Fine-Grained Rewards for Visual CoT: Mitigating Hallucinations in Vision-Language Models

Jimyung Park
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
Email: victorious_s@yonsei.ac.kr

Minhyuk Jeong
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
Email: wjd9496@yonsei.ac.kr

Dongjun Kim
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
Email: dj991108@yonsei.ac.kr

Hyunjun Yuh
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
Email: hyunjunyuh@yonsei.ac.kr

Jeonghoon Mo*
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
Email: j.mo@yonsei.ac.kr

***Abstract*—Large Vision-Language Models (LVLMs) often hallucinate objects, relations, or attributes not grounded in the input image. Existing approaches such as cross-entropy training and response-level preference optimization (e.g., RLHF, DPO) fail to explicitly target hallucinations within the reasoning process. We propose a fine-grained reinforcement learning framework for Visual Chain-of-Thought (CoT), decomposing responses into [objects] → [relations] → [answer] with stage-specific rewards. Using Group Relative Preference Optimization (GRPO), our method directly penalizes hallucinations at each stage while ensuring stable training. Experiments on POPE and VQAv2 show substantial hallucination reduction with competitive task performance, demonstrating the benefit of stage-wise penalization for aligning LVLM reasoning with visual evidence.**

***Index Terms*—Vision-Language Models, Hallucination Mitigation, Reinforcement Learning, Chain-of-Thought**

I. Introduction

Recent advancements in Large Vision-Language Models (LVLMs) have enabled significant progress in multimodal reasoning tasks such as Visual Question Answering (VQA), image captioning, and referring expression comprehension (REC). Models like GPT-4V [1], Gemini [2], Qwen-VL [3], and LLaVA [4] demonstrate impressive capabilities by jointly processing images and text, powering applications in education, healthcare, autonomous driving, and industrial automation.

Despite these remarkable achievements, LVLMs remain prone to a critical issue known as **hallucination**, where the model generates objects, attributes, or relationships inconsistent with the input image. Hallucination manifests in three primary forms: (1) *Object hallucination*, where the mentioned objects are inconsistent with the objects in the image, (2) *Attribute hallucination*, where incorrect properties (e.g., color, size) are assigned, and (3) *Relation hallucination*, where relationships between objects are misrepresented. These errors are prevalent even in state-of-the-art LVLMs [5], [6], posing serious risks in safety-critical domains such as medical image analysis, autonomous vehicles, and CCTV surveillance.

Most existing methods operate at a coarse, response-level granularity, making it difficult to localize and penalize hallucinations at the stage where they occur, e.g., during object recognition, relational reasoning, or final answer generation. For instance, recent approaches such as supervised fine-tuning (SFT) with next-token prediction (NTP), multimodal reinforcement learning with human feedback (RLHF), and direct preference optimization (DPO) have shown *partial* success in mitigating hallucinations by incorporating visual signals or preference-based feedback. However, these improvements remain fundamentally limited due to the structural mismatch between their training objectives and hallucination suppression.

This study proposes a new framework that introduces *stage-wise*, *fine-grained* penalties aiming to mitigate hallucinations at their source. To directly resolve the structural mismatch identified above, we align the training objective with the actual goal of hallucination suppression. We enforce a stage-wise training signal that localizes penalties at the stage where hallucinations arise.

The response generation from our framework is structured as a Visual Chain-of-Thought (CoT), decomposed into `[objects]` → `[relations]` → `[answer]`. This decomposition enables stage-specific supervision instead of coarse, response-level signals.

Next, the following four rewards are computed: First, *Object Grounding Precision* penalizes object-level inconsistencies by evaluating whether the predicted objects in the [objects] block are visually consistent with the image, directly targeting the object hallucination (i.e., objects inconsistent with the image). Second, *Question-Groundedness* rewards question-relevant entities/attributes and penalizes spurious mentions, aligning content selection with the question intent. Third, *Reasoning Validity* evaluates the consistency of relation triples with grounded

*Corresponding author

boxes and basic geometry (e.g., left-of/on/overlap), targeting the relational hallucination. Fourth, *Answer Accuracy* assesses final answer faithfulness to prevent fluency-only optimization. Finally, these stage-specific rewards are summed to a composite reward.

