agent self-check framework for scalable hallucination mitigation in VLMs. Our key insight is that, in the visual domain, rewarding *whether* claims are correct is insufficient; one must also regulate *what* is claimed. MARVEL verifies each claim against the image through a Checker that never observes the response (Madaan et al., 2023; Li et al., 2026), an information asymmetry that reduces confirmation bias (Stechly et al., 2025; Jhaveri et al., 2026). To make these rewards reliable, we introduce a Visual Claim Regulation (VCR) mechanism, implemented as a **Regulator** agent that audits the claim set before it is rewarded: it removes redundant claims, rewards non-redundant coverage, and penalizes claims that remain answerable from a degraded image. Experimental results across 12 benchmarks show that MARVEL consistently outperforms existing methods without relying on ground-truth. Our main contributions are:

- We identify why text-based self-verification fails on VLMs: with the image as the only evidence, image-agnostic and redundant claims make the consistency reward exploitable.
- We propose **MARVEL**, the first visual multi-agent reinforced self-check framework for VLM hallucination mitigation, and **VCR**, a claim-set regulation mechanism that stabilizes self-verification rewards by removing redundant claims and penalizing image-agnostic ones.
- Experiments and ablations on 12 benchmarks show that MARVEL consistently outperforms baselines, offering a scalable pathway to self-supervised VLM hallucination mitigation.

2 Related Work

VLM Hallucination. VLM hallucination, i.e., generated text that contradicts the visual input, is typically measured by object- and attribute-level benchmarks (Li et al., 2023; Wang et al., 2023). Supervised alignment methods adapt RLHF or DPO with human preference data (Sun et al., 2024; Yu et al., 2024; Fu et al., 2025; Yang et al., 2025c; Wu et al., 2025), but costly labels limit their scalability to unlabeled corpora. Inference-time methods, including contrastive decoding (Leng et al., 2024; Wang et al., 2024), attention intervention (Huang et al., 2024; An et al., 2025; Zou et al., 2024; Cho et al., 2025), correction decoding (Wang et al., 2025; Yin et al., 2025; Tang et al., 2025; Anand et al., 2026; Chen et al., 2024a), and post-hoc self-checking (Yin et al., 2024; Wu et al., 2024), leave model weights unchanged and are thus complementary to MARVEL. Representation editing (Jiang et al., 2025; Wang et al., 2026b) steers internal activations toward factual directions, which require explicit supervision to define. Closer to our setting, annotation-free self-training builds preference pairs from corrupted-image negatives (Deng et al., 2024) or self-injected hallucinations (Lu et al., 2025), and visual evidence prompting (Li et al., 2025) grounds generation in salient regions. Vision-Zero (Wang et al., 2026a) and EvoLMM (Thawakar et al., 2025) self-evolve VLM reasoning via self-play, and MARCH (Li et al., 2026) applies multi-agent self-checking to text-only LLMs; none addresses how self-verification rewards are exploited when the image is the only evidence. MARVEL instead derives fine-grained rewards from the model's own outputs without annotation or external verifiers, and VCR guards these rewards against exploitation by redundant and image-agnostic claims.

3 METHOD

3.1 PROBLEM FORMULATION

We consider reinforcement learning for vision-language generation without ground-truth labels. Let $\mathcal{D} = \{(\boldsymbol{I}, \boldsymbol{q})\}$ denote a collection of image-question pairs, where $\boldsymbol{I} = \{I_1, I_2, \ldots, I_K\}$ is a set of input images and $\boldsymbol{q}$ is a textual instruction. A vision-language policy π_θ takes $\boldsymbol{x} = (\boldsymbol{I}, \boldsymbol{q})$ as input and generates a response $\boldsymbol{y} \sim \pi_\theta(\cdot|\boldsymbol{I}, \boldsymbol{q})$. Our goal is to improve the factual grounding of π_θ by encouraging $\boldsymbol{y}$ to be consistent with $\boldsymbol{I}$, reducing hallucinated objects, attributes, and relations.

In the standard RLVR setting, a scalar reward $r(\boldsymbol{x}, \boldsymbol{y})$ can be computed by verifying the model output against a ground-truth answer or an executable checker. The policy is then optimized to maximize the expected reward while regularizing its deviation from a reference policy π_{ref}:

$$\max_{\pi_\theta} \mathbb{E}_{\boldsymbol{y}\sim\pi_\theta(\cdot|\boldsymbol{I},\boldsymbol{q})} \left[r(\boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y})\right] - \beta \mathrm{KL}[\pi_\theta \| \pi_{\text{ref}}],$$

where β controls the strength of KL regularization, omitted in the following for simplicity.

We instantiate this objective with Group Relative Policy Optimization (GRPO) (Shao et al., 2024). For each input $x = (I, q)$, GRPO samples a group of G responses from the rollout policy $\pi_{\bar{\theta}}$:

$$\{\boldsymbol{y}_i\}_{i=1}^{G} \sim \pi_{\bar{\theta}}(\cdot|\boldsymbol{I}, \boldsymbol{q}),$$

where each response receives a scalar reward $r_i = r(\boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}_i)$. GRPO then normalizes the rewards within the group to obtain the relative advantage $\hat{A}_i = (r_i - \mu_r)/(\sigma_r + \epsilon_{\text{std}})$, where μ_r and σ_r are the mean and standard deviation of the rewards in the group, and ϵ_{std} is a small constant for numerical stability. The policy is updated using the clipped surrogate objective $\mathcal{L}_{\text{GRPO}}(\theta)$:

$$-\mathbb{E}\Big[\frac{1}{G}\sum_{i=1}^{G} \min(\rho_i \hat{A}_i,\ \text{clip}(\rho_i, 1-\varepsilon, 1+\varepsilon)\hat{A}_i)\Big] \tag{1}$$

where $\rho_i = \frac{\pi_\theta(\boldsymbol{y}_i|\boldsymbol{I},\boldsymbol{q})}{\pi_{\bar{\theta}}(\boldsymbol{y}_i|\boldsymbol{I},\boldsymbol{q})}$, and $\varepsilon > 0$ is the clipping parameter.

The key challenge in our setting is that the reward $r(\boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y})$ is not directly available. Unlike mathematical reasoning or code generation, where correctness can be verified by an exact answer or program execution, we cannot directly determine whether $\boldsymbol{y}$ faithfully describes the visual content of $\boldsymbol{I}$. Existing self-verification approaches can be understood as constructing a surrogate reward R_{SV} to approximate the latent reward function $r(\boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y})$. Following this perspective, we propose MARVEL, a visual multi-agent self-verification framework that provides scalable fine-grained reward signals for annotation-free policy optimization, as detailed in the following subsections.

3.2 MARVEL OVERVIEW

MARVEL adapts the Proposer–Checker paradigm (Wang et al., 2026a; Li et al., 2026) to the visual domain and extends it with a dedicated **Regulator** agent. This extension is necessary: with the image as the only evidence, consistency alone rewards claims that need no perception or that repeat each other, so the Regulator governs the quantity and quality of the claim set. As illustrated in Figure 3, MARVEL orchestrates four agents in a sequential verification pipeline:

- **Solver.** Generates a response $\boldsymbol{y}$ conditioned on $(\boldsymbol{I}, \boldsymbol{q})$ and is the only agent updated by GRPO.
- **Proposer.** Decomposes $\boldsymbol{y}$ into a set of atomic factual claims $\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m\}$ and converts each claim into a visual question-answer pair, yielding the raw claim set $\mathcal{Q} = \{(\boldsymbol{v}_i, \boldsymbol{a}_i)\}_{i=1}^{m}$.
- **Regulator.** Audits $\mathcal{Q}$ by (i) removing semantically redundant claims to prevent reward inflation via repetition, yielding the filtered set $\mathcal{Q}^* = \{(\boldsymbol{v}_i, \boldsymbol{a}_i)\}_{i=1}^{n}$, $n \le m$; (ii) rewarding sufficient claim coverage via a claim-count signal r^{cnt}; and (iii) penalizing image-agnostic claims that can be resolved without genuine visual perception via a quality signal r^{qual}. We refer to the claim-level quantity and quality control as **Visual Claim Regulation** (VCR), detailed in Section 3.4.
- **Checker.** Receives the original image I and questions $\{\boldsymbol{v}_i\}$ from $\mathcal{Q}^*$, with no access to $\boldsymbol{y}$, an information asymmetry that reduces confirmation bias. It generates image-grounded reference answers $\{\boldsymbol{a}'_i\}$ and verifies consistency with the claim-implied answers $\{\boldsymbol{a}_i\}$ via majority voting, producing a consistency reward r^{cons} that enforces factual grounding (Section 3.5).

