CogRef: Cognitive Reflection for Hallucination Mitigation in Multimodal Large Language Models

Anonymous authors
Paper under double-blind review

Abstract

Hallucination remains a major challenge for multimodal large language models (MLLMs). Existing mitigation methods often modify the generation process of MLLMs or rely on external verification tools, potentially constraining the original capabilities of the base MLLM. Motivated by human cognitive learning, where past experiences are reflected upon and reused to guide subsequent behavior, we propose CogRef, a cognitive reflection framework that decouples multimodal generation from hallucination supervision. A frozen MLLM serves as the cognitive core, while a learned supervisory circuit evaluates its response and provides structured feedback for final generation. To train the circuit, we transform past response cases into reusable multimodal experience, and internalize it through Concordance-Calibrated On-Policy Self-Distillation (CC-OPSD). CC-OPSD coordinates experience- and evidence-privileged teachers, enabling the supervisory circuit to internalize both response-handling experience and fine-grained visual grounding capability. At inference, CogRef requires neither privileged information nor external verification tools. Experiments across diverse MLLM families and multimodal benchmarks demonstrate that CogRef consistently mitigates hallucinations while preserving general multimodal capability, with the largest improvement reaching $\mathbf{45.32}$ points (MHaluBench: $\mathbf{31.45}\% \rightarrow \mathbf{76.77}\%$). Moreover, the supervisory circuit generalizes across different MLLMs without model-specific retraining, while selectively correcting hallucinated content and preserving visually grounded responses.

1 Introduction

Recent multimodal large language models (MLLMs) have demonstrated increasingly strong capabilities in visual understanding and multimodal reasoning (Tian et al., 2026; Team et al., 2026; Google DeepMind, 2026). However, their generations often introduce statements that are unsupported by or inconsistent with the input image. This phenomenon, commonly referred to as multimodal hallucination, poses a fundamental challenge to the reliability of MLLMs (Li et al., 2023; Zhou et al., 2024). Existing methods mitigate multimodal hallucination mainly from three directions. Training-based methods can internalize hallucination mitigation into the model parameters, but updating the base MLLM may alter its pretrained knowledge and general generation ability (Zhang et al., 2024; Wu et al., 2025; Wada et al., 2026). Decoding-based methods avoid full model training, yet their interventions can be unstable across generation steps (Im et al., 2026; Leng et al., 2024). Post-hoc methods can preserve the base model, but their correction capability often depends on external visual tools or complex verification pipelines at inference (Yin et al., 2024; Jeong et al., 2026). These trade-offs motivate a key question: *can hallucination be mitigated without relying on external visual tools or modifying the pretrained knowledge of the base MLLM?*

A natural inspiration comes from human learning, where cognition adapts by integrating knowledge acquired from past experience to guide future behavior (Sharpe et al., 2021; Sun et al., 2023). Building on this experience-driven perspective, we propose CogRef, a *Cog*nitive *Ref*lection framework for multimodal hallucination mitigation, which internalizes prior hallucination-related experience into an intrinsic supervisory capability, enabling the model to regulate unreliable generations without relying on external tools. CogRef consists of a *cognitive core* and a learned *supervisory cir-*

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

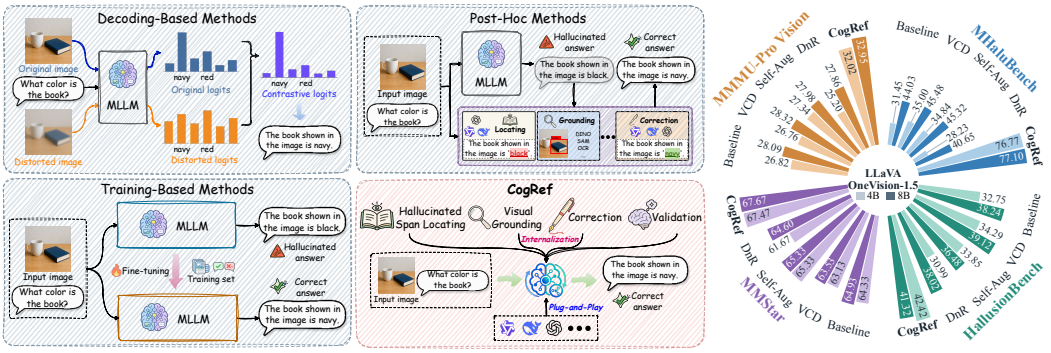


Figure 1: **Left:** Conceptual comparison of hallucination mitigation paradigms. Unlike decoding-, post-hoc-, and training-based methods, CoGREF internalizes hallucination-aware grounding and correction capabilities, while preserving the capabilities of the base MLLM. **Right:** Performance comparison of different methods on LLaVA-OneVision-1.5. CoGREF achieves consistently stronger performance across the evaluated benchmarks.