For model training, we adopt a progressive stabilization approach according to the following three-step curriculum:

1) **Step 1: Object Grounding SFT** to establish reliable grounding.
2) **Step 2: Visual CoT SFT** to learn the full reasoning format while retaining stage boundaries.
3) **Step 3: Fine-Grained Group Relative Preference Optimization (GRPO)** to compute localized rewards rather than a single response-level scalar.

Unlike former approaches like NTP, DPO, and RLHF, which apply uniform or response-level signals, our stage-wise rewards (i) localize penalties to the exact stage where hallucinations occur (objects, relations, or answer), (ii) align optimization with visual faithfulness, and (iii) preserve interpretability via explicitly decomposing responses into `[objects]`, `[relations]`, and `[answer]` (See details in Fig. 1). Consequently, our approach can alleviate the mismatch between the training objective and the hallucination mitigation.

II. Related Work

A. Chain-of-Thought for LVLMs

Visual CoT [7] represents step-by-step reasoning grounded on explicit regional evidence by iteratively cropping and feeding localized visual cues. This strategy improves recognition of small or spatially confined objects and encourages models to expose intermediate reasoning. However, the supervision signal is largely organized around the output format and final responses, failing to attribute penalties to specific object, attribute, or relation mistakes that occur at different stages of the reasoning process.

Ground-R1 [8] extends Visual CoT with longer internal thoughts. It optimizes with GRPO using composite rewards that consider adherence to structure, answer accuracy, and grounding-aware terms, improving interpretability and stability. However, the reward aggregation typically operates at the response-level; consequently, errors at different stages (object existence, relations, attributes) are not always disentangled during training.

B. Preference-Based Alignment for Hallucination Mitigation

RLHF with an explicit reward model is one of the recent multimodal methods for preference-based alignment. RLHF-V [9] leverages human-edited responses and a multimodal reward model to emphasize reliable, visually consistent outputs. LLaVA-RLHF [10] trains a reward model from human preference pairs and updates the policy through proximal policy optimization (PPO). These approaches effectively integrate visual signals into alignment and have shown strong empirical gains. However, their scalar rewards are typically applied at the response-level, which poses challenges for attributing hallucinations to specific objects, relations, and intermediate steps.

DPO, which learns from pairwise preferences without an explicit reward model, is another approach to preference-based alignment. DPO variants optimize directly from preference pairs and avoid policy-gradient loops. POVID [11] improves image–text consistency by constructing negative samples that induce hallucinations and discouraging them during training. HA-DPO [12] contrasts non-hallucinatory and hallucinatory responses for the same image to steer the model toward faithful outputs. These methods provide effective alignment signals with simpler training dynamics. In practice, however, supervision often remains coarse at the level of whole responses, motivating efforts toward more targeted, fine-grained signals that better localize the error sources.

C. Self-Feedback Guided Revision

VOLCANO [13] mitigates hallucination by having a single LVLM run a *Critique* $\rightarrow$ *Revise* $\rightarrow$ *Decide* loop. The model generates an answer, produces self-feedback to flag inconsistencies, and revises the output before finalizing. This behavior is learned through SFT on a corpus of feedback–revision pairs. This reduces hallucinations without training a separate reward model and can dynamically refocus attention on relevant regions during reasoning. However, it trades additional inference-time iterations for improved faithfulness and typically provides supervision at the response-level rather than explicitly tying penalties to particular reasoning stages.

D. Fine-Grained Preference Optimization

Beyond response-level alignment, several works explore more granular supervision. FGAIF [14] decomposes outputs into atomic facts (e.g., color or presence), verifies each fact with an auxiliary model, and performs preference optimization at the fact level to emphasize visual faithfulness. These directions move supervision closer to error sources and are synergistic with approaches that make reasoning steps explicit. However, in most cases, the granularity is not yet aligned to an explicit stage-wise structure.

