ReLoop: "Seeing Twice and Thinking Backwards" via Closed-loop Training to Mitigate Hallucinations in Multimodal understanding

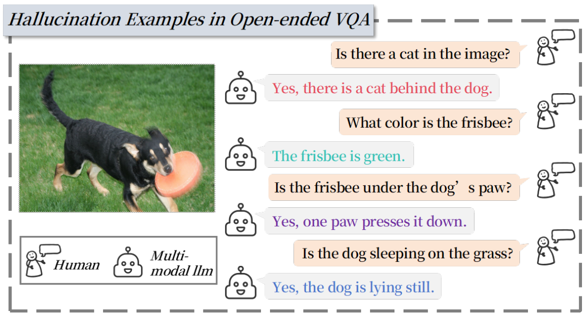
Anonymous ACL submission

Abstract

While Multimodal Large Language Models (MLLMs) have achieved remarkable progress in open-ended visual question answering, they remain vulnerable to hallucinations. These are outputs that contradict or misrepresent input semantics, posing a critical challenge to the reliability and factual consistency. Existing methods often rely on external verification or post-hoc correction, lacking an internal mechanism to validate outputs directly during training. To bridge this gap, we propose **ReLoop**, a unified closed-loop training framework that encourages multimodal consistency for cross-modal understanding in MLLMs. ReLoop adopts a ring-shaped structure that integrates three complementary consistency feedback mechanisms, obliging MLLMs to **"seeing twice and thinking backwards"**. Specifically, ReLoop employs the frozen Consistency Feedback Plugin (CFP), comprising semantic reconstruction, visual description, and an attention supervision module for attention alignment. These components collectively enforce semantic reversibility, visual consistency, and interpretable attention, enabling the model to correct its outputs during training. Extensive evaluations and analyses demonstrate the effectiveness of ReLoop in reducing hallucination rates across multiple benchmarks, establishing a robust method for hallucination mitigation in MLLMs. We will release our source code and data in the camera-ready version.

1 Introduction

In recent years, MLLMs (Liu et al., 2023b; OpenAI, 2023; Li et al., 2023a) have demonstrated significant progress in bridging vision and language, addressing tasks such as visual question answering (VQA), image captioning, and instruction adherence. However, a fundamental difficulty that persists is hallucination, where the generation of outputs that are inconsistent with or unsupported by visual inputs (Kalavasis et al., 2024). Hallucinations are especially common in open-ended VQA circumstances, in which unclear or underspecified questions can result in factual mistakes. These hallucinations span diverse categories, including *Object*, *Attribute*, *Relation*, and *Event*. Figure 1 illustrates that a singular image of "a dog grasping an orange frisbee" can elicit various forms of hallucination: a fictitious "cat" (object), an incorrectly identified "green" frisbee (attribute), an erroneous spatial relation "under the paw" (relation), or a temporal misrepresentation "sleeping" (event). These errors are semantically plausible yet visually unfounded, posing major challenges for trustworthiness and safety of MLLMs across critical applications, including medical decision-making (Kim et al., 2025), robotic perception (Park et al., 2023), and autonomous navigation (Alsulaimawi, 2025).



Is there a cat in the image?

Yes, there is a cat behind the dog.

What color is the frisbee?

The frisbee is green.

Is the frisbee under the dog' s paw?

Yes, one paw presses it down.

Is the dog sleeping on the grass?

Yes, the dog is lying still.

Figure 1: Illustration of four major hallucination types in open-ended VQA. Despite being visually grounded, MLLMs produce fluent but hallucinated responses across object, attribute, relation, and event dimensions.

Existing works (Sun et al., 2023; Ayala and Béchard, 2024; Sun et al., 2024) often regard hallucination as an output-level anomaly that is corrected post hoc, overlooking its underlying cause. In practice, hallucinations frequently arise from misalignment between the input, visual content, and the model's latent reasoning. Without an internal supervision mechanism, models may produce fluent yet ungrounded answers. We argue that hallucination stems from the model's inability to validate its own output across modalities and recommend injecting this ability directly into training.

We subsequently derive inspiration from human cognitive processes. When answering visual questions, individuals rarely rely on a single forward guess. Instead, after answering, they may reassess the question's intent, examine the visual scene, and refine conclusions—especially in the face of ambiguity or uncertainty. However, most models operate in a unidirectional manner, mapping $(Q, I \rightarrow A)$. As a result, once the model makes a prediction, there is no structured way to assess whether it actually understood the question, if the answer aligns with the visual evidence, or whether the model attended to the right regions in the image.

To address this issue, we propose **ReLoop**, a cognitively inspired unified training framework that encourages multimodal consistency for cross-modal understanding in MLLMs. ReLoop implements a feedback-driven closed-loop supervision process, allowing the model to reassess its predictions and validate their consistency with the original input through multi-level supervision during training. Specifically, after MLLMs produce an answer from the image-question pair, Reloop enables the model to recapture input semantics and assess internal consistency via: a Consistency Feedback Plugin (CFP), comprising two frozen modules: (1) CFP-Lang reconstructs the question $\hat{Q}^*$ from (A, I) to supervise semantic alignment, and (2) CFP-Vis generates a description I^* to assess factual grounding. In parallel, an attention supervision module extracts the model's token-to-image attention map $\mathcal{H}$ and compares it with an entropy-based pseudo-ground truth. All signals are integrated as differentiable losses in the overall optimization objective. This design encourages the model to "see twice and think backward"—first see to answer $(Q, I \rightarrow A)$, see twice to reassess $(A, I \rightarrow \hat{Q}^*, I^*, \mathcal{H})$, and finally to correct $(\hat{Q}^*, I^*, \mathcal{H} \approx Q, I, \mathcal{H}_{\text{pseudo}})$.

ReLoop bridges the gap between perception and output. It turns the black-box understanding process into an interpretable, feedback-aware loop that continuously refines the model's internal representations. Our key contributions can be summarized clearly as follows:

- We propose **ReLoop**, a cognitively inspired closed-loop training framework that ensures consistency among image, question, and answer modalities, effectively mitigating hallucinations in MLLMs.
- We introduce three complementary consistency signals: semantic reconstruction, visual description, and attention alignment, to emulate the humanlike "reversible thinking" process and improve cross-modal consistency during training.
- We provide a novel use of pretrained vision-language models by repositioning them as frozen Consistency Feedback Plugins (CFPs) in the training loop. Rather than functioning as typical forward-only encoders, they now perform in a reflective, backward supervisory role, producing feedback signals to guide the main model's alignment with multimodal semantics.