cuit. Given a multimodal input, the cognitive core first generates a candidate response using the original capabilities of the base MLLM. The supervisory circuit evaluates the candidate response and produces structured feedback for the cognitive core. The cognitive core then uses this feedback to generate the final response. In this way, the pretrained MLLM retains its original capability, while hallucination monitoring and correction are handled through a separate feedback pathway. To equip the supervisory circuit with reliable hallucination-aware feedback, we propose *Concordance-Calibrated OPSD (CC-OPSD)*, which internalizes reliable visual grounding and selective response-handling capabilities from reusable multimodal experience. As shown in Figure 1, CoGREF ① preserves the pretrained knowledge and generation ability of the base MLLM by keeping the cognitive core frozen and ② internalizes model-specific correction experience into a learned supervisory circuit without relying on external verification tools at inference. Experimental results show that CoGREF delivers substantial gains across both hallucination and general multimodal benchmarks. On hallucination evaluation, the largest improvements reach $\mathbf{31.45}\% \rightarrow \mathbf{76.77}\%$ on MHaluBench and $\mathbf{32.75}\% \rightarrow \mathbf{42.42}\%$ on HallusionBench. These gains extend to general multimodal evaluation, with the largest improvements reaching $\mathbf{53.00}\% \rightarrow \mathbf{64.87}\%$ on MMStar and $\mathbf{16.40}\% \rightarrow \mathbf{25.49}\%$ on MMMU-Pro Vision. **Our main contributions are summarized as follows:**

❶ We propose CoGREF, a cognitive feedback framework inspired by human cognitive reflection, which decouples hallucination supervision from multimodal task execution by keeping the cognitive core frozen and learning a dedicated supervisory circuit.

❷ We construct reusable multimodal experience that captures both correction and preservation behavior and internalize it into the supervisory circuit, requiring no external verification tools or privileged information at inference.

❸ We propose CC-OPSD to reliably internalize experience under heterogeneous privileged supervision, using complementary experience and visual evidence while preventing local conflicts from shifting the supervision target.

❹ Experiments across multiple multimodal benchmarks and MLLM families show that CoGREF consistently mitigates hallucinations while preserving grounded responses and transfers across different MLLMs without model-specific retraining.

2 Related Work

2.1 Multimodal Hallucination Mitigation

Existing multimodal hallucination mitigation methods mainly intervene during training, decoding, or post-generation processing. Training-based approaches improve hallucination awareness through additional supervision. REVERIE introduces reflective instruction tuning with rationales for visual

grounding (Zhang et al., 2024), while REVERSE further incorporates retrospective self-verification to detect and revise hallucinated content (Wu et al., 2025). Decoding-based methods modify token generation without updating the base model. VCD contrasts predictions from original and distorted visual inputs to suppress language priors (Leng et al., 2024), and Self-Aug further adapts contrastive decoding through query-dependent augmentation and entropy-aware control (Im et al., 2026). Post-hoc methods refine generated responses using external tools. DnR introduces question-conditioned visual analysis with 4 visual experts such as Grounding DINO (Liu et al., 2024) and SAM (Kirillov et al., 2023) (Jeong et al., 2026), while Woodpecker builds an explicit verification pipeline using GPT-3.5-turbo, Grounding DINO, and BLIP-2-FlanT5 (Yin et al., 2024). Despite their effectiveness, these methods require modifying the base model or depend on multiple external tools.

2.2 On-Policy Self-Distillation

On-Policy Distillation (OPD) trains a student on states induced by its own policy, reducing the distribution mismatch between teacher-generated training trajectories and student inference (Agarwal et al., 2024). On-Policy Self-Distillation (OPSD) further constructs the teacher and student from the same model by conditioning them on different information, allowing privileged context to provide dense token-level supervision without a separate teacher model (Zhao et al., 2026). Recent work extends on-policy distillation to MLLMs. Vision-OPD transfers fine-grained regional perception from a crop-conditioned teacher to a full-image student (Yuan et al., 2026), while Visual-OPSD uses privileged visual-thought traces to improve multimodal reasoning without retaining the expensive visual generation process at inference (Li et al., 2026). These methods show that privileged information can substantially improve multimodal on-policy learning. Our work considers a different problem arising from heterogeneous privileged supervision, for which we propose *Concordance-Calibrated OPSD (CC-OPSD)* to enable reliable experience internalization and improve visual grounding and response handling in the supervisory circuit.