E. Positioning Our Work

Prior studies have advanced alignment along three complementary axes: using evidence explicitly via CoT, optimizing from human feedback or preferences, and reducing hallucination through hallucination-free training data for SFT. We build on these insights and pursue a stage-wise Visual CoT framework (`[objects]` $\rightarrow$ `[relations]` $\rightarrow$ `[answer]`) coupled with fine-grained rewards optimized via GRPO. By associating supervision signals with specific stages and error types (object existence, attributes, relations), our training objective directly targets the origins of hallucination while remaining compatible with CoT-, preference-, and self-feedback–based advances.

Question: Is there a cup in the image?

• **Baseline Output**
Yes, there appears to be a cup on the table in the foreground of the image. It is located near the center-right side of the table and seems to be a standard drinking cup, possibly for coffee or tea.

• **Ours Output**
`[objects]` table at (0.38, 0.47, 1.0, 0.69); bottle at (0.88, 0.38, 0.95, 0.52); glass window at (0.38, 0.0, 1.0, 0.49); bottle at (0.94, 0.4, 0.97, 0.49); bottle at (0.97, 0.4, 1.0, 0.51); counter at (0.38, 0.47, 1.0, 0.68); chair at (0.15, 0.42, 0.35, 0.73); `[/objects]`

`[relations]`
bottle - ON - table
chair - near - table
`[/relations]`

`[answer]` No, there are no cups in the picture. `[/answer]`

Question: How many elephants in the image?

• **Baseline Output**
There are two elephants in the image.

• **Ours Output**
`[objects]` green, grassy field at (0.0, 0.43, 1.0, 0.86); large, gray elephant at (0.6, 0.41, 0.67, 0.51); small, gray elephant at (0.55, 0.43, 0.58, 0.48); small, gray elephant at (0.5, 0.42, 0.54, 0.49); `[/objects]`

`[relations]`
elephant - IN - field
elephant - IN - field
elephant - IN - field
tree - behind - elephant
`[/relations]`

`[answer]` Three. `[/answer]`

Fig. 1. Qualitative comparison between **Baseline Output** and **Ours Output**. Top: Example of hallucination suppression in binary detection. Bottom: Example of accurate multi-object counting.

III. Method

Our goal is to mitigate hallucinations in LVLMs by explicitly supervising *where* errors occur in the reasoning process. To this end, we design a three-stage structured output, a progressive training pipeline, and a fine-grained reinforcement learning scheme. Fig. 2 provides an overview of our method.

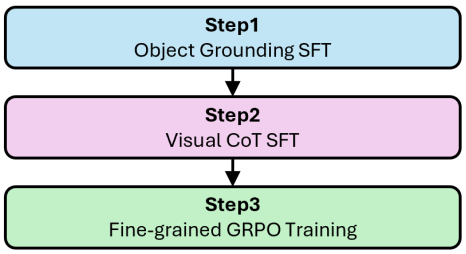

Fig. 2. **Overview of our three-stage training pipeline.** Step 1 trains the model to generate only the `[objects]` block for explicit grounding. Step 2 extends this to full Visual CoT generation. Step 3 applies GRPO with four fine-grained rewards (Grounding Precision, Question-Groundedness, Reasoning Validity, Answer Coherence) to directly mitigate hallucinations across all stages.

A. Training Data Construction from Visual Genome

To train a model that generates visually grounded reasoning steps, we require structured supervision that explicitly captures *visual evidence*, *logical reasoning*, and *final answers*. We construct training tuples $(I, \mathbf{o}, \mathbf{r}, a)$ from the Visual Genome (VG) dataset, where I is an image, $\mathbf{o}$ is a set of grounded objects, $\mathbf{r}$ is a set of relationships between these objects, and a is a short free-form answer to a given question. The detailed pipeline for data construction is as follows.

a) Structured output format: Each sample is serialized into a three-stage output format that aligns with our Visual CoT design.

Descriptions:

- `[objects]`: a list of detected objects with class labels and normalized bounding boxes, augmented with key attributes such as color and size.
- `[relations]`: spatial or semantic relationships between objects, e.g., `ON`, `HOLDING`.
- `[answer]`: a concise natural-language response to the given question.