2 Related Work

Hallucination Mitigation in MLLMs. Multimodal LLMs frequently produce hallucinations—responses conflicting with visual inputs, such as inventing entities or misaligning semantics (Li et al., 2023b). Recent mitigation efforts combine post-hoc correction and architectural refinement. Retrieval-augmented methods like (Mala et al., 2025) grounds outputs in external knowledge via hybrid retrievers, while (Ayala and Béchard, 2024) reduces hallucinations in structured outputs. Architectural solutions such as OPERA (Huang et al., 2024) penalize over-trust during decoding, and preference-aligned training like TPO (Gu and Wang, 2025) enhances vision grounding. Post-generation frameworks, including Woodpecker (Yin et al., 2023), further improve factuality through structured verification.

Semantic Reversibility and Bidirectional Supervision. Human cognition leverages bidirectional reasoning to validate hypotheses—a principle termed "cognitive reversibility" (Johnson-Laird, 1983). Recent works explore this idea through decoding-time strategies: Self-RAG (Asai et al., 2023) integrates retrieval-augmented generation with self-reflection, enabling models to critique and refine their outputs iteratively, while DeepSeek-Math employs Group Relative Policy Optimization (GRPO) (Shao et al., 2024), enhancing mathematical reasoning by optimizing policy decisions based on group sampling strategies. Similarly, back-translation methods (Sennrich et al., 2016) enforce answer-question consistency through round-trip translation. Training-time solutions remain limited:

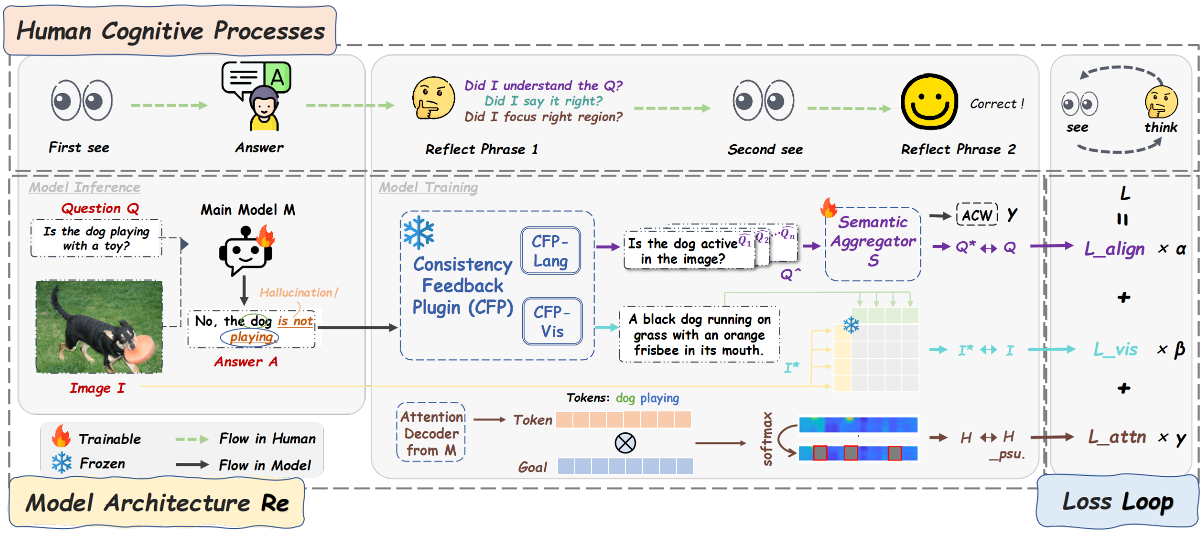


Figure 2: **Seeing Twice and Thinking Backwards: ReLooping Hallucination Suppression in Multimodal Language Models.** This diagram aligns human cognitive phases (left) with model modules (right) in a closed-loop process. The main model M produces an answer which is then introspected via CFP-Lang (language reconstruction), CFP-Vis (visual description), and internal cross-attention maps. Semantic aggregation, CLIP similarity, and entropy-based soft masks produce feedback losses that are summed and back-propagated to update M and the semantic aggregator S.

CycleConsistency (Pang and Wang, 2020) aligns forward-backward pathways via joint training but struggles with error accumulation in open-domain settings.

Cross-modal Consistency. Ensuring cross-modal consistency is vital for mitigating hallucinations in multimodal large language models (MLLMs). Recent methods enhance visual-text alignment to reduce semantic drift. Visual Contrastive Decoding (VCD) (Leng et al., 2024) contrasts outputs from original and perturbed images to promote grounding and reduce unimodal bias. Hallucination-Augmented Contrastive Learning (HACL) (Jiang et al., 2024) treats hallucinated captions as hard negatives to improve alignment. EAGLE (Villa et al., 2025) further refines visual encoders post-pretraining, yielding better grounding and fewer hallucinations.

3 Preliminaries

3.1 Task Formulation: Open-ended Visual Question Answering

We consider the task of open-ended VQA, where the model receives an image I and a natural language question Q, and produces a free-form answer A. Unlike multiple-choice settings, this task requires the model to produce linguistically coherent and visually grounded responses without predefined options.

In this case, hallucination refers to answers that contradict the image I, misinterpret the question Q, or introduce unsupported content.

3.2 Consistency Signals

To encourage faithful understanding, we supervise the model using three types of cross-modal consistency signals:

Linguistic Consistency. We verify whether the model's answer A implies the same question intent as the original Q, by attempting to reconstruct Q from (A, I). This tests whether the model understood the question meaningfully.

Visual Consistency. We evaluate whether the answer A is factually grounded in image I, by generating a descriptive caption I^* based on (A, I) and checking its alignment with the image. This ensures that the response reflects the actual visual content.

Attention Consistency. We examine whether the model attends to the correct regions of the image while producing A. This is assessed by comparing its internal attention map $\mathcal{H}$ with a soft pseudo-ground truth $\mathcal{H}_{\text{pseudo}}$ derived from entropy-based cues.

Together, these consistency signals serve as indirect evidence of whether the model truly grasps both the visual input and the question semantics.

4 ReLoop Framework: Reflect, Recapturing, and Optimize through a Closed-Loop Process

We introduce **ReLoop**, a unified training framework aimed at reducing hallucinations in MLLMs for open-ended VQA answering. As illustrated in Figure 2, the framework incorporates three complementary consistency feedback mechanisms: **semantic reconstruction**, **visual description**, and **attention alignment** to supervise the model toward producing answers faithful to both the question and the image.