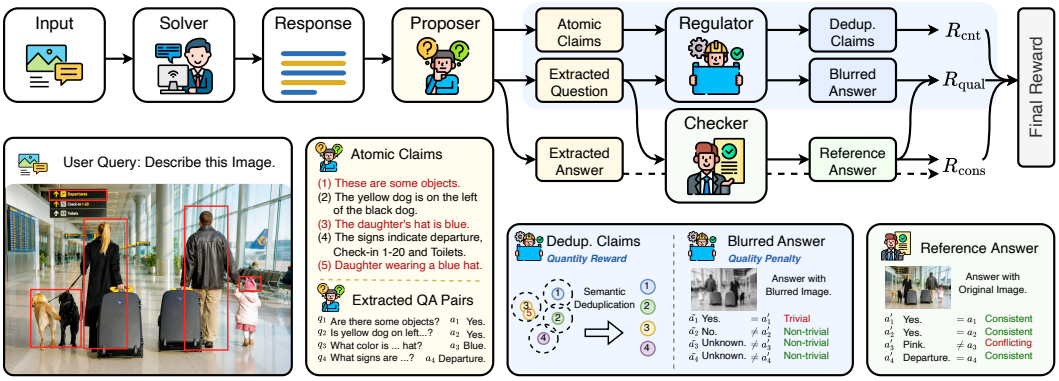


Figure 3: Overview of the MARVEL pipeline. Given a **Solver**-generated response, the **Proposer** decomposes it into atomic factual claims and converts each into a visual question-answer pair. The **Regulator** performs Visual Claim Regulation (VCR) by deduplicating semantically redundant claims, rewarding claim quantity, and penalizing image-agnostic claims that can be answered without genuine visual perception. The **Checker** re-answers each question using the original image, and verifies consistency with the Solver's answers. The final self-verifying reward aggregates the Checker's consistency signal and the Regulator's quantity and quality signals.

The final self-verifying reward aggregates signals from the Regulator and the Checker:

$$r^{\text{SV}}(\boldsymbol{I}, \boldsymbol{q}, \boldsymbol{y}) = \underbrace{r^{\text{cons}}}_{\text{Checker}} + \underbrace{r^{\text{cnt}} + r^{\text{qual}}}_{\text{Regulator}}. \tag{2}$$

This reward is the Solver's GRPO training signal, with no learned reward model or annotation.

3.3 Proposer: Claim Extraction

Given $\boldsymbol{y}$, the Proposer extracts a set of atomic factual claims:

$$\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m\} = \text{Extract}(\boldsymbol{y}), \tag{3}$$

where each c_i is a short, self-contained statement about a single visual fact (e.g., object existence, color, quantity, or spatial relation). Each claim is then converted into a visual question-answer pair:

$$(\boldsymbol{v}_i,\ \boldsymbol{a}_i) = \text{QA}(\boldsymbol{c}_i), \qquad i = 1, \ldots, m, \tag{4}$$

where $\boldsymbol{a}_i$ is the answer implied by the claim. The resulting raw claim set $\mathcal{Q} = \{(\boldsymbol{v}_i, \boldsymbol{a}_i)\}_{i=1}^{m}$ is then passed to the Regulator for quantity and quality control (Section 3.4).

Example. For the query "Describe this image" applied to an airport scene, the Solver might generate a response $\boldsymbol{y}$. The Proposer then extracts atomic claims and converts each to a QA pair, for instance: *"The yellow dog is on the left of the black dog"* $\rightarrow$ (Q: *Is the yellow dog on the left of the black dog?*, A: *Yes*); *"The daughter's hat is blue"* $\rightarrow$ (Q: *What color is the daughter's hat?*, A: *Blue*).

3.4 Regulator: Visual Claim Regulation

The Regulator addresses degenerate behaviors that undermine visual self-verification if left unchecked (Figure 6): a model can (1) repeat near-identical claims to inflate r^{cons}; (2) produce easily verified claims that earn a stable reward while remaining under-informative; or (3) generate image-agnostic claims (e.g., *"the image contains objects"*) whose answers follow from language priors rather than visual perception. The Regulator counteracts (1) through semantic deduplication, and (2) and (3) through complementary quantity and quality rewards, which together form VCR.

Semantic Deduplication. The Regulator removes semantically overlapping claims from $\mathcal{Q}$ using cosine similarity over sentence embeddings $\phi(\cdot)$:

$$\text{Sim}(\boldsymbol{c}_i, \boldsymbol{c}_j) = \frac{\phi(\boldsymbol{c}_i)^\top \phi(\boldsymbol{c}_j)}{\|\phi(\boldsymbol{c}_i)\| \cdot \|\phi(\boldsymbol{c}_j)\|}. \tag{5}$$

Pairs exceeding a type-aware threshold are merged, yielding the deduplicated claim set $\mathcal{Q}^* = \{(\boldsymbol{v}_i, \boldsymbol{a}_i)\}_{i=1}^{n}, n \leq m$.

Example. In the airport-scene example, *"Daughter wearing a blue hat"* is a near-duplicate of *"The daughter's hat is blue"* and is removed, reducing the claim set from $m = 5$ to $n = 4$.

Quantity Reward. The Regulator then adds a claim-count reward proportional to the number of deduplicated claims n, capped at a target n_{target}:

$$r^{\text{cnt}} = \lambda_{\text{cnt}} \cdot \min\left(\frac{n}{n_{\text{target}}}, 1\right), \tag{6}$$

with $\lambda_{\text{cnt}} = 0.5$ so that consistency remains dominant. Since n counts only unique claims, repetition (1) cannot inflate r^{cnt}, while r^{cnt} and r^{qual} discourage under-informative responses (2).

Quality Penalty. To address (3), the Regulator constructs a degraded image $\tilde{\boldsymbol{I}}$ by applying heavy Gaussian blur, grayscale conversion, and horizontal flipping, and re-answers each question $\boldsymbol{v}_i \in \mathcal{Q}^*$ as $\tilde{\boldsymbol{a}}_i = \text{Ans}(\tilde{\boldsymbol{I}}, \boldsymbol{v}_i)$. If $\tilde{\boldsymbol{a}}_i$ remains consistent with the reference answer $\boldsymbol{a}'_i$ from the original image (Section 3.5), the claim is trivial, i.e., answerable without meaningful visual content, and incurs a fixed penalty:

$$r^{\text{qual}} = -\lambda_{\text{qual}} \sum_{i=1}^{n} \mathbf{1}\left[\mathcal{F}(\tilde{\boldsymbol{a}}_i, \boldsymbol{a}'_i) = 1\right], \tag{7}$$

pushing the Solver toward responses that require genuine image grounding.

Example. The claim *"There are some objects in the image"* still yields *"Yes"* on the degraded image, consistent with the reference answer, and thus incurs a penalty. By contrast, *"The yellow dog is on the left of the black dog"* cannot be answered from the degraded image and is not penalized.