Example:

`[objects]`

```
green bottle at (0.88, 0.38, 0.95, 0.52);
wooden table at (0.38, 0.47, 1.00, 0.69)
[/objects]
[relations]
bottle - ON - table
[/relations]
[answer]
No, there are no cups in the picture.
[/answer]
```

This explicit structure forces the model to first ground visual evidence in the `[objects]` stage, then reason about relationships in the `[relations]` stage, and finally synthesize the answer in the `[answer]` stage. This format is also directly compatible with the fine-grained rewards introduced in Step 3.

b) Region-based object selection: Each question in VG is associated with a region of interest (ROI) via region descriptions and QA-to-region mappings. We first locate the region bounding box and select all objects whose bounding boxes overlap with the region by an IoU greater than $\tau_{\text{IoU}} = 0.01$. This ensures that only objects visually relevant to the question are included. Objects are described using canonicalized class labels with up to two key attributes (e.g., "red", "small") from VG's attribute annotations. Bounding box coordinates are normalized by image width W and height H:

$$\tilde{b}_i = \Big(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}\Big), \quad \tilde{b}_i \in [0,1]^4$$

where (x_1, y_1) and (x_2, y_2) are the top-left and bottom-right corners of the object box. The final `[objects]` block contains up to 10 objects, ordered left-to-right and top-to-bottom.

c) Relation extraction and filtering: From the VG relationship annotations, we extract triplets (s, p, o) where s (subject) and o (object) are among the region-selected objects. Predicates p are normalized into a compact vocabulary $\mathcal{P} = \mathcal{P}_{\text{spatial}} \cup \mathcal{P}_{\text{semantic}}$, covering both spatial (`ON`, `IN`, `LEFT OF`) and semantic (`HOLDING`, `WEARING`) relations. To reduce noise, we keep only the top-five most relevant relations based on region overlap and frequency. The following is an example relation serialization:

$$\texttt{"person - HOLDING - umbrella"}.$$

d) Answer alignment: For QA-style supervision, the `[answer]` stage uses the ground-truth short answer provided by VG.

e) Curriculum-aware filtering: To ensure high-quality supervision for curriculum learning, we apply three filtering steps. First, we remove QA samples where no objects overlap with the region. Second, we discard relations where neither subject nor object is visually grounded in the selected region. Lastly, we limit the number of objects and relations to prevent long, noisy outputs.

f) Benefits of this design: Our data construction pipeline offers two key benefits:

- **Explicit grounding:** The normalized bounding boxes in `[objects]` provide precise visual evidence.
- **Stage-wise error attribution:** By separating reasoning into distinct stages, hallucination penalties can be localized.

This design bridges the gap between raw multimodal data and fine-grained reinforcement learning, enabling robust hallucination mitigation in subsequent training steps.

B. Training Pipeline

Our training proceeds in three progressive stages:

a) Step 1. Object Grounding SFT: The model input in this step is image I only, and the target block is `[object]`. We train the model to output serialized label–box pairs using teacher forcing and cross-entropy:

$$\mathcal{L}_{\text{step1}} = \sum_{t \in S_o} -\log \pi_\theta(y_t \mid y_{<t}, I),$$

where S_o represents tokens of the `[objects]` stage. This step isolates visual grounding and prevents leakage of spurious linguistic priors.

b) Step 2. Visual CoT SFT: In this step, the model takes image I and question x as inputs, and the target is the full structured output: `[objects]` $\rightarrow$ `[relations]` $\rightarrow$ `[answer]`. We fine-tune the model to jointly generate all stages using cross-entropy:

$$\mathcal{L}_{\text{step2}} = \sum_{t \in S_o \cup S_r \cup S_a} -\log \pi_\theta(y_t \mid y_{<t}, x, I),$$

where S_o, S_r, and S_a represent tokens for the `[objects]`, `[relations]`, and `[answer]` stages respectively.

c) Step 3. Fine-grained GRPO Training: Building on Step 2, Step 3 optimizes the same structured outputs using fine-grained reinforcement learning. For each input, we sample $N = 8$ candidate responses. Next, instead of a single global reward, we decompose the responses into stage-specific components. If the response can be successfully parsed into the required `[objects]`, `[relations]`, and `[answer]` blocks, we compute the fine-grained composite reward:

$$r = \lambda_{gp} r_{gp} + \lambda_{qg} r_{qg} + \lambda_{rv} r_{rv} + \lambda_{ac} r_{ac}.$$

where each reward term evaluates a specific stage:

- r_{gp}: Grounding Precision for `[objects]` stage,
- r_{qg}: Question-Groundedness of selected objects,
- r_{rv}: Reasoning Validity in the `[relations]` stage,
- r_{ac}: Answer Coherence for the `[answer]` stage.