These feedback signals are instantiated through a frozen **Consistency Feedback Plugin (CFP)**: semantic reconstruction (CFP-Lang) and visual description (CFP-Vis), and **attention supervision** from the model itself. The CFP module is broadly compatible with a range of encoder-decoder or decoder-only MLLMs. During inference (*First See* → *Answer*), the model receives a question-image pair and produces an initial answer. The training process then begins with *Reflect* → *Second See* → *Correct*: the model examines its output through structured consistency feedback. Specifically, it "introspectively" asks:

- *"Did I understand the Q?"* (→ semantic reconstruction)
- *"Did I say it right?"* (→ visual description)
- *"Did I focus the right region?"* (→ attention alignment)

ReLoop decomposes hallucination mitigation into two interacting components:

- ***"Re"*** emphasizes recapturing details, encouraging the model to reassess the semantic and visual cues from both question and image through CFP modules and token-level attention heatmaps.
- ***"Loop"*** denotes a feedback-driven training loop. After each forward prediction, feedback from the three consistency pathways is aggregated into the loss function (L_{align}, L_{vis}, L_{attn}), driving iterative updates that refine the model's multimodal grounding and answer reliability.

4.1 A Closed-loop Training

The entire training process follows a closed-loop pattern, emulating "seeing twice and thinking backward". Each training step proceeds as follows:

1. **First See:** The main model M takes the image I and question Q as input to produce an initial answer A.
2. **Reflect:** The model introspects on A by reconstructing a proxy question $\hat{Q}$, generating a visual description I^*, and extracting token-level attention $\mathcal{H}$.
3. **Second See:** The reconstructions are compared against the original inputs to compute consistency losses, capturing discrepancies in semantics, visual grounding, and attention focus.
4. **Correct:** All feedback signals are aggregated into L_{total} to update M and the semantic aggregator S via backpropagation.

This multi-stage loop is repeated across training epochs, leading to the model M that gradually reduces hallucinations.

4.2 Re: Recapturing Details for Consistency Supervision

This stage corresponds to the training-time processes of "Reflect" and "Second See", where the model reassesses its answers to recapture overlooked semantic and visual details. Three feedback pathways modules examine whether the model understood the question, correctly grounded its answer in the image, and attended to salient regions.

4.2.1 CFP-lang: Language Reconstruction and Adaptive Consistency Weighting

To evaluate whether the model correctly interprets the input question, we introduce a frozen language reconstruction module, CFP-lang. Given the answer-image pair (A, I), CFP-lang produces a set of candidate reverse questions $\{\hat{Q}_1, \hat{Q}_2, \dots, \hat{Q}_k\}$ that approximate possible intents underlying the predicted answer. A lightweight semantic aggregator S, composed of a BERT encoder and a single-layer MLP, scores each candidate against the original question Q using BERTScore. The highest-ranked proxy $\hat{Q}^*$ is selected to reflect the model's inferred intent.

However, directly enforcing alignment on all reconstructed questions may introduce noise, particularly when the produced answer is short or underspecified. To mitigate this, we introduce an Adaptive Consistency Weighting (ACW) mechanism, which adjusts the attention supervision (mentioned in section 4.2.3) strength based on the similarity between Q and $\hat{Q}^*$:

$$\gamma = \begin{cases} 1.0 & \text{if BERTScore}(Q, \hat{Q}^*) \geq 0.8 \\ 0.1 & \text{if } 0.6 \leq \text{BERTScore}(Q, \hat{Q}^*) < 0.8 \\ 0.01 & \text{if BERTScore}(Q, \hat{Q}^*) < 0.6 \end{cases} \tag{1}$$

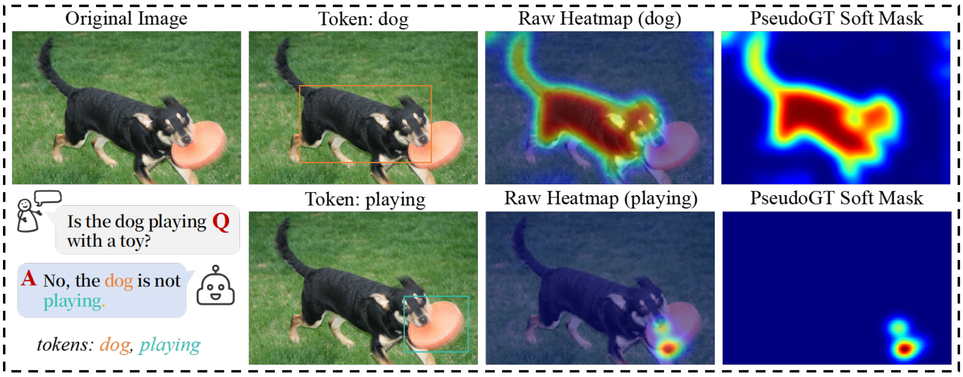


Figure 3: **Token-Level Attention Supervision.** Visualization of predicted attention $\mathcal{H}$ and entropy-based pseudo ground truth $\mathcal{H}_{\text{pseudo}}$ for two key answer tokens: *dog* (top row) and *playing* (bottom row).

Rather than discarding low-confidence pairs, this soft weighting ensures that stronger semantic matches contribute more prominently to the learning objective. The language consistency loss is computed as:

$$L_{\text{align}} = 1 - \text{BERTScore}(Q, \hat{Q}^*) \quad (2)$$

4.2.2 CFP-visual: Visual Description and Similarity Supervision

To validate whether the produced answer A is visually grounded in the image I, we employ a frozen visual description module, CFP-visual. Given (A, I), it generates a caption I^* describing the image content implied by the answer. We then compute the cosine similarity between the CLIP-encoded vectors of I and I^*, and derive the visual consistency loss as:

$$L_{\text{vis}} = 1 - \cos(\text{CLIP}_{\text{img}}(I), \text{CLIP}_{\text{text}}(I^*)) \quad (3)$$

4.2.3 Attention Supervision via Heatmap Consistency

To enhance interpretability and mitigate hallucinations arising from inattentive or unstable decoding, we explicitly supervise the model's token-level cross-attention patterns. From the decoder of the main model M, we extract attention maps $\mathcal{H}$, which indicate the spatial focus during answer generation. We construct a soft pseudo-ground-truth heatmap $\mathcal{H}_{\text{pseudo}}$ using entropy-based masking. This method preserves uncertainty information and avoids brittle hard labels. As illustrated in Figure 3, well-grounded tokens (e.g., *dog*) yield concentrated heatmaps aligned with visual evidence, while hallucinated tokens (e.g., *playing*) produce offset patterns. We enforce alignment between $\mathcal{H}$ and $\mathcal{H}_{\text{pseudo}}$ by minimizing the KL divergence:

$$L_{\text{attn}} = \text{KL}(\mathcal{H} \parallel \mathcal{H}_{\text{pseudo}}) \quad (4)$$

4.3 Loop: Feedback Aggregation, Alignment, and Optimization

After consistency signals are computed from language, vision, and attention supervision, ReLoop aggregates them into a unified training objective. This stage corresponds to the "Correction" step in the loop, where the model updates its parameters based on multi-perspective feedback. The total loss combines standard supervision with the three consistency terms:

$$L_{\text{total}} = L_{\text{sft}} + \alpha \cdot L_{\text{align}} + \beta \cdot L_{\text{vis}} + \gamma \cdot L_{\text{attn}} + \lambda \cdot \Omega(\theta) \quad (5)$$

where L_{sft} is the token-level cross-entropy loss, and $\Omega(\theta)$ is an L2 regularization term. The consistency weights are empirically set as $\alpha = 1.0$, $\beta = 0.7$, $\lambda = 10^{-5}$ and γ is defined in Equation 1.