3.5 Checker: Image-Grounded Verification

The Checker receives the filtered claim set $\mathcal{Q}^*$ and the original image $\boldsymbol{I}$, with no access to $\boldsymbol{y}$. For each question $\boldsymbol{v}_i \in \mathcal{Q}^*$, it generates an image-grounded reference answer $\boldsymbol{a}'_i = \text{Ans}(\boldsymbol{I}, \boldsymbol{v}_i)$ and verifies consistency with the claim-implied answer $\boldsymbol{a}_i$ via majority voting over K independent runs:

$$z_i = \mathbf{1}\Big[\sum_{k=1}^{K} z_i^{(k)} > \frac{K}{2}\Big], \qquad z_i^{(k)} = \mathcal{F}(\boldsymbol{a}_i, \boldsymbol{a}'_i). \tag{8}$$

The resulting consistency reward is

$$r^{\text{cons}} = \frac{1}{n} \sum_{i=1}^{n} z_i, \tag{9}$$

i.e., the fraction of claims in $\mathcal{Q}^*$ supported by the image; if $n = 0$, we set $r^{\text{cons}} = 0$.

Final Reward. Combining the Checker and Regulator signals, the total reward for GRPO training is equation 2, whose three terms are complementary: r^{cons} enforces factual grounding, r^{cnt} promotes broader visual coverage, and r^{qual} filters out trivial and image-agnostic claims.

4 Experiment

The four functions in our method, Extract($\boldsymbol{y}$), QA($\boldsymbol{c}_i$), Ans($\boldsymbol{I}, \boldsymbol{v}_i$), and $\mathcal{F}(\boldsymbol{a}_i, \boldsymbol{a}'_i)$, are implemented via four distinct prompt templates. The details of templates are provided in Appendix A.1.

Models, Datasets, and Baselines. We evaluate MARVEL on three sizes of Qwen2.5-VL (3B, 7B, and 32B) (Bai et al., 2025b). For the main experiments, we primarily report results on Qwen2.5-VL-7B to ensure fair comparison with baselines at the same model scale. We compare against nine baselines across three categories: (1) the **Base Model**; (2) four **self-evolving** models that require no ground-truth, including **Vision-Zero** (Wang et al., 2026a) (CLEVR, Chart, and Real) and **EvoLMM** (Thawakar et al., 2025); and (3) four models post-trained via RLVR on human-labeled data: **R1-OneVision** (Yang et al., 2025b), **VLAA-Thinker** (Chen et al., 2025), **MM-Eureka** (Meng

Table 1: Performance comparison between MARVEL and baselines on multiple hallucination datasets. AMBER evaluates attribute recognition, object existence, and spatial relation understanding. MME assesses color, position, and scene perception. POPE measures object hallucination, and VSR evaluates visual-spatial reasoning. **Bold** denotes the best result and underline denotes the second best. Δ indicates the improvement of MARVEL over the base model (Qwen2.5-VL-7B).

	AMBER				MME			POPE	VSR	Avg.
Method	Attribute	Existence	Relation	Accuracy	Color	Position	Scene	–	–	–
Qwen2.5-VL-7B	84.45	94.90	75.66	85.01	90.00	82.50	77.50	86.51	90.64	85.24
VisionZero(CLEVR)	85.28	94.62	78.67	86.19	90.00	82.50	79.00	86.78	91.96	86.11
VisionZero(Chart)	84.65	95.04	74.88	84.86	90.00	82.50	77.50	86.23	90.82	85.16
VisionZero(RealWorld)	84.49	94.92	75.90	85.11	90.00	80.00	77.13	86.41	90.21	84.91
EvoLMM	85.29	94.23	77.28	85.60	92.50	81.67	79.75	**87.29**	91.35	86.11
R1-Onevision	79.39	87.02	81.43	82.61	82.50	60.83	74.50	82.42	72.73	78.16
VLAA-Thinker	84.87	95.29	75.36	85.17	92.50	85.00	77.13	86.19	90.52	85.78
MM-Eureka	85.02	94.62	75.48	85.04	92.50	82.50	77.50	85.71	90.91	85.48
OpenVLThinker	81.27	**95.55**	75.06	83.96	92.50	75.00	71.13	79.25	86.12	82.20
MARVEL	**86.37**	94.58	**84.50**	**88.48**	**95.00**	**86.67**	**82.25**	87.04	**92.03**	**88.54**
Δ	↑1.92	↓0.32	↑8.84	↑3.47	↑5.00	↑4.17	↑4.75	↑0.53	↑1.39	↑3.30

Table 3: Performance comparison between MARVEL and baselines on multiple mixed multi-task benchmarks. BLINK evaluates low-level visual perception including IQ test, jigsaw, multi-view reasoning, relative depth, and reflectance. NaturalBench measures compositional reasoning with overall accuracy, group, image, and question-level metrics. Higher is better.

	MMB	MMMB	BLINK					NaturalBench				SEED	Avg.
Method	–	–	IQ	Jigsaw	MVR	Depth	Rel.Refl.	Acc	G-Acc	I-Acc	Q-Acc	–	–
Qwen2.5-VL-7B	80.62	82.83	0.67	46.67	33.83	71.77	29.85	77.97	31.79	60.39	57.89	76.36	54.22
VZ(CLEVR)	80.99	82.93	1.33	51.33	24.06	70.97	27.61	78.18	32.58	60.76	58.37	76.48	53.80
VZ(Chart)	80.16	82.98	0.67	36.67	32.33	69.35	22.39	78.09	31.63	60.50	58.03	76.48	52.44
VZ(RealWorld)	80.42	83.13	0.00	44.00	39.85	73.39	27.61	78.08	32.26	60.50	58.18	76.59	54.50
EvoLMM	80.88	83.48	4.67	**62.00**	39.10	75.00	31.34	78.07	31.74	60.32	58.16	76.48	56.77
R1-Onevision	69.61	80.96	0.00	56.00	29.32	28.23	11.19	60.91	8.68	28.13	26.26	69.91	39.10
VLAA-Thinker	80.10	82.27	1.33	35.33	20.30	71.77	28.36	78.33	32.84	61.08	58.66	76.54	52.24
MM-Eureka	80.62	83.23	1.33	32.00	12.03	76.61	15.67	78.37	33.00	**61.29**	58.71	76.42	50.77
OpenVLThinker	75.12	79.14	0.67	56.67	20.30	51.61	20.15	66.59	16.95	37.95	36.42	72.25	44.48
MARVEL	**81.61**	**84.34**	**9.33**	56.67	**41.35**	**79.03**	**38.06**	**78.53**	**33.32**	61.16	**59.18**	**76.95**	**58.29**
Δ	↑0.99	↑1.52	↑8.67	↑10.00	↑7.52	↑7.26	↑8.21	↑0.55	↑1.53	↑0.76	↑1.29	↑0.59	↑4.07

et al., 2025), and **OpenVLThinker** (Deng et al., 2026). For training, we sample 32K examples from the GQA training split (Hudson & Manning, 2019). For evaluation, we report results on hallucination relevant benchmarks including AMBER (Wang et al., 2023), POPE (Li et al., 2023), VSR (zero-shot) (Liu et al., 2023a), and MME (Fu et al., 2026). To provide a more comprehensive assessment across capabilities, we further conduct extensive evaluation on A-OKVQA (Schwenk et al., 2022), RealWorldQA, GQA (test) (Hudson & Manning, 2019), BLINK (Fu et al., 2024), NaturalBench (Li et al., 2024a), SEEDBench (Li et al., 2024b), MMMB (Sun et al., 2025), and MMBench (Liu et al., 2024b). We implement MARVEL using the Verl framework (Sheng et al., 2025), and evaluate all models using the VLMEvalKit (Duan et al., 2024). We use a learning rate of 2×10^{-6}, a batch size of 128, and a group size of 8. We set the deduplication threshold to 0.8, $n_{\text{target}} = 5$, and $\lambda_{\text{qual}} = 0.1$; results are robust to them (Figure 7). Further details are given in Appendix A.2.

4.1 Main Results

We first evaluate on hallucination-related benchmarks in Table 1, followed by real-world VQA benchmarks in Table 2, and finally evaluate on multiple mixed benchmarks in Table 3.