Otherwise (i.e., when parsing fails), we assign a fixed penalty:

$$r = -0.2.$$

Subsequently, rewards are normalized within each group, and token-level updates are applied using the following PPO-style clipped objective:

$$\mathcal{L}_{\text{grpo}} = -\,\mathbb{E}_{i,t}\big[\min\big(\rho_{i,t}\,\hat{A}_i,\ \text{clip}(\rho_{i,t}, 1-\epsilon, 1+\epsilon)\,\hat{A}_i\big)\big],$$

where $\rho_{i,t}$ is the ratio of new vs. old policy likelihoods, and $\hat{A}_i$ is the group-normalized advantage. This design ensures stable optimization while directly addressing hallucinations in each stage.

IV. Experimental Setup

A. Evaluation Datasets

We evaluate the proposed model on two benchmarks:

- **POPE** [5]: hallucination probe with *Random*, *Popular*, and *Adversarial* splits.
- **VQAv2** [15]: 3,000 samples randomly drawn (random seed = 42) from the standard open-ended Visual Question Answering (VQAv2) validation set.

POPE [5] probes object hallucination by polling candidate objects for a given image, yielding object-level metrics. On the other hand, VQAv2 [15] measures overall question answering performance. Together, these benchmarks allow us to quantify object-level hallucination while monitoring general multimodal capability; they complement, rather than replace, our analysis of stage-wise behavior during training.

B. Metrics

In terms of POPE, we report accuracy and F1 for each split, plus the mean across splits. For VQAv2, we report accuracy.

C. Implementation Details

Backbone is Qwen2.5-VL [16] with 448×448 vision encoder. We train with 2×A100 (BF16). Steps 1 and 2 use AdamW (lr=5×10^{-5}); Step 3 uses GRPO (lr=5×10^{-6}). Batch size: 16 (SFT) / 8 (RL). For Step 1 (object grounding) and Step 2 (Visual CoT supervision), training runs for 1 epoch on our curated 43,080 Visual Genome–based samples. For Step 3, we perform GRPO training with fine-grained rewards on the GQA dataset, running for 2,240 optimization steps. Fine-grained rewards are computed by a frozen unified reward model (UnifiedReward-Think-7b [17]).

V. Results and Discussion

A. Main Results on POPE

Table I compares our method against representative 7B models, including LLaVA 1.5 [4], FGAIF [14], HA-DPO [12], LLAVA-RLHF [10], and VOLCANO [13]. Our Step3 (GRPO) attains the best mean accuracy and F1, with especially strong gains on the *Adversarial* split. Specifically, compared to strong RL/PO methods, Step3 achieves the best mean F1 (88.0) and Accuracy (88.3). While HA-DPO slightly leads on *Random*, our model is markedly stronger on the *Adversarial* split (F1: 86.1 vs. 82.5), driving the average performance and reflecting the improved resistance to hallucination. Unlike other models, our model explicitly outputs visual evidence it further reinforces this evidence through fine-grained rewards; as a result, our model can exclude plausible yet nonexistent objects from its responses.

B. POPE Ablation across Training Stages

Table II presents the ablation results of our staged training procedure on POPE. Within our pipeline, the mean F1 on POPE improves monotonically from the Baseline (82.8) → Step2 (84.3; +1.5) → Step3 (88.0; +3.7 over Step2, +5.2 over Baseline). The mean accuracy shows the same trend (85.0 → 86.0 → 88.3; +3.3 total). These gains align with our reward design: r_{gp} reduces object hallucination at the `[objects]` stage, r_{rv} filters inconsistent `[relations]`, and r_{ac} encourages answers consistent with grounded evidence.

TABLE I
POPE results across three settings. All models are 7B. Best scores are **bolded** and second ones are underlined.