3 Method

Our method consists of three stages, as shown in Figure 2. We first transform collected cases into structured data that captures both hallucination correction and response preservation. We then internalize experience into the supervisory circuit through CC-OPSD. At inference, the supervisory circuit assesses the candidate response and provides feedback to generate the final response.

3.1 Multimodal Experience Construction

We first construct multimodal experience data for training the supervisory circuit. Inspired by how humans learn from past experience, we regard model-generated responses as valuable learning signals. However, raw cases provide only coarse correctness labels, without revealing which visual claims determine response reliability or how they should be handled. Model-based memory studies show that raw corpora can be transformed into structured question–answer supervision for knowledge internalization (Quek et al., 2026). We extend this principle to multimodal hallucination and derive structured experience from raw response cases.

Visual Claim Structuring. We start from 1,000 response cases sampled from REVERIE (Zhang et al., 2024), with positive and negative examples equally represented. Each case contains an image I_i, a question Q_i, and a candidate response A_i. Since a free-form response may contain multiple visual statements with different grounding quality, we use Qwen3.6-35B to decompose A_i into atomic visual claims $\mathcal{C}_i = \{c_{i1}, c_{i2}, \ldots, c_{iK_i}\}$, where each c_{ik} represents a single statement that can be independently examined against the visual input. Compound descriptions are separated into distinct claims covering visual content such as object existence, attributes, counts, relations, and actions. Semantically redundant or overlapping claims are further consolidated into standardized claim units $\mathcal{U}_i = \{u_{i1}, u_{i2}, \ldots, u_{iM_i}\}$. Each u_{ij} retains its correspondence to the original response while representing a distinct visual proposition. No correctness judgment is introduced at this stage; the goal is to establish a fine-grained and consistent interface for subsequent grounding assessment.

Grounded Experience Derivation. For each structured claim unit u_{ij}, Qwen3.6-35B assesses its grounding using the visual input together with the reference information available during data

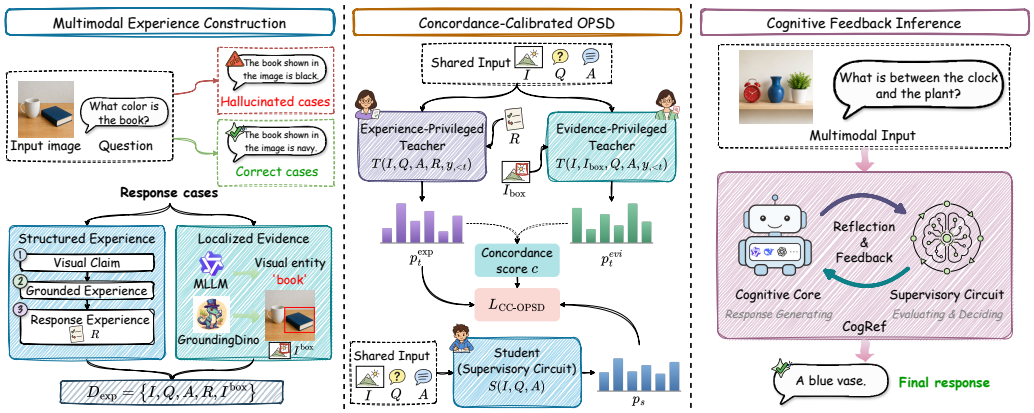


Figure 2: Overview of CogRef. Multimodal experience is first constructed from response cases and then internalized into the supervisory circuit through concordance-calibrated OPSD, enabling plug-and-play cognitive feedback at inference.

construction. We assign one of four grounding states,

$$z_{ij} \in \{\textsc{Supported}, \textsc{Contradicted}, \textsc{Absent}, \textsc{Insufficient}\}, \tag{1}$$

which distinguish visually grounded content from claims that conflict with the evidence, introduce nonexistent content, or lack sufficient support. Each state is associated with a corresponding handling action. Beyond the instance-level decision, we further abstract each verified unit into an experience pattern $e_{ij} = \phi(u_{ij}, z_{ij}, g(z_{ij}))$, where $\phi(\cdot)$ captures the underlying grounding principle behind the decision. Claims requiring correction, removal, or hedging yield reusable anti-hallucination experience, while supported claims yield preservation experience that discourages unnecessary modification. This abstraction reduces dependence on a particular wording or visual instance and exposes the response-handling principle that can transfer to future cases.