Mitigating Hallucination. Table 1 reports results on hallucination-related benchmarks. MARVEL achieves the best average score of 88.54, outperforming all baselines by a clear margin and improving over the base model by 3.30 points. The gains are particularly pronounced in spatial and relational understanding: MARVEL improves Relation recognition by 8.84 points and Position perception by 4.17 points over the base model. The color dimension and attribute recognition also see substantial boosts of 5.0 and 1.92 points, respectively. Notably, RLVR methods trained on labeled

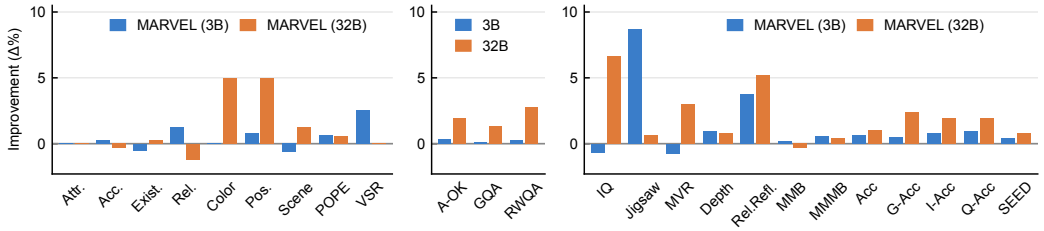


Figure 5: Per-metric improvements ($\Delta\%$) of MARVEL over the respective Qwen2.5-VL-3B and -32B base models, across (a) Hallucination, (b) Real-World VQA, and (c) Mixed Benchmarks.

data (R1-OneVision, OpenVLThinker) underperform the base model on several metrics, indicating that indiscriminate supervision can introduce hallucinations while targeting other capabilities.

Strong Performance on Real-World VQA. In Table 2, MARVEL achieves the best performance on all three benchmarks, averaging 73.16 and surpassing the strongest baseline, EvoLMM, by 0.86 points. The largest gain is on RealWorldQA (+3.53 over the base model), which requires understanding complex real-world scenes, suggesting that MARVEL develops more robust visual grounding beyond template-driven or synthetic training signals.

Table 2: Performance comparison between MARVEL and baselines on open-ended VQA benchmarks. A-OKVQA (A-OK) evaluates knowledge-grounded visual question answering. RealWorldQA (RWQA) measures question answering over complex real-world visual scenes.

Method	A-OK	GQA	RWQA	Avg.
Qwen2.5-VL-7B	86.38	60.46	67.45	71.43
VZ(CLEVR)	86.81	<u>60.61</u>	67.97	71.80
VZ(Chart)	85.76	60.28	67.97	71.34
VZ(RealWorld)	86.29	60.35	67.97	71.54
EvoLMM	<u>86.90</u>	60.59	<u>69.41</u>	<u>72.30</u>
VLAA-Thinker	85.68	60.34	65.49	70.50
MM-Eureka	85.50	60.25	65.10	70.28
MARVEL	**87.60**	**60.92**	**70.98**	**73.16**
Δ	↑1.22	↑0.45	↑3.53	↑1.74

Comprehensive Gains on Multiple Mixed Benchmarks. Table 3 presents results on general-purpose benchmarks spanning low-level visual perception, compositional reasoning, multi-image understanding, and holistic multimodal understanding. MARVEL achieves the highest average score of 58.29, surpassing the previous best (EvoLMM, 56.77) by 1.52 points and improving over the base model by 4.07 points. In BLINK, MARVEL achieves the best or second-best score on all five subtasks. On NaturalBench, MARVEL again leads overall with improvements across all four metrics. On the broader benchmarks MMBench and MMBench-Multi, MARVEL scores 81.61 and 84.34, respectively, and the highest SEEDBench score of 76.95, demonstrating well-rounded capability across standard multimodal evaluation suites.

4.2 Model Generalization

Consistent gains across all scales. To evaluate whether MARVEL generalizes across model scales, we apply the identical training recipe to Qwen2.5-VL-3B and 32B without any scale-specific modifications, and measure per-metric improvements relative to the respective base models. As summarized in Figure 4, every cell of the 3×3 grid is positive, demonstrating that the claim-level self-verification signal benefits models regardless of capacity. The 32B model improves substantially across all three categories ($+1.17$, $+1.99$, $+2.04$), with particularly strong per-metric gains of $+5.00$ on both MME/Color and MME/Position shown in Figure 5. Even the 3B model benefits consistently: it achieves $+8.67$ on BLINK/Jigsaw and $+2.52$ on VSR zero-shot, yielding a $+1.33$ average improvement on Mixed benchmarks. These results confirm that MARVEL generalizes reliably from 3B to 32B, without any scale-specific adaptations.



Figure 4: Average improvements ($\Delta\%$) of MARVEL over the Qwen2.5-VL base model across three model scales and three benchmark categories. All cells are positive, demonstrating consistent gains at every scale and category.

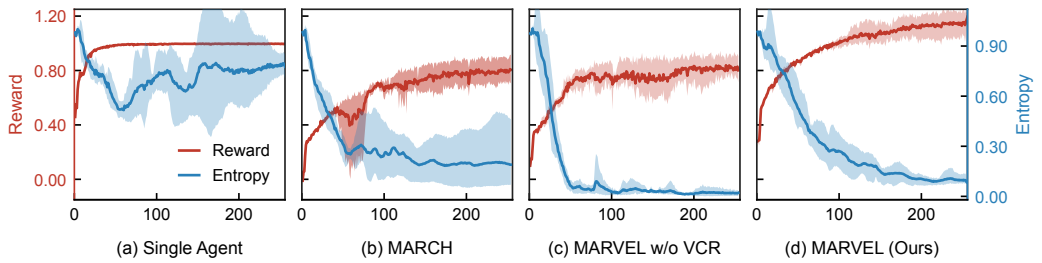


Figure 6: Reward and entropy curves over 256 training steps for (a) Single Agent SelfCheck, (b) MARCH (Li et al., 2026), (c) MARVEL w/o VCR, and (d) **MARVEL**. The comparison demonstrates that claim-level verification with VCR is necessary for stable, non-degenerate training.

4.3 Ablation Studies

VCR Analysis. To assess the contribution of each VCR component, we ablate them individually and report results in Table 4. **Dedup.** proves most critical for the Hallucination benchmark (-1.92): without semantic deduplication, the model exploits near-identical claims to inflate the consistency reward without genuinely improving perceptual accuracy, a form of reward hacking that degrades grounded perception. r^{cnt} exerts the broadest influence, yielding the largest drops on Mixed (-2.85) and notable degradation on Real-World VQA (-1.63) and Hallucination (-1.71); this pattern suggests that incentivizing sufficient claim diversity is essential for comprehensive visual coverage across task types. r^{qual} produces comparatively modest but consistent declines across all three categories (-0.71, -0.87, -1.28), confirming its complementary role in steering the Solver away from image-agnostic claims toward genuinely visually grounded responses. Removing VCR entirely, i.e., directly transferring text-based self-verification with a consistency-only reward, causes the largest drops and even falls below the base model on Mixed (53.96 vs. 54.22), so the gains come from VCR rather than from the transfer itself; VCR also improves free-form COCO captioning (Appendix A.5). A human audit of 200 claims further confirms reward reliability: claim extraction is 94.5% faithful and the consistency reward r^{cons} correlates with human factuality judgments at Pearson $r = 0.83$ (Appendix A.5).

Table 4: Ablation study of the VCR: claim deduplication (Dedup.), claim quantity reward (r^{cnt}), and claim quality penalty (r^{qual}). Each variant removes one component, and w/o VCR removes all three (consistency reward only); Decrease denotes the drop relative to the full MARVEL.

Method	Hallucination	VQA(s)	Mixed(s)
MARVEL	**88.54**	**73.16**	**58.29**
— w/o Dedup.	86.63	72.03	57.86
Decrease	-1.92	-1.13	-0.43
— w/o r^{cnt}	86.83	71.53	55.45
Decrease	-1.71	-1.63	-2.85
— w/o r^{qual}	87.83	72.30	57.01
Decrease	-0.71	-0.87	-1.28
— w/o VCR	86.62	71.59	53.96
Decrease	-1.92	-1.57	-4.33

Granularity of Self-Verification. A key design choice in MARVEL is the use of *claim-level* self-verification. To empirically validate the superiority of our decomposed approach, we implement a response-level baseline that replaces the multi-stage claim pipeline. Further implementation details are provided in Appendix A.6. As shown in Table 5, claim-level verification consistently outperforms its response-level counterpart across all three benchmark categories. These results demonstrate that decomposing verification into atomic claims enables more precise reward signals during training, leading to more hallucination-resistant models.