Model	Random		Popular		Adversarial		mean Acc	mean F1
	Acc.	F1	Acc.	F1	Acc.	F1		
Qwen2.5-VL [16]	85.5	83.1	85.3	83.2	84.1	82.2	85.0	82.8
LLaVA 1.5 [4]	88.2	87.3	87.3	86.2	85.2	84.2	86.9	85.9
FGAIF [14]	87.0	86.7	84.0	83.7	79.6	79.9	83.5	83.4
HA-DPO [12]	**90.5**	**90.2**	87.9	88.1	81.5	82.5	86.6	86.9
LLAVA-RLHF [10]	84.8	83.3	83.3	81.8	80.7	79.5	82.9	81.5
VOLCANO [13]	89.9	89.4	**88.5**	87.9	86.2	85.7	88.2	87.7
Ours (Step2)	87.0	85.3	85.9	84.2	85.1	83.4	86.0	84.3
Ours (Step3, GRPO)	90.2	89.6	87.9	**88.2**	**86.7**	**86.1**	**88.3**	**88.0**

TABLE II
POPE by training stage (ours). Best scores are **bolded** and second ones are underlined.

Model	Random		Popular		Adversarial		mean Acc	mean F1
	Acc.	F1	Acc.	F1	Acc.	F1		
Baseline (Qwen2.5-VL) [16]	85.5	83.1	85.3	83.2	84.1	82.2	85.0	82.8
Step2 (SFT – Full Data)	87.0	85.3	85.9	84.2	85.1	83.4	86.0	84.3
Step3 (GRPO)	**90.2**	**89.6**	**87.9**	**88.2**	**86.7**	**86.1**	**88.3**	**88.0**

C. VQAv2 Results

Table III shows the accuracy of our model on VQAv2 [15]. Step3 improves $+1.2$ points over the Baseline and $+4.5$ over Step2 by reducing parsing errors and aligning reasoning with visual evidence. Notably, Step2 drops to 81.4%; this is probably due mainly to format parsing errors (i.e., missing or malformed `[answer]` blocks), which cause valid predictions to be discarded. Introducing fine-grained rewards in Step3—including a format-validity term—stabilizes structure and lifts accuracy to 84.9%, showing that structural guidance complements content rewards.

TABLE III
VQAv2 accuracy by training stage. Best score is **bolded**.

Model	VQAv2 Accuracy (%)
Baseline (Qwen2.5-VL) [16]	83.7
Step2 (SFT – Full Data)	81.4
Step3 (GRPO)	**84.9**

D. Takeaway

Our results indicate that stage-wise GRPO substantially improves hallucination robustness (POPE) while enhancing the downstream task utility (VQA). Enforcing the structured output and rewarding stage-specific correctness produce (i) stronger robustness on adversarial hallucination probes, and (ii) better end-task accuracy, without relying on global, response-level signals alone. Numbers for external methods are reproduced from their reports under comparable 7B settings [4], [10], [12]–[14], [16]; absolute values may vary with implementation details, but the trends are consistent: stage-wise, fine-grained optimization yields the best overall POPE mean and the highest VQA accuracy in our setup.

VI. Limitations and Future Work

While our approach significantly mitigates hallucinations and improves reasoning accuracy, several limitations remain. First, our training relies on datasets such as Visual Genome, which contain explicit object and relation annotations. These datasets are limited in domain coverage and may not capture the full diversity of real-world scenarios, leading to potential performance degradation when deployed in open-world environments. Second, the fine-grained rewards are computed using a frozen reward model that itself may be imperfect or biased. Incorrect judgments from the reward model can propagate errors during GRPO training, potentially over-penalizing correct outputs or reinforcing spurious correlations. Lastly, our model is primarily trained on question–answer pairs from the Visual Genome dataset. As a result, the model exhibits weaker instruction-following ability compared to models trained on broader instruction-tuning corpora.

In future research, we plan to (i) expand our training to include large-scale, diverse datasets for broader coverage, (ii) improve the reward model through various types of hallucination data to reduce bias and noise, and (iii) explore alternative response formats beyond the conventional "objects + relations + answer," including natural language explanations, explicit reasoning chains, evidence-supported responses, and structured graphs, to assess which format best enhances grounding reliability and mitigates hallucinations.