Only the parameters of the main model M and the semantic aggregator S are updated during training. All feedback modules, including CFP-Lang, CFP-Vis, attention supervision, and CLIP, remain frozen.

5 Experimental Setup

Training Data. We curate 30K high-quality $\{I, Q, A\}$ from LLaVA-Instruct-150K. To simulate hallucination supervision, we generate contrastive examples by perturbing key semantics (e.g., *objects, attributes, relations, event*), enabling fine-grained control over hallucination types. Details are in Appendix A.1 A.2.

Evaluation Benchmarks and Metrics. We evaluate ReLoop on a broad range of hallucination and multimodal understanding benchmarks, including POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023), MMHal-B (Sun et al., 2023), HallusionBench (Guan

Type	Module	Signal Type	Baseline	ReLoop	ΔMean	Baseline Hallu.	ReLoop Hallu.	ΔRate
Object	Visual	CLIP(I, I^*)	28.02 ± 3.10	29.46 ± 3.27	↑1.44	24.5%	10.3%	↓14.2%
	Language	BERT$(Q, \hat{Q})$	0.862 ± 0.022	0.873 ± 0.024	↑0.011			
	Attention	Entropy$(\mathcal{H})$	1.31 ± 0.40	1.28 ± 0.45	↓0.03			
Attribute	Visual	CLIP(I, I^*)	26.59 ± 3.31	26.81 ± 3.41	↑0.22	7.3%	4.0%	↓3.3%
	Language	BERT$(Q, \hat{Q})$	0.868 ± 0.025	0.894 ± 0.028	↑0.026			
	Attention	Entropy$(\mathcal{H})$	1.36 ± 0.46	1.32 ± 0.52	↓0.04			
Relation	Visual	CLIP(I, I^*)	27.22 ± 3.26	28.01 ± 3.38	↑0.79	13.2%	7.6%	↓5.6%
	Language	BERT$(Q, \hat{Q})$	0.855 ± 0.020	0.875 ± 0.023	↑0.020			
	Attention	Entropy$(\mathcal{H})$	1.39 ± 0.43	1.34 ± 0.50	↓0.05			
Event	Visual	CLIP(I, I^*)	26.63 ± 3.08	26.94 ± 3.37	↑0.31	10.4%	5.2%	↓5.2%
	Language	BERT$(Q, \hat{Q})$	0.861 ± 0.024	0.877 ± 0.029	↑0.016			
	Attention	Entropy$(\mathcal{H})$	1.33 ± 0.42	1.51 ± 0.55	↑0.18			

Table 1: Effect of ReLoop on consistency and hallucination reduction across different hallucination types. We compare MiniGPT-4 (baseline) and ReLoop in terms of signal outputs from three frozen feedback modules: visual grounding (CLIP similarity), semantic alignment (BERTScore), and attention focus (entropy). Δ denotes the absolute change in signal quality after applying ReLoop.

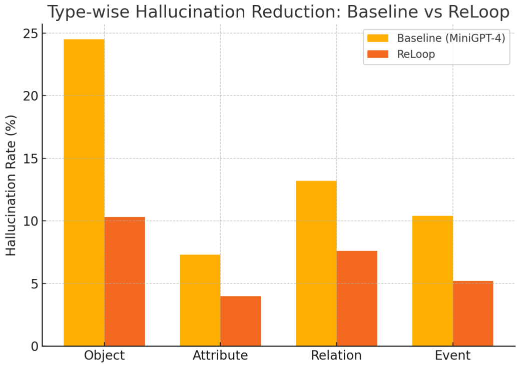


Figure 4: Type-wise hallucination rates (%) for baseline (MiniGPT-4) and ReLoop models.

et al., 2024), Faith/FaithS (Jing et al., 2024), and MME (Fu et al., 2023). Full definitions are in Appendix A.3 A.4.

Baselines. We use MiniGPT-4 as the baseline model in Experiment 6.1 and compare against LLaVA-1.5 variants trained with LLaVA-RLHF (Sun et al., 2023), HA-DPO (Zhao et al., 2023) , and POVID (Zhou et al., 2024). All baselines share the same backbone and training setup for a fair comparison. Implementation details are provided in Appendix A.5.

6 Results and Analysis

6.1 Identify Internal Causes of Hallucinations: Module Signals vs. Hallucination States

We first aim to pinpoint internal representation deficiencies that drive hallucination behaviors across different hallucination types. We analyze consistency signal deviations produced by ReLoop's frozen supervision modules, with hallucinated versus non-hallucinated samples. Responding: *"Did I understand the question?"* (language, via BERTScore); *"Did I say it right?"* (visual, via CLIP similarity); *"Did I focus the right region?"* (attention, via entropy).

Multimodal hallucinations stem from structured, modality-specific representation gaps. As shown in Table 1, hallucinated responses are consistently associated with lower CLIP similarity (–2.25), reduced BERTScore (–0.034), and higher attention entropy (+0.31). Figure 5 reveals distinct signal patterns associated with different hallucination types. Object hallucinations correspond to a clear leftward shift in CLIP similarity, indicating weaker visual grounding. Attribute hallucinations are marked by lower BERTScore, reflecting reduced semantic alignment. Event hallucinations show higher attention entropy, suggesting that the model distributes focus more broadly, which may help in capturing complex scenes but also increases the risk of focusing on irrelevant regions.