Table 5: Comparison of claim-level and response-level self-verification on the three benchmark categories. Decrease denotes the performance drop relative to the claim-level.

Method	Hallucination	VQA(s)	Mixed(s)
Claim-Level	**88.54**	**73.16**	**58.29**
Response-Level	87.70	72.51	56.99
Decrease	-0.84	-0.66	-1.30

Training Dynamics. Figure 6 compares training stability across four configurations to isolate the contributions of VCR and claim-level reward structure. MARVEL pipeline is the only configuration that achieves both monotonically increasing reward and controlled entropy decay above 0.1, reflecting genuine policy improvement rather than exploitative shortcutting. This training-time stability,

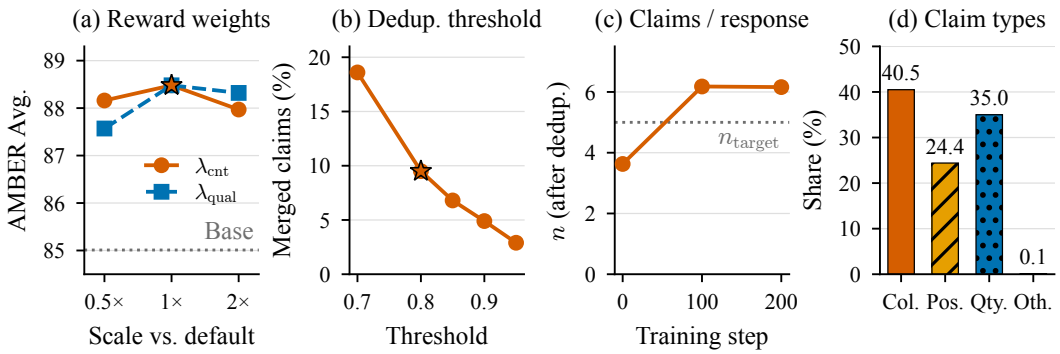


Figure 7: VCR is robust to its hyperparameters and does not induce over-claiming. (a) AMBER average when scaling λ_{cnt} (default 0.5) or λ_{qual} (default 0.1) by $0.5\times$, $1\times$, or $2\times$. (b) Share of raw claims merged when re-applying deduplication with different thresholds. (c) Deduplicated claims per response during training. (d) Share of kept claims by type. Stars mark default values.

conferred by the claim quantity reward r^{cnt} and quality penalty r^{qual} of VCR, directly underlies the performance advantage MARVEL demonstrates across all three benchmark categories.

Robustness and Rollout Behavior. Figure 7 tests the sensitivity of VCR and whether its rewards invite over-claiming. Halving or doubling λ_{cnt} or λ_{qual} changes the AMBER average by at most 0.91 points, and every setting stays $+2.6$ to $+3.5$ points above the base model (a). For deduplication, only 2.8% of 85,777 claim pairs from 2,835 rollouts exceed a similarity of 0.8, and lowering the threshold to 0.7 nearly doubles the merge rate and starts merging distinct claims, so 0.8 sits at the elbow (b). During training, the number of deduplicated claims per response rises from 3.63 to about 6.2 and then plateaus just above n_{target}, where r^{cnt} saturates (c), while response length settles at about 126 words; claims stay spread across color, position, and quantity (d), so the policy neither over-claims nor collapses onto one easily verified claim type, in line with the stable training in Figure 6.

Computational Overhead. Table 6 reports the wall-clock time per training step for each method. EvoLMM (Thawakar et al., 2025) trains separate Proposer and Solver models, and the switching overhead more than doubles the average step time compared to SelfCheck. Its training time is therefore estimated as twice that of SelfCheck, although the actual cost may be even higher. MARVEL comprises four agents, with only the Solver being trainable and the other three (Proposer, Regulator, Checker) frozen. This design results in a per-step runtime of 281.21 s, representing a $1.64\times$ overhead compared to SelfCheck (GT) (171.24 s). Given the consistent gains across all evaluation dimensions, this overhead is a favorable trade-off between compute cost and reward quality.

Table 6: Wall-clock time per training step, along with configuration #Agent(trained/total) and relative overhead compared to SelfCheck (GT). *Estimated as twice the step time of SelfCheck.

Method	Time / Step (s)	#Agents	Overhead
SelfCheck (GT)	171.24	1/2	×1.00
SelfCheck	180.02	1/2	×1.05
EvoLMM*	360.04	2/2	×2.10
MARVEL	281.21	1/4	×1.64

5 Conclusion

We presented MARVEL, a visual multi-agent self-verification framework that eliminates the need for ground-truth annotations by decomposing model responses into atomic visual claims and rewarding their consistency with the source image. The central contribution, Visual Claim Regulation, addresses two practical failure modes in claim-level verification: reward hacking through repetitive claims and the generation of image-agnostic claims. Mitigating these issues is essential for maintaining stable and non-collapsing training dynamics. Across 12 hallucination, VQA, and mixed benchmarks, MARVEL consistently outperforms competitive baselines, and scales reliably from 3B to 32B parameters. We hope this work encourages broader adoption of structured, annotation-free self-verification as a practical path toward more grounded and trustworthy vision-language models.

AI USE STATEMENT

In this work, we used generative AI tools for translation. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims or proofs, propose or refine hypotheses, design research methodology or experiments, implement methods, clean or reformat datasets, perform qualitative or thematic data analysis, or interpret results. Additionally, we used generative AI tools to polish the writing of the paper to improve readability. We have reviewed all AI-assisted text and verified that it faithfully reflects our methods and results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work aims to improve the factual reliability of vision-language models, with potential benefits in high-stakes application domains such as medical image interpretation, legal document analysis, and accessibility tools for visually impaired users. All experiments rely exclusively on publicly available academic datasets and pretrained models; no new data were collected, no human subjects were involved, and no personally identifiable information was processed. As with any advance in the fidelity of generative models, more accurate visual descriptions could in principle be misused to produce more convincing multimodal misinformation, though the primary intent and expected application of this work is the reduction of model hallucination. Finally, the multi-agent verification pipeline incurs a modest additional computational cost (approximately $1.64\times$ that of a single-agent baseline), which may lead to a larger carbon footprint; we encourage future work to explore more efficient verification architectures.

REPRODUCIBILITY STATEMENT

To ensure reproducibility, we provide full implementation details in the appendix, including training hyperparameters (Appendix A.2) and the complete prompt templates used by the Proposer, Regulator, and Checker agents (Appendix A.1). Dataset construction, evaluation protocols, and benchmark details are described in Appendix A.3. All experiments use publicly available datasets and pretrained model checkpoints. Our code, training configurations, and evaluation scripts are available at an anonymous repository: https://anonymous.4open.science/r/MARVEL.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 2025.

Neeraj Anand, Samyak Jha, Udbhav Bamba, and Rahul Rahaman. CRoPS: A training-free hallucination mitigation framework for vision-language models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=KQSoZDPVGX.

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, 2025b. URL https://arxiv.org/abs/2502.13923.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Hardy Chen, Haoqin Tu, Fali Wang, Hui Liu, Xianfeng Tang, Xinya Du, Yuyin Zhou, and Cihang Xie. SFT or RL? an early investigation into training R1-Like reasoning large vision-language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=wZI5qkQeDF.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*, 2024a. URL https://openreview.net/forum?id=EYvEVbfoDp.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, and Sungzoon Cho. Do you keep an eye on what i ask? mitigating multimodal hallucination via attention-guided ensemble decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ziw5bzg2NO.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93, 2024.

Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, Quanquan Gu, James Zou, Kai-Wei Chang, and Wei Wang. Enhancing large vision language models with self-training on image comprehension. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=FZW7Ctyjm3.