VII. Conclusion

In this paper, we proposed a fine-grained, stage-wise reinforcement learning framework to mitigate hallucinations in LVLMs. Our approach decomposes model outputs into three explicit stages, `[objects]` $\rightarrow$ `[relations]` $\rightarrow$ `[answer]`, enabling precise localization and penalization of hallucinations at their sources.

By integrating four carefully designed reward components—grounding precision, question-groundedness, reasoning validity, and answer coherence—we align the training objective directly with hallucination suppression. Extensive experiments on VQAv2 and POPE benchmarks demonstrate that our method not only achieves significant improvements in hallucination mitigation but also enhances overall reasoning performance.

This work highlights the importance of stage-wise supervision for faithful multimodal reasoning and provides a foundation for developing more reliable LVLMs. In future research, we plan to extend our approach to diverse real-world datasets and further refine the reward modeling process to reduce bias and enhance scalability.

Acknowledgement

This research has been funded (i) by the Industrial Technology Innovation Program (P246800032, Development of Multi-Modal Foundational Models and AI Accelerators for Zero-shot Intelligent Surveillance System) of the Ministry of Industry, Trade and Energy of Korea and (ii) by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education (MOE, Korea) and National Research Foundation of Korea (NRF).

References

[1] OpenAI, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[2] R. Anil *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023. [Online]. Available: https://arxiv.org/abs/2312.11805

[3] P. Wang *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. [Online]. Available: https://arxiv.org/abs/2409.12191

[4] H. Liu *et al.*, "Visual instruction tuning," in *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023. [Online]. Available: https://arxiv.org/abs/2304.08485

[5] Y. Li *et al.*, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023, pp. 292–305. [Online]. Available: https://aclanthology.org/2023.emnlp-main.20/

[6] Z. Sun *et al.*, "Aligning large multimodal models with factually augmented RLHF," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 13 088–13 110, introduces MMHal-Bench. [Online]. Available: https://aclanthology.org/2024.findings-acl.775/

[7] H. Shao *et al.*, "Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning," *arXiv preprint arXiv:2403.16999*, 2024. [Online]. Available: https://arxiv.org/abs/2403.16999

[8] M. Cao *et al.*, "Ground-r1: Incentivizing grounded visual reasoning via reinforcement learning," *arXiv preprint arXiv:2505.20272*, 2025. [Online]. Available: https://arxiv.org/abs/2505.20272

[9] T. Yu *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 (CVPR)*, 2024. [Online]. Available: https://arxiv.org/abs/2312.00849

[10] Z. Sun *et al.*, "Aligning large multimodal models with factually augmented RLHF," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 13 088–13 110. [Online]. Available: https://aclanthology.org/2024.findings-acl.775/

[11] Y. Zhou *et al.*, "Aligning modalities in vision large language models via preference fine-tuning," *arXiv preprint arXiv:2402.11411*, 2024. [Online]. Available: https://arxiv.org/abs/2402.11411

[12] Z. Zhao *et al.*, "Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization," *arXiv preprint arXiv:2311.16839*, 2023. [Online]. Available: https://arxiv.org/abs/2311.16839

[13] S. Lee *et al.*, "Volcano: Mitigating multimodal hallucination through self-feedback guided revision," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), Long Papers*, 2024, pp. 391–404. [Online]. Available: https://aclanthology.org/2024.naacl-long.23/

[14] L. Jing *et al.*, "Fgaif: Aligning large vision-language models with fine-grained ai feedback," *Transactions on Machine Learning Research (TMLR)*, 2025. [Online]. Available: https://jmlr.org/tmlr/papers/

[15] Y. Goyal *et al.*, "Making the v in VQA matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6904–6913. [Online]. Available: https://arxiv.org/pdf/1612.00837

[16] P. Wang *et al.*, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024. [Online]. Available: https://arxiv.org/abs/2409.12191

[17] U.-R. Team, "Unified-reward: Open and unified reward model for llms/lvlms," https://codegoat24.github.io/UnifiedReward/, 2025, we used the UnifiedReward-Think-7B checkpoint in our experiments.