Signal dynamics vary by hallucination type. (1) *Object hallucinations* are primarily rooted in the visual module. They often manifest as hallucinated entities not present in the image. ReLoop yields a significant gain in CLIP similarity (↑1.44) and a decrease in attention entropy (↓0.03), suggesting enhanced image-text alignment and focused visual

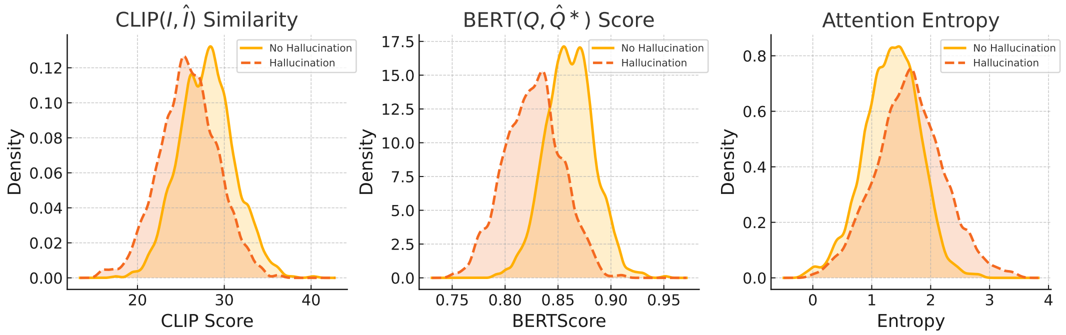


Figure 5: KDE distributions of CLIP similarity, BERTScore, and attention entropy for hallucinated and non-hallucinated samples. ReLoop's frozen modules exhibit sharp signal shifts that serve as reliable supervision sources.

Model	Hallucination Suppression			Cross-modal Faithfulness		
	POPE↑	CHAIR_s ↓	CHAIR_i ↓	F1↑	Faith↑	FaithS↑
MiniGPT-4	82.3	49.0	22.7	63.2	86.7	68.5
+ ReLoop	**83.9**	**38.8**	**20.5**	**69.9**	**88.6**	**71.3**
InstructBLIP	83.8	47.8	20.6	**68.4**	87.3	69.8
+ ReLoop	**85.3**	**36.9**	**17.5**	67.0	**88.5**	**73.2**
LLaVA	85.7	53.5	24.2	65.8	**89.5**	**75.8**
+ ReLoop	**86.3**	**40.2**	**16.2**	**70.3**	89.2	75.3
mPLUG-owl	89.1	62.5	31.0	58.9	**88.3**	**72.7**
+ ReLoop	**90.9**	**42.5**	**21.8**	**66.5**	87.9	71.0
ShareGPT4V	88.2	50.2	21.8	68.0	88.2	73.6
+ ReLoop	**89.7**	**44.9**	**21.5**	**69.2**	**89.3**	**74.8**

Table 2: Performance comparison of various LVLMs with and without ReLoop. Hallucination is measured by POPE, CHAIR_s, and CHAIR_i, cross-modal faithfulness is evaluated using F1, Faith, and FaithS.↓ indicates lower is better; ↑ indicates higher is better.

grounding. (2) *Attribute hallucinations* show the largest improvement in BERTScore (↑0.026) and only a slight change in CLIP similarity (↑0.22), indicating that semantic reconstruction plays a more important role than visual grounding. This aligns with their nature: attributes often relate to textual misinterpretation (e.g., *color or size*), even when visual cues are present. (3) *Relation hallucinations* involve complex spatial or relational semantics and display moderate improvements across all three signals (CLIP↑0.79, BERT↑0.020, Entropy↓0.05), suggesting that ReLoop's multi-signal supervision addresses cross-modal misalignment collaboratively. (4) *Event hallucinations* are primarily tied to attention misallocation. ReLoop improves CLIP (↑0.31) and BERT (↑0.016) slightly, but entropy increases (↑0.18), reflecting broader attention scopes. This likely helps avoid fixation on irrelevant regions, especially in dynamic or temporally inferred scenes. Figure 4 shows that ReLoop successfully mitigates hallucinations compared to MiniGPT-4 across four hallucination types.

6.2 Effects of Structured Feedback in ReLoop

Motivated by earlier findings, we evaluate how effectively ReLoop's structured feedback enhances semantic grounding across five representative LVLMs (Table 2). The observed improvements span models with diverse architectures and training paradigms, showing that ReLoop is broadly compatible and easily integrable into various LVLMs.

Hallucination Suppression. ReLoop significantly reduces references to non-existent entities. MiniGPT-4's CHAIR_s/CHAIR_i drop by 20.8%/9.7% (POPE +1.9%), InstructBLIP's by 22.7%/15.0%. LLaVA also benefits, showing a 24.9% and 33.1% drop in the same metrics. These reductions confirm that ReLoop effectively mitigates object-level and spatial hallucinations by enhancing visual consistency.

Cross-modal Faithfulness. ReLoop also enhances cross-modal faithfulness. For MiniGPT-4 and LLaVA, F1 scores improve by 10.6% and 6.8%, respectively, while Faith/FaithS rising by 4.1%, 4.9%, and 3.4% in MiniGPT-4, InstructBLIP, and

Ablation Version	Hallucination Suppression			Cross-modal Faithfulness		
	POPE↑	CHAIR$_s$ ↓	CHAIR$_i$ ↓	F1↑	Faith↑	FaithS↑
MiniGPT-4	83.0	49.0	22.7	60.2	84.3	64.2
w/o Consistency Supervision	84.2	47.4	21.6	60.7	86.7	68.5
w/o Gating & Aggregator	**85.4**	39.8	19.7	60.4	88.1	71.6
w/o Attention Supervision	83.6	40.2	20.1	61.9	86.3	67.5
Full ReLoop	84.9	**38.3**	**18.9**	**63.1**	**88.6**	**72.8**

Table 3: Performance comparison of ReLoop under different ablation configurations on MiniGPT-4. Removing consistency supervision results in the worst faithfulness and hallucination rate, while full ReLoop delivers the best overall performance. Although gating removal slightly improves POPE, it hurts precision (F1) and consistency.

Method	Hallucination Suppression			Cross-modal Faithfulness		
	POPE↑	CHAIR$_s$ ↓	CHAIR$_i$ ↓	F1↑	Faith↑	FaithS↑
LLaVA-1.5	83.5	53.9	23.5	63.2	86.9	70.5
+ LLaVA-RLHF	**88.2**	44.5	20.1	67.0	89.0	74.4
+ HA-DPO	86.7	52.3	21.6	65.4	88.4	73.5
+ POVID	84.3	53.2	24.2	64.7	87.3	71.8
+ ReLoop	87.9	**42.0**	**19.5**	**67.4**	**89.5**	**75.1**

Table 4: Performance comparison of ReLoop with various alignment-enhancing baselines for LLaVA-1.5 on metrics measuring hallucination suppression and cross-modal faithfulness. Best scores are in bold and the second are underlined.

ShareGPT4V. These gains suggest that the model not only grounds responses more accurately in the image but also maintains semantic alignment with the question intent.