Yihe Deng, Hritik Bansal, Fan Yin, Nanyun Peng, Wei Wang, and Kai-Wei Chang. OpenVLThinker: Complex vision-language reasoning via iterative SFT-RL cycles. *Advances in Neural Information Processing Systems*, 38:123817–123846, 2026.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=DgH9YCsqWm.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16563–16577, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.850. URL https://aclanthology.org/2025.findings-acl.850/.

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

Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6700–6709, 2019.

Ayush Rajesh Jhaveri, Anthony GX-Chen, Ilia Sucholutsky, and Eunsol Choi. Failing to falsify: Evaluating and mitigating confirmation bias in language models. *arXiv preprint arXiv:2604.02485*, 2026.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, et al. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating vision-language models on natural adversarial samples. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024a. URL https://openreview.net/forum?id=Dx88A9Zgnv.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, June 2024b.

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 Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.205. URL https://aclanthology.org/2025.acl-long.205/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhuo Li, Yupeng Zhang, Pengyu Cheng, Jiajun Song, Mengyu Zhou, Hao Li, Shujie Hu, Yu Qin, Erchao Zhao, Xiaoxi Jiang, et al. MARCH: Multi-agent reinforced self-check for LLM hallucination. *arXiv preprint arXiv:2603.24579*, 2026.

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.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

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. *Advances in neural information processing systems*, 36:34892–34916, 2023b.

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. Springer, 2024b.

Yifan Lu, Ziqi Zhang, Chunfeng Yuan, Jun Gao, Congxuan Zhang, Xiaojuan Qi, Bing Li, and Weiming Hu. Mitigating hallucinations in large vision-language models by self-injecting hallucinations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13861–13877, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.746. URL https://aclanthology.org/2025.findings-emnlp.746/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

Vipula Rawte, Aryan Mishra, Amit Sheth, and Amitava Das. Defining and quantifying visual hallucinations in vision-language models. In Trista Cao, Anubrata Das, Tharindu Kumarage, Yixin Wan, Satyapriya Krishna, Ninareh Mehrabi, Jwala Dhamala, Anil Ramakrishna, Aram Galystan, Anoop Kumar, Rahul Gupta, and Kai-Wei Chang (eds.), *Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pp. 501–510, Albuquerque, New Mexico, May 2025. Association for Computational Linguistics. ISBN 979-8-89176-233-6. doi: 10.18653/v1/2025.trustnlp-main.32. URL https://aclanthology.org/2025.trustnlp-main.32/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pp. 1279–1297, 2025.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the self-verification limitations of large language models on reasoning and planning tasks. In *International Conference on Learning Representations*, volume 2025, pp. 98190–98243, 2025.

Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Parrot: Multilingual visual instruction tuning. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=Yx0bl5OFvc.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in MLLMs with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26147–26159, 2025.

Omkar Thawakar, Shravan Venkatraman, Ritesh Thawkar, Abdelrahman Shaker, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. EvoLMM: Self-evolving large multimodal models with continuous rewards. *arXiv preprint arXiv:2511.16672*, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

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

Qinsi Wang, Bo Liu, Tianyi Zhou, Jing Shi, Yueqian Lin, Yiran Chen, Hai Helen Li, Kun Wan, and Wentian Zhao. Vision-Zero: Scalable VLM self-evolution via multi-agent self-play. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=s00SNXREV6.

Tianbo Wang, Yuqing Ma, Kewei Liao, Zhange Zhang, Simin Li, Jinyang Guo, and Xianglong Liu. AFTER: Mitigating the object hallucination of LVLM via adaptive factual-guided activation editing. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=ggycXmhrrG.

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 2024*, pp. 15840–15853, 2024.

Xumeng Wen, Zihan Liu, Shun Zheng, Shengyu Ye, Zhirong Wu, Yang Wang, Zhijian Xu, Xiao Liang, Junjie Li, Ziming Miao, et al. Reinforcement learning with verifiable rewards implicitly incentivizes correct reasoning in base LLMs. *arXiv preprint arXiv:2506.14245*, 2025.

Jiulong Wu, Zhengliang Shi, Shuaiqiang Wang, Jizhou Huang, Dawei Yin, Lingyong Yan, Min Cao, and Min Zhang. Mitigating hallucinations in large vision-language models via entity-centric multimodal preference optimization. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19445–19461, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.982. URL https://aclanthology.org/2025.emnlp-main.982/.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6944–6962, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.414. URL https://aclanthology.org/2024.findings-acl.414/.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=Bjq4W7P2Us.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-OneVision: Advancing generalized multimodal reasoning through cross-modal formalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2376–2385, 2025b.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via DPO: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025c.

Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14625–14634, 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, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=4OsgYD7em5.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

A APPENDIX

Organization. In this appendix, we provide additional details and descriptions that are not included in the main paper. Specifically, §A.2 presents full implementation details including training hyperparameters and prompt templates, §A.3 provides additional information on datasets and evaluation protocols, and §A.5 reports hyperparameter sensitivity, a human audit of the reward, free-form captioning results, and rollout statistics.

A.1 PROMPT TEMPLATES.

Our reward pipeline consists of four prompts applied to evaluate the factual accuracy of model-generated captions with respect to the input image.

- Extract$(\cdot)$ (CLAIMEXTRACT): Extracts verifiable atomic claims explicitly stated in the caption, strictly restricted to color, position, and quantity categories.
- QA$(\cdot)$ (CLAIMTOQA): Converts each extracted atomic claim into a structured question-answer pair based on predefined formats without reinterpreting the original claim.
- Ans$(\cdot)$ (IMAGEQA): Answers each generated question directly from the image context, utilizing the original image for reference or a degraded image for quality assessment.
- $\mathcal{F}(\cdot,\cdot)$ (ANSWERCHECK): Compares the caption-derived answer against the image-grounded answer, allowing simple linguistic variants while enforcing strict semantic checking to produce a binary correctness signal.

CLAIMEXTRACT: Extract verifiable claims from caption

```
Extract ONLY verifiable claims explicitly stated in the caption,
limited to: color, position, and quantity.

Guidelines: - Only extract claims that are explicit and directly
stated in the caption. - Do NOT infer unstated attributes or any
information not present in the caption. - Each claim MUST be grounded
in an exact text span from the caption as "evidence". - Each claim
MUST belong to exactly one of three types: color, position, or
quantity. - Each claim MUST clearly associate the attribute with a
specific object. - If no valid claim exists, output [].

[Color] Include explicit colors (e.g., "red", "light gray"). Example:
evidence: "a black cat" → claim: "The cat is black."
[Position] Allowed: left, right, center, top, bottom only. Format:
"The [object] is at the [position] of the image."
[Quantity] Explicit numbers only; exclude vague quantifiers (e.g.,
"several", "many").

Input: [Caption]: {response}
Output: [{"evidence": "...", "claim": "...", "type":
"color/position/quantity"}, ...]
```

CLAIMTOQA: Convert claims into question–answer pairs

```
Convert claims into Question-Answer pairs using templates.

Guidelines: - Do NOT rewrite or reinterpret the claim. - The
question MUST be directly derived from the claim. - The answer MUST
be a substring or direct normalization of the claim.

[Color] Q: What color is the [OBJECT]? A: [color phrase]
[Position] Q: Where is the [OBJECT] in the image? A: left / right /
top / bottom / center
[Quantity] Q: How many [OBJECT] are in the image? A: [number]

Input: [Claims]: {caption_claim}
Output: [{"claim": "...", "question": "...", "answer": "..."},
...]
```

IMAGEQA: Answer questions grounded in the image

```
Answer the questions based on the image.
Guidelines: - Answer each question based only on the image. - Color
questions: answer with a color word or phrase only. - Position
questions: answer with ONLY one of: left, right, center, top,
bottom. - Quantity questions: answer with a number word or numeral
only. - If the answer cannot be determined reliably from the image,
answer "unknown".

Input: [Questions]: {question}
Output: [{"question": "...", "answer": "..."}, ...]
```

ANSWERCHECK: Compare caption answer vs. image answer

```
Compare Answer1 (from caption) and Answer2 (from image).
Rules: - Output 1 if they mean the same thing; output 0 otherwise. -
Output 0 if Answer2 is "unknown" or either answer is empty.
Matching guidelines: - Ignore format differences: "3" == "three"
- Allow simple variants: "blue" == "dark blue" - Allow position
variants: "left" == "on the left" - Otherwise, be strict.

Input: [Answer Pairs]: {check_list}
Output: [{"answer1": "...", "answer2": "...", "is_correct": 1/0},
...]
```

A.2 IMPLEMENTATION DETAILS

Training Parameters. We train MARVEL using GRPO (Shao et al., 2024) on top of the VeRL framework (Sheng et al., 2025). The key hyperparameters are summarized in Table 7.