Response-Level Experience Synthesis. A candidate response may contain both reliable and unreliable claims. We therefore synthesize the claim-level grounding decisions and experience patterns into a unified response-level representation,

$$R_i = \Psi\left(\{u_{ij}, z_{ij}, g(z_{ij}), e_{ij}\}_{j=1}^{M_i}\right), \tag{2}$$

where $\Psi(\cdot)$ denotes the synthesis process performed by Qwen3.6-35B. The resulting R_i summarizes how the candidate response should be assessed and provides targeted guidance for its subsequent handling. For grounded responses, R_i mainly captures preservation experience. For hallucinated or partially grounded responses, it combines preservation of supported content with correction guidance for unreliable claims. Importantly, R_i is not used as a revised final answer. It serves as structured privileged experience for teaching the supervisory circuit how to evaluate a response and provide appropriate feedback to the cognitive core.

Localized Evidence Construction. In parallel with the structured experience R_i, we construct a question-conditioned visual view that provides complementary grounding evidence. Given I_i and Q_i, Qwen3-VL-32B (Bai et al., 2025) first identifies the visual entities most relevant to the question and generates semantic localization targets. Grounding DINO (Liu et al., 2024) then localizes the corresponding regions on the original image. We highlight the selected regions with bounding boxes while preserving the complete scene context, producing a localized evidence image I_i^{box}. For questions that depend on the global scene or do not admit a reliable local target, the original image is retained without additional localization. The localized view is used only during training and is not required at inference time.

After construction, the training data are represented as

$$\mathcal{D}_{\text{exp}} = \left\{\left(I_i, Q_i, A_i, R_i, I_i^{\text{box}}\right)\right\}_{i=1}^{N}. \tag{3}$$

Here, R_i forms the *experience privilege*, encoding correction and preservation experience, while I_i^{box} forms the *evidence privilege*, highlighting the relevant visual evidence. The next section describes how these two privileged views are coordinated to train the supervisory circuit.

3.2 Experience Internalization for Supervisory Circuit Learning

Fixed offline supervision can suffer from a mismatch between training states and those induced by the learner in sequential prediction (Ross et al., 2011; Agarwal et al., 2024). To reliably internalize the constructed multimodal experience into the supervisory circuit, we adopt OPSD, where a privileged teacher provides dense supervision along trajectories generated by the current student policy. We employ the structured multimodal experience R_i as *experience privilege*, enabling the supervisory circuit to internalize when and how to intervene. Motivated by prior findings that region-level supervision can be distilled into fine-grained perception from the full image (Wei et al., 2026), the localized visual evidence I_i^{box} serves as *evidence privilege*, enabling the circuit to identify hallucination-relevant regions without external grounding models, crop-and-zoom tools, or region annotations at inference. Consequently, these privileges provide complementary signals for internalizing response-handling knowledge and visual grounding cues into the supervisory circuit.

Joint-Privilege Conflict. A natural design is to provide both privileges to a single teacher (Yuan et al., 2026; Li et al., 2026). Given a student generated prefix $y_{i,<t}$, the resulting jointly privileged teacher distribution is

$$p_{i,t}^{\text{joint}} = T\left(\cdot \mid I_i, I_i^{\text{box}}, Q_i, A_i, R_i, y_{i,<t}\right). \tag{4}$$

Here, R_i provides response handling experience derived from previous errors and grounded responses, while I_i^{box} supplies focused visual evidence for the current query. Their joint use appears desirable because the two privileges provide different information about the same candidate response. However, we find that joint conditioning does not consistently yield stronger supervision. In particular, a teacher conditioned on both privileges can lead to worse distillation than one using the experience privilege alone. This observation suggests that heterogeneous privileges should not be combined unconditionally. We further analyze this issue in Section 4.3.

Concordance-Calibrated OPSD. To make the contribution of each privilege explicit, we decompose the jointly privileged teacher into two separate privileged views. For each training instance, the supervisory student only observes the original image I_i, question Q_i, and candidate response A_i. A trajectory $y_i \sim S_\theta\left(\cdot \mid I_i, Q_i, A_i\right)$ is first sampled from the current student policy. At decoding step t, the corresponding student distribution is $p_{i,t}^{S} = S_\theta\left(\cdot \mid I_i, Q_i, A_i, y_{i,<t}\right)$. All teacher distributions are evaluated on the same student generated prefix $y_{i,<t}$, so that supervision is provided on states actually visited by the supervisory circuit.