6.3 Ablation Study

To assess the contribution of each component in ReLoop, we perform a coarse-grained ablation study over four configurations (Table 3). Removing consistency supervision leads to the highest hallucination rates (CHAIR$_s$: 47.4) and lowest semantic faithfulness (FaithS: 68.5), highlighting its central role. Attention supervision also proves important, as its removal moderately reduces FaithS. While removing gating slightly improves POPE, it harms F1 and hallucination suppression. Full ReLoop achieves the best overall results, reducing CHAIR$_s$ by 10.7 and increasing FaithS by 8.6 over the baseline. These findings underscore the complementary roles of all modules and the importance of structured feedback for robust alignment.

6.4 Unified Comparison with Alignment Strategies

We compare ReLoop with representative alignment methods, LLaVA-RLHF, HA-DPO, and POVID on both fine-grained hallucination metrics and broader benchmark evaluations. As shown in Table 4, ReLoop consistently outperforms alternatives on POPE, CHAIR, F1, and faithfulness metrics, indicating stronger hallucination suppression and cross-modal faithfulness. On benchmark-level evaluations (Table 5), ReLoop leads on AMBER, MMHal-B, and HallusionBench, while remaining competitive on MME. The slight MME drop may reflect a common trade-off between alignment supervision and low-level perception, also observed in other alignment-based methods like LLaVA-RLHF. These findings underscore ReLoop's effectiveness across both targeted and comprehensive settings.

Method	AMBER↑	MME↑	MMHal-B↑	Hallu-B↑
LLaVA-1.5	73.9	**1513**	65.4	48.6
+ LLaVA-RLHF	73.8	1231	64.3	43.2
+ HA-DPO	77.2	1374	65.6	49.9
+ POVID	75.8	1421	65.9	51.4
+ ReLoop	**80.3**	1505	**68.9**	**52.3**

Table 5: Benchmark-level comparison of ReLoop with alignment strategies across four evaluation baselines.

7 Conclusion

We present **ReLoop**, a closed-loop training framework that mitigates hallucinations in MLLMs by enforcing semantic and visual consistency through bidirectional feedback. By incorporating language reconstruction, visual description, and attention alignment, ReLoop allows models to verify and refine predictions during training. Experiments show consistent gains in hallucination suppression and interpretability, establishing ReLoop as a general foundation for building more reliable MLLMs.

Potential Limitations

Performance Variability Across Hallucination Types. While ReLoop substantially improves hallucination suppression in object and attribute categories, its effectiveness on relation and event hallucinations remains relatively modest. These hallucination types often involve higher-order reasoning and temporal or spatial understanding, which are less easily corrected through current consistency signals. Future extensions may incorporate specialized supervision tailored to relational semantics or causal cues to address this gap.

Supervision Dependency and Domain Adaptability. ReLoop relies on access to paired image–question–answer data to compute consistency signals. This requirement poses challenges in domains with limited high-quality supervision, such as medical or scientific imaging. Moreover, the training framework assumes reasonably clean and grounded reference answers, which may not hold in low-resource or noisy environments. Reducing ReLoop's dependence on strongly supervised inputs and exploring semi-supervised or synthetic feedback generation remain promising directions for broader applicability.

Reliance on Pretrained Vision–Language Modules. The effectiveness of ReLoop hinges on auxiliary modules such as CLIP and BLIP-2 to produce semantic feedback. These pretrained models must offer reasonably accurate visual-textual alignment; otherwise, the resulting supervision may be noisy or misleading. This dependence limits ReLoop's deployment in specialized or domain-shifted settings where existing pretrained modules underperform. Enhancing the adaptability or self-calibration of feedback modules could help mitigate this limitation.

Ethics Statement

All datasets utilized in this work are either publicly released or ethically sourced, ensuring full compliance with associated data usage policies. For evaluation purposes, we additionally include AI-generated content produced under controlled prompting conditions. These samples are clearly labeled and subjected to careful human verification to ensure factual accuracy and annotation quality. We acknowledge the broader implications of hallucination mitigation in AI systems and advocate for responsible model development that prioritizes reliability, fairness, and interpretability.

References

Zahir Alsulaimawi. 2025. Feedback-enhanced hallucination-resistant vision-language model for real-time scene understanding. *arXiv preprint arXiv:2504.04772*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Orlando Marquez Ayala and Patrice Béchard. 2024. Reducing hallucination in structured outputs via retrieval-augmented generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 228–238. Association for Computational Linguistics.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Jihao Gu and Yingyao Wang. 2025. Token preference optimization with self-calibrated visual-anchored rewards for hallucination mitigation. *arXiv preprint arXiv:2412.14487*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14375–14385. IEEE.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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 (CVPR)*. Highlight Paper.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27036–27045. Code available at `https://github.com/X-PLUG/mPLUG-HalOwl/tree/main/hacl`.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguis-*

tics: EMNLP 2024, pages 5042–5063. Association for Computational Linguistics.

Philip N. Johnson-Laird. 1983. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Harvard University Press, Cambridge, MA.

Alkis Kalavasis, Anay Mehrotra, and Grigoris Velegkas. 2024. On the limits of language generation: Trade-offs between hallucination and mode collapse. *arXiv preprint arXiv:2411.09642*.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Lizhou Fan, Eugene Park, Tristan Lin, Joonsik Yoon, Wonjin Yoon, Maarten Sap, Yulia Tsvetkov, Paul Liang, Xuhai Xu, and 6 others. 2025. Medical hallucinations in foundation models and their impact on healthcare. *arXiv preprint arXiv:2503.05777*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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 (CVPR)*, pages 1316–1325. Poster Highlight.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Chandana Sree Mala, Gizem Gezici, and Fosca Giannotti. 2025. Hybrid retrieval for hallucination mitigation in large language models: A comparative analysis. *arXiv preprint arXiv:2504.05324*.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11831–11838.

Jin-Soo Park, Xuesu Xiao, Garrett Warnell, Harel Yedidsion, and Peter Stone. 2023. Learning perceptual hallucination for multi-robot navigation in narrow hallways. In *Proceedings of the 2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, England.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 371–376, Berlin, Germany. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Zhongxiang Sun, Xiaoxue Zang, Kai Zheng, Yang Song, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. 2024. Redeep: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*.

Andrés Villa, Juan León Alcázar, Motasem Alfarra, Vladimir Araujo, Alvaro Soto, and Bernard Ghanem. 2025. Eagle: Enhanced visual grounding minimizes hallucinations in instructional multimodal models. *arXiv preprint arXiv:2501.02699*.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*.