Image Augmentation Pipeline. To enhance the diversity of training data and improve the model's robustness to visual perturbations, we apply a stochastic image augmentation pipeline to input images during training. Each of the seven augmentation operations is applied independently with probability $p = 0.125$, with parameters drawn uniformly from the specified ranges at runtime. The full configuration is summarized in Table 9. The Figure 8 illustrates example augmented images generated by the pipeline. At each training step, every enabled operation is applied sequentially to the input image with its corresponding probability, yielding a stochastically augmented view while preserving semantic content. The augmentation parameters are drawn uniformly from the specified ranges at runtime, ensuring sufficient variation across training samples.

A.3 DATASET AND EVALUATION DETAILS

We utilize the VLMEvalKit (Duan et al., 2024) evaluation toolkit for our assessments, and the following sections provide detailed information on the evaluation protocols and datasets employed in our experiments.

Evaluation Protocol. All models are served via a local vLLM inference server with tensor parallelism and evaluated using the VLMEvalKit framework. For open-ended and multiple-choice questions requiring semantic judgment, we employ `Qwen2.5-32B-Instruct` as the judge model, accessed through a dedicated API endpoint. Evaluation is conducted in parallel across benchmarks using 64 concurrent API workers (`--api-nproc 64`) to ensure efficient throughput. Results are logged and persisted to a designated output directory for reproducibility.

Datasets Statistics. Our evaluation suite (v2) comprises 12 benchmarks spanning a diverse range of visual understanding tasks, including perception, reasoning, grounding, and multimodal question answering. Table 8 summarizes the benchmark names, primary task categories, and the number of evaluation samples for each dataset. The total evaluation set consists of 74,160 samples across all 12 benchmarks.

Table 7: Key training hyperparameters used in all experiments.

Hyperparameter	Value
Optimization	
Learning rate	2×10^{-6}
LR schedule	Cosine decay
LR warmup steps	2
LR decay steps	350
Weight decay	0.1
Gradient clipping	1.0
Training	
Training algorithm	GRPO
Global batch size	128
Rollouts per prompt (n)	8
Training epochs	5
Max prompt length	8,192 tokens
Max response length	2,048 tokens
Loss aggregation	Token mean
PPO / GRPO	
Advantage estimator	GRPO
KL coefficient	0.02
KL loss coefficient	0.0
Clip ratio (low / high)	0.2 / 0.2
Entropy coefficient	0.0
Generation	
Temperature	1.0
Top-p	1.0
Top-k	-1 (disabled)
GPU memory utilization	0.6
Parallelism (7B)	
Tensor parallel size	2
Pipeline parallel size	2
Generation tensor parallel	2

Table 8: Statistics of the 12 evaluation benchmarks used in our experiments. Benchmarks are sorted by number of samples.

Benchmark	#Samples
RealWorldQA	765
A-OKVQA (Schwenk et al., 2022)	1,145
VSR (zero-shot) (Liu et al., 2023a)	1,222
MMMB (en) (Sun et al., 2025)	1,980
MME (Fu et al., 2026)	2,374
POPE (Li et al., 2023)	5,127
BLINK (circular) (Fu et al., 2024)	5,622
MMBench-TEST (V1.1) (Liu et al., 2024b)	7,299
NaturalBench (Li et al., 2024a)	7,600
GQA (Hudson & Manning, 2019)	12,578
AMBER (Wang et al., 2023)	14,216
SEEDBench (Li et al., 2024b)	14,232
Total	**74,160**

A.4 Ablation Details

A.5 Additional Analyses

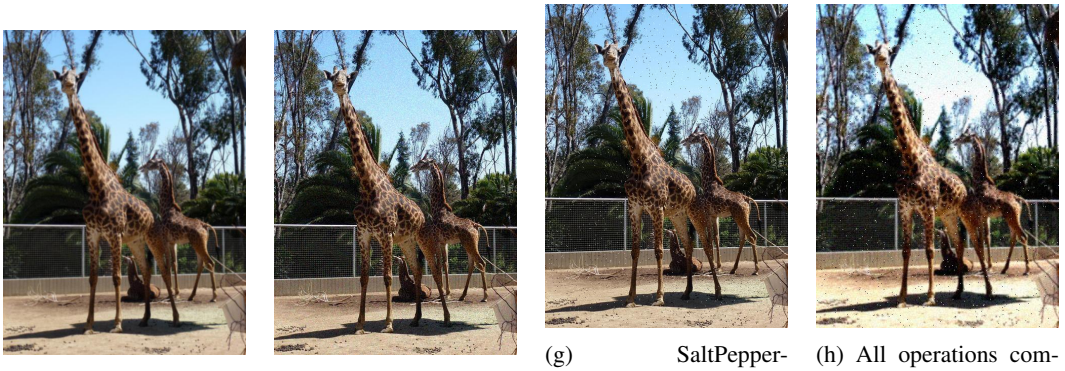
Hyperparameter Sensitivity. We vary one VCR hyperparameter at a time on Qwen2.5-VL-7B and report the AMBER average in Figure 7(a). Every setting outperforms the base model by $+2.6$

Table 9: Image augmentation operations and their configurations.

Operation	**Range**	**Probability**
Resize	$[0.80,\ 1.20]$	0.125
Brightness	$[0.80,\ 1.20]$	0.125
Contrast	$[0.80,\ 1.20]$	0.125
Sharpness	$[0.00,\ 2.00]$	0.125
Gaussian Blur	$[0.00,\ 1.00]$	0.125
Gaussian Noise	$[0.00,\ 0.05]$	0.125
Salt-and-Pepper Noise	$[0.00,\ 0.01]$	0.125



(a) Resize (factor=0.8). (b) Brightness (factor=1.2). (c) Contrast (factor=1.2). (d) Sharpness (factor=2.0).



(e) GaussianBlur(1.0). (f) GaussianNoise(0.05). (g) SaltPepperNoise(0.01). (h) All operations combined.

Figure 8: Example augmented images generated by our pipeline. The original image is transformed using each of the seven augmentation operations independently, as well as all operations combined (bottom-right).

Table 10: Ablation study on the augmentation (*aug*) component. "Decrease" denotes the performance change after removing the augmentation module from MARVEL. Negative values indicate performance degradation.

	AMBER				MME			POPE	VSR	Avg.
Method	Attribute	Existence	Relation	Accuracy	Color	Position	Scene	–	–	–
MARVEL	**86.37**	**94.58**	**84.50**	**88.48**	**95.00**	**86.67**	**82.25**	87.04	**92.03**	**88.54**
MARVEL w/o aug	86.39	94.21	84.07	88.23	92.50	**86.67**	80.63	**87.25**	91.76	87.97
Decrease	0.03	−0.37	−0.42	−0.25	−2.50	0.00	−1.63	0.21	−0.26	−0.58

to $+3.5$ points, and the spread across settings is at most 0.91 points, so the gains do not hinge on a particular choice. The default values lie at the top of both curves; increasing λ_{cnt} to 1.0 slightly lowers the score, indicating that over-weighting claim count is counterproductive. For the dedupli-

Table 11: Ablation study on the augmentation (*aug*) component for open-ended VQA benchmarks. "Decrease" denotes the performance change after removing the augmentation module from MARVEL. Negative values indicate performance degradation.