The *experience privileged teacher* additionally observes the structured error experience R_i,

$$p_{i,t}^{\text{exp}} = T\left(\cdot \mid I_i, Q_i, A_i, R_i, y_{i,<t}\right), \tag{5}$$

where R_i contains both correction experience for erroneous or insufficiently grounded content and preservation experience for visually supported content. The *evidence privileged teacher* instead receives the localized evidence image while retaining the original visual context,

$$p_{i,t}^{\text{evi}} = T\left(\cdot \mid I_i, I_i^{\text{box}}, Q_i, A_i, y_{i,<t}\right). \tag{6}$$

This decomposition exposes the two privileged effects separately and makes their token level agreement directly measurable. We then use the experience privileged teacher as the supervision anchor, as R_i explicitly encodes how the candidate response should be preserved or corrected. Accordingly, the distillation target is always defined by $p_{i,t}^{\text{target}} = p_{i,t}^{\text{exp}}$. The evidence privileged teacher does not directly modify this target. Instead, it determines how strongly each token should contribute to distillation based on the support provided by visual evidence. Specifically, we measure the token level disagreement between the two privileged teachers using the Jensen–Shannon divergence, $d_{i,t} = \text{JS}\left(p_{i,t}^{\text{exp}}, p_{i,t}^{\text{evi}}\right)$. We then convert this disagreement into a concordance score $c_{i,t} = \exp\left(-\frac{d_{i,t}}{\tau}\right)$, where τ controls the sensitivity to privilege disagreement. A small divergence produces a high concordance score, indicating that the localized visual evidence supports the response handling experience. Conversely, a large divergence reduces the influence of the evidence privilege. The concordance score is used to calibrate the token level distillation weight:

$w_{i,t} = 1 + \lambda c_{i,t}$, where λ controls the maximum contribution of evidence based calibration. The resulting CC-OPSD objective is

$$\mathcal{L}_{\text{CC-OPSD}} = \mathbb{E}_{x_i \sim \mathcal{D}} \mathbb{E}_{y_i \sim S_\theta(\cdot|x_i)} \left[\frac{\sum_{t=1}^{T_i} w_{i,t} D_{\text{KL}} \left(p_{i,t}^{\text{exp}} \| p_{i,t}^{S} \right)}{\sum_{t=1}^{T_i} w_{i,t}} \right]. \tag{7}$$

When the two privileged teachers agree, the corresponding token receives a larger relative contribution to the distillation objective. When disagreement is high, $c_{i,t}$ approaches zero and the token retains only its baseline experience-supervision weight. This allows the supervisory circuit to internalize knowledge from both privileges while avoiding conflicts between their signals.

3.3 CogRef: Cognitive Feedback Inference

Candidate Response Generation. At inference time, CogRef first uses the base MLLM as the *cognitive core* to generate a candidate response. Given an image I and a question Q, the cognitive core $\mathcal{M}$ produces$A \sim \mathcal{M}(\cdot \mid I, Q)$. The cognitive core is not trained for hallucination supervision and therefore retains its original pretrained knowledge and generation capability. It remains responsible for both the initial response and the final output. This functional separation allows hallucination-specific experience to be introduced without modifying the base MLLM.

Supervisory Assessment and Feedback. Given the candidate response A, the learned supervisory circuit $\mathcal{S}_\theta$ evaluates its consistency with the original visual input and produces structured feedback $F = \mathcal{S}_\theta(I, Q, A)$. Unlike the privileged teachers used during training, the supervisory circuit has access only to the original image, question, and candidate response at inference time. It examines the visual claims expressed in A and identifies content that is unsupported, inconsistent, or insufficiently grounded. At the same time, visually supported content is explicitly retained in the feedback to reduce unnecessary modification. The resulting F summarizes the response assessment and provides targeted guidance on how the candidate response should be handled. Importantly, the supervisory circuit does not generate the final answer; it only provides feedback to guide the final generation.

Feedback-Guided Final Generation. The structured feedback F is finally returned to the cognitive core to guide the generation of the user-facing response. Conditioned on the original image, question, candidate response, and supervisory feedback, the cognitive core produces the final response $A^* \sim \mathcal{M}(\cdot \mid I, Q, A, F)$. Since the final response is still generated by the original base MLLM, the supervisory circuit only provides assessment and corrective guidance without taking over the generation process. No privileged information or external tools are required at inference time. The inference process of CogRef can be summarized as

$$(I, Q) \xrightarrow{\mathcal{M}} A \;\rightarrow\; (I, Q, A) \xrightarrow{\mathcal{S}_\theta} F \;\rightarrow\; (I, Q, A, F) \xrightarrow{\mathcal{M}} A^*, \tag{8}$$

where the cognitive core $\mathcal{M}$ performs both candidate and final response generation, while the supervisory circuit $\mathcal{S}_\theta$ converts the internalized error experience into feedback for the final generation. This modular design enables plug-and-play integration with different MLLMs, including