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

A Additional Experimental Details

A.1 Implementation Details

Backbone and Setup. We apply ReLoop to five representative LVLMs with diverse architectures: MiniGPT-4, InstructBLIP, LLaVA-1.5, mPLUG-owl, and ShareGPT4V. Importantly, we do not alter the internal structures of these models. ReLoop is introduced as a lightweight, external consistency-supervision framework during training. All backbones are initialized with their public checkpoints and keep their visual encoders (e.g., *ViT, CLIP*) frozen.

ReLoop Components. ReLoop introduces three frozen feedback modules: (1) CFP-Lang (MiniGPT-4-based reverse question reconstructor); (2) CFP-Vis (BLIP-2-based visual describer); (3) Attention Supervision that aligns decoder attention maps with entropy-based soft pseudo-labels. A frozen BERT encoder plus an MLP scorer serves as a lightweight semantic aggregator. All feedback modules remain frozen; only the backbone and the aggregator are updated.

Training Details. Experiments are performed on 8×A100 GPUs (80GB) using mixed-precision training (fp16) for 3 epochs. We adopt the AdamW optimizer with parameters $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 0.05. The effective batch size is 128, with a gradient accumulation step of 8. The initial learning rate is set to 5×10^{-5}, along with 1,000 warm-up steps and cosine learning rate decay scheduling.

Loss Function. The overall objective is

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sft}} + \alpha\, \mathcal{L}_{\text{align}} + \beta\, \mathcal{L}_{\text{vis}} + \gamma\, \mathcal{L}_{\text{attn}} + \lambda\, \Omega(\theta), \tag{6}$$

We set the hyper-parameters as $\alpha = 1.0$, $\beta = 0.7$, and $\lambda = 10^{-5}$. The weight γ is dynamically adjusted by the Adaptive Consistency Weighting (ACW) mechanism, which modulates γ based on the BERTScore between the original and reconstructed questions (see Section 4.2.1).

A.2 Training Dataset Construction

We curated approximately 30K high-quality QA-image triplets from the LLaVA-Instruct-150K corpus (Liu et al., 2023a), each containing an image, an open-ended question, and a human-annotated answer. To simulate hallucination supervision, we generated semantically contradictory answers by modifying key elements (e.g., *objects, attributes, or relations*) in the references. These hallucinated samples were automatically constructed and manually verified for quality and type diversity. In Experiment 6.1, we selected 500 representative QA-image pairs from the filtered validation set based on POPE and MMHalBench, equally split between hallucinated and non-hallucinated cases. In Experiment 6.2, we evaluated five LVLMs on this curated set to assess the impact of ReLoop. Models with open alignment architectures (e.g., *MiniGPT-4, InstructBLIP*) showed the greatest improvement, while high-performing black-box models (e.g., *ShareGPT4V*) saw minimal gains, suggesting ReLoop's effectiveness hinges on alignment signal compatibility.

A.3 Evaluation Metrics

To comprehensively evaluate the effectiveness of ReLoop in mitigating hallucinations and enhancing visual grounding, we adopt a structured set of metrics covering both hallucination suppression and cross-modal consistency. In particular, shown in Table 4, we group the metrics into two key categories: *Hallucination Suppression*, which quantifies the presence of non-existent or spurious content, and *Cross-modal Faithfulness*, which assesses the semantic and perceptual alignment between generated text and visual input.

A.3.1 Metrics on Hallucination Suppression

For hallucination evaluation, we incorporate CHAIR (Rohrbach et al., 2018) to measure hallucination frequencies at instance levels and include POPE (Li et al., 2023b), a probing-based diagnostic benchmark to evaluate object hallucinations through direct VQA-style interactions. Together, these metrics allow us to holistically assess ReLoop's ability to suppress hallucinated content while preserving descriptive quality.

- **CHAIR** (Rohrbach et al., 2018) (Caption Hallucination Assessment with Image Relevance) quantifies hallucinations by detecting whether the model-generated captions mention objects that do not exist in the image. It provides two variants:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects}\}|}, \tag{7}$$

$$\text{CHAIR}_S = \frac{|\{\text{hallucinated responses}\}|}{|\{\text{all responses}\}|}, \tag{8}$$

where CHAIR_I measures instance-level hallucination (object granularity) and CHAIR_S measures sentence-level hallucination (response granularity).

- **POPE** (Li et al., 2023b) (Polling-based Object Probing Evaluation) automates hallucination detection via instance-level object probing. It:
  - Segments objects in the image;
  - Asks the model about object existence and introduces distractor queries;
  - Computes metrics such as F1 score to measure detection precision.

  POPE offers direct insights into a model's visual grounding capability through objective visual questioning.

A.3.2 Metrics on Cross-modal Faithfulness

On the side of Cross-modal Faithfulness, we adopt Faith and Faith_S (Jing et al., 2024), which evaluate how well the generated text is grounded in the visual input. Faith focuses on overall alignment, while Faith_S specifically checks whether statements are supported by the visual evidence in a token-level or segment-wise manner. In addition, we report the $F1$ score, a standard metric that captures the harmonic mean of precision and recall between the predicted and reference entities. In our context, it reflects how well the model identifies relevant visual content without fabricating or omitting essential elements, thus serving as a practical indicator of the model's grounding precision and completeness.

- **F1 Score** reflects the harmonic mean of precision and recall in detecting whether queried objects exist. High F1 indicates accurate recognition and rejection of hallucinated entities:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}$$

- **Faith** (Jing et al., 2024) measures the overall semantic alignment between image and response. It uses automated matching or human verification to assess whether the content is factually grounded in the image:

$$\text{Faith} = \frac{|\text{Aligned Statements}|}{|\text{Total Statements}|} \tag{10}$$

- **Faith_S** (Jing et al., 2024) extends Faith to a finer granularity by evaluating the support of specific sentence segments or tokens using cross-modal supervision or saliency alignment:

$$\text{Faith}_S = \frac{|\text{Grounded Segments or Tokens}|}{|\text{Total Segments or Tokens}|} \tag{11}$$

A.4 Evaluation Benchmark

Besides, to provide a fine-grained and multi-perspective assessment of ReLoop's effectiveness in suppressing hallucinations and enhancing cross-modal faithfulness, we adopt four complementary benchmarks. AMBER (Wang et al., 2023) targets object-level hallucinations, while MMHal-B (Sun et al., 2023) and HallusionBench (Guan et al., 2024) assess errors in attributes, spatial relations, and perceptual consistency. MME (Fu et al., 2023) covers general multimodal capabilities such as OCR and counting. These benchmarks collectively evaluate generative and discriminative capabilities, entity grounding, perceptual consistency, and multimodal reasoning:

- **AMBER** (Wang et al., 2023): An LLM-free multi-dimensional benchmark that diagnoses hallucinations in both generative and discriminative tasks. It explicitly tests object *existence*, *attributes*, and *relations*, allowing us to assess ReLoop's object-level grounding fidelity, attribute correctness, and relational accuracy. This supports the evaluation of semantic precision in visual grounding.

- **MMHal-B** (Sun et al., 2023): A benchmark built upon fact-augmented reinforcement learning (RLHF) that penalizes hallucinated attributes and spatial configurations. MMHal-B offers targeted diagnostics for hallucination suppression in factual and compositional dimensions, particularly assessing whether ReLoop can resist overgeneralization and maintain factual grounding under complex prompts.

- **HallusionBench** (Guan et al., 2024): A benchmark that probes visual-linguistic robustness under ambiguous image-text settings. It emphasizes contextual grounding, requiring models to handle subtle visual cues and nuanced linguistic traps. HallusionBench evaluates ReLoop's ability to maintain perceptual consistency and reject misleading contextual cues that typically trigger hallucinations.

- **MME** (Fu et al., 2023): A broad-spectrum benchmark measuring multimodal perception and cognition across 14 sub-tasks, including OCR, object counting, spatial reasoning, and commonsense grounding. MME validates whether ReLoop's structured supervision translates into generalized improvements in visual understanding and multimodal reasoning, beyond hallucination mitigation.

Together, these benchmarks offer layered supervision signals from fine-grained object hallucination detection to holistic multimodal cognition, providing strong empirical evidence of ReLoop's reliability across diverse real-world tasks.

A.5 Baseline Implementation

To evaluate ReLoop's generalizability and additive benefit, we compare it with three representative alignment-based hallucination mitigation strategies: LLaVA-RLHF (Sun et al., 2023), HA-DPO (Zhao et al., 2023) , and POVID (Zhou et al., 2024). These baselines span a diverse range of supervision paradigms, from reinforcement learning to contrastive grounding. Importantly, all methods are applied on top of the same backbone (LLaVA-1.5) with consistent training configurations, ensuring fair comparison.

- **LLaVA-RLHF** (Sun et al., 2023) aligns responses to human preferences through reinforcement learning from human feedback. While effective for improving general fluency and tone, it does not explicitly penalize visual or factual inconsistencies.

- **HA-DPO** (Zhao et al., 2023) adopts hallucination-aware preference optimization by contrasting faithful versus hallucinated generations. This method introduces targeted loss signals during fine-tuning, encouraging the model to avoid semantically spurious content.

- **POVID** (Zhou et al., 2024) enhances visual grounding via perturbed image inputs, injecting contrastive visual signals to reduce reliance on textual priors and promote visual fidelity.

Results from both fine-grained hallucination metrics (Table 4) and benchmark-level evaluations (Table 5) demonstrate that ReLoop consistently outperforms all competing methods. These results validate ReLoop as a robust and generalizable framework capable of enhancing multimodal model performance beyond what is achievable by current alignment-based techniques alone.

B Case Study

We present a qualitative case study to analyze how ReLoop mitigates hallucination across four representative types:

- **Object Hallucination**: The baseline model incorrectly asserts the presence of a "referee stand" which is not in the image. ReLoop corrects this by recognizing the absence of such an entity.

- **Attribute Hallucination**: An animal is mislabeled as "dog" instead of "chihuahua." ReLoop identifies the finer-grained attribute correctly.

- **Relation Hallucination**: The spatial relationship "on a sofa" is incorrectly predicted; ReLoop grounds the child's location more accurately.

- **Event Hallucination**: The action "not playing" contradicts visual evidence; ReLoop revises the answer to match the depicted motion.

As shown in Figure 6, baseline models such as MiniGPT-4 frequently produce fluent yet inaccurate answers that are not grounded in the image. ReLoop corrects these errors by leveraging consistency feedback to align its answers with both the question intent and visual content. The examples highlight ReLoop's capacity to suppress diverse hallucination patterns and improve factual reliability in open-ended VQA.

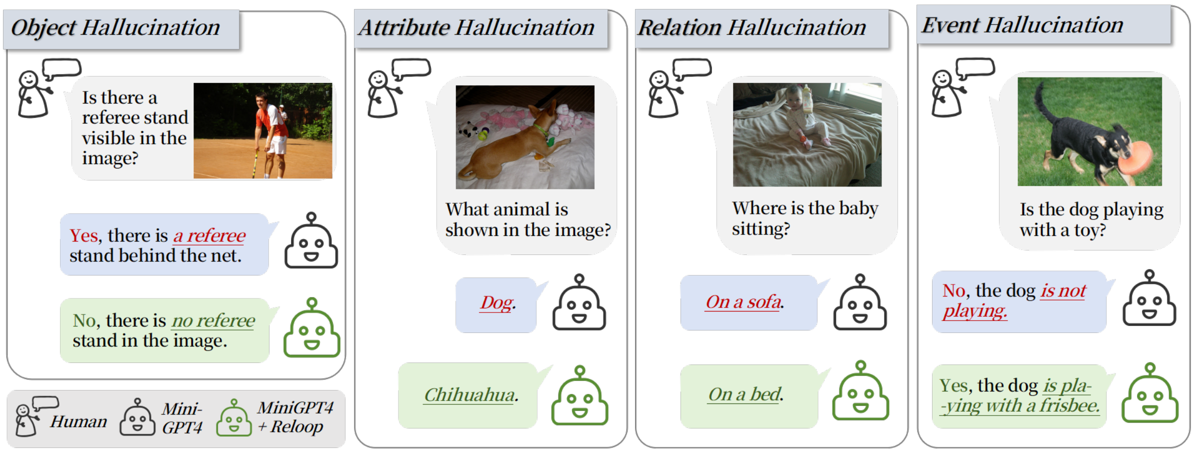

Object Hallucination
Is there a referee stand visible in the image?
Yes, there is a referee stand behind the net.
No, there is no referee stand in the image.
Human
Mini-GPT4
MiniGPT4 + Reloop
Attribute Hallucination
What animal is shown in the image?
Dog.
Chihuahua.
Relation Hallucination
Where is the baby sitting?
On a sofa.
On a bed.
Event Hallucination
Is the dog playing with a toy?
No, the dog is not playing.
Yes, the dog is pla- -ying with a frisbee.

Figure 6: **Case Study:** Comparison between MiniGPT-4 and ReLoop across four types of hallucination in open-ended VQA: *Object*, *Attribute*, *Relation*, and *Event*. ReLoop produces more accurate and grounded responses by aligning its outputs with both the visual evidence and the question semantics.