Method	A-OK	GQA	RWQA	Avg.
MARVEL	**87.60**	60.92	**70.98**	**73.16**
MARVEL w/o aug	86.81	**60.99**	68.76	72.19
Decrease	−0.79	0.07	−2.22	−0.98

Table 12: Ablation study on the augmentation (*aug*) component for general benchmarks. BLINK evaluates low-level visual perception including IQ test, jigsaw, multi-view reasoning, relative depth, and reflectance. NaturalBench measures compositional reasoning with overall accuracy, group, image, and question-level metrics. "Decrease" denotes the performance change after removing the augmentation module from MARVEL. Negative values indicate performance degradation.

	MMB	MMMB	BLINK					NaturalBench				SEED	Avg.
Method	–	–	IQ	Jigsaw	MVR	Depth	Rel.Refl.	Acc	G-Acc	I-Acc	Q-Acc	–	–
MARVEL	**81.61**	**84.34**	**9.33**	56.67	**41.35**	**79.03**	**38.06**	**78.53**	33.32	**61.16**	**59.18**	**76.95**	**58.29**
MARVEL w/o aug	81.14	83.99	6.00	**58.00**	36.84	77.42	34.33	78.37	**33.42**	61.13	58.76	76.56	57.16
Decrease	−0.47	−0.35	−3.33	1.33	−4.51	−1.61	−3.73	−0.16	0.11	−0.03	−0.42	−0.39	−1.13

Table 13: Ablation study of different data processing components on hallucination benchmarks. "w/o dedup" removes the data deduplication strategy, "w/o count" removes the object-count balancing strategy, and "w/o quality" removes the quality filtering strategy.

	AMBER				MME			POPE	VSR	Avg.
Method	Attribute	Existence	Relation	Accuracy	Color	Position	Scene	–	–	–
Qwen2.5-VL-7B	84.45	94.90	75.66	85.01	90.00	82.50	77.50	86.51	90.64	85.24
MARVEL	86.37	94.58	84.50	88.48	95.00	86.67	82.25	87.04	92.03	88.54
MARVEL-nodedup	85.51	94.90	81.25	87.22	92.50	82.50	77.13	86.83	91.81	86.63
MARVEL-nocount	86.05	94.13	83.11	87.77	90.00	84.17	77.50	87.07	91.69	86.83
MARVEL-noquality	85.66	94.52	82.69	87.62	95.00	85.83	81.88	86.24	91.04	87.83

Table 14: Ablation study of different data processing components on real-world VQA benchmarks.

Method	A-OK	GQA	RWQA	Avg.
Qwen2.5-VL-7B	86.38	60.46	67.45	71.43
MARVEL	87.60	60.92	70.98	73.16
MARVEL-nodedup	86.64	60.57	68.89	72.03
MARVEL-nocount	85.68	61.21	67.71	71.53
MARVEL-noquality	86.81	61.32	68.76	72.30

Table 15: Ablation study of different data processing components on mixed multi-task benchmarks.

	MMB	MMMB	BLINK					NaturalBench				SEED	Avg.
Method	–	–	IQ	Jigsaw	MVR	Depth	Rel.Refl.	Acc	G-Acc	I-Acc	Q-Acc	–	–
Qwen2.5-VL-7B	80.62	82.83	0.67	46.67	33.83	71.77	29.85	77.97	31.79	60.39	57.89	76.36	54.22
MARVEL	81.61	84.34	9.33	56.67	41.35	79.03	38.06	78.53	33.32	61.16	59.18	76.95	58.29
MARVEL-nodedup	81.61	83.33	3.33	58.67	41.35	79.84	35.82	78.62	34.21	61.63	59.21	76.69	57.86
MARVEL-nocount	81.09	83.03	1.33	60.67	25.56	76.61	29.85	78.14	33.21	61.05	58.45	76.35	55.45
MARVEL-noquality	81.66	83.79	7.33	56.00	30.08	77.42	38.06	78.57	33.84	61.63	59.08	76.71	57.01

cation threshold, we log all pairwise claim similarities in 2,835 training responses (85,777 claim pairs, steps 0/100/200). Pairwise similarities are concentrated at low values (median ≈ 0.22); only 6.1% of pairs exceed 0.7 and 2.8% exceed 0.8, so near-duplicates form a small, well-separated tail. Re-applying deduplication with different thresholds (Figure 7(b)) shows that 0.8 is the elbow: low-

ering it to 0.7 nearly doubles the merge rate and starts merging genuinely distinct claims (e.g., *"The van is white"* and *"There is one van"*, similarity 0.75), whereas pairs above 0.8 are paraphrases or exact repeats, and the curves flatten beyond 0.8.

Human Audit of the Reward. To validate the reward independently of the Checker, we manually audit 200 randomly sampled claims from training rollouts (Table 16). Claim extraction is faithful to the response in 94.5% of cases, the triviality label used by r^{qual} agrees with human judgment in 89.5% of cases, and r^{cons} correlates strongly with human factuality judgments (Pearson $r = 0.83$). Disagreements are dominated by borderline cases rather than systematic errors.

Table 16: Human audit of the reward pipeline on 200 randomly sampled claims.

Stage	Agreement with human
Claim extraction faithful to response	94.5%
Triviality label (r^{qual})	89.5%
r^{cons} vs. human factuality (Pearson r)	0.83

Free-Form Generation. All main benchmarks are discriminative. To test whether VCR also improves free-form descriptions, we evaluate open-ended captioning on COCO val (Lin et al., 2014) (Table 17). MARVEL improves BLEU-1 by $+6.81$ and ROUGE-L by $+6.51$ over the base model, whereas the variant without VCR stays close to the base model, so the gain comes from VCR. Absolute BLEU scores are low for all models because Qwen2.5-VL writes long paragraph captions that share few n-grams with the short COCO references; we therefore interpret these numbers as a controlled relative comparison.

Table 17: Open-ended captioning on COCO val with Qwen2.5-VL-7B.

Method	BLEU-1	BLEU-4	ROUGE-L
Base model	15.65	2.58	14.72
MARVEL w/o VCR	15.18	3.05	14.52
MARVEL	**22.46**	**4.10**	**21.23**

Rollout Statistics. Figure 7(c,d) summarizes 2,835 training responses (15,428 kept claims) at steps 0, 100, and 200. The number of deduplicated claims per response rises from 3.63 to about 6.2 and then stays flat just above $n_{\text{target}} = 5$, where r^{cnt} saturates, and response length settles at about 126 words after an initial increase; the policy therefore does not over-claim or grow verbose to inflate the reward. Claims cover all three types (only 0.1% fall outside color, position, and quantity); 86.2% of responses contain all three types and 96.6% contain at least two. No response yields an empty claim set ($n = 0$). Trivial claims are rare (0.85 per response on average), the mean magnitude of r^{qual} is 0.085, well below the cap of r^{cnt} (0.5), and only 0.39% of rollouts have $|r^{\text{qual}}| \geq 0.5$, so the quality penalty rarely cancels the quantity reward.

A.6 Response-Level Self-Verification

In addition to the claim-level self-verification described in the main paper, we also explore a response-level self-verification approach where the model evaluates the overall factual accuracy of its generated response with respect to the input image. Specifically, we design a response-level reward pipeline that directly assesses the correctness of the entire caption without decomposing it into atomic claims.

ResponseJudge: Holistic response-level verification

```
Evaluate the following image description for factual accuracy based on
the image.

Scoring Guidelines:
- Score 1.0: Fully accurate and consistent with the image. No
hallucinations or factual errors.
```

```
- Score 0.5: Partially accurate. Some details are correct but there
are minor errors or omissions.
- Score 0.0: Significant hallucinations, contradictions, or largely
inconsistent with the image.
Evaluation Criteria:
- Object existence: Are the mentioned objects actually present in the
image?
- Attribute accuracy: Are colors, sizes, positions, and other
attributes correctly described?
- Actions and relationships: Are the described actions and spatial
relationships accurate?
- No fabrication: Does the description avoid mentioning things not
present in the image?

Output ONLY a JSON object:
{''score'': <0.0--1.0>, ''reason'': ''brief explanation''}

Image Description to Evaluate: {response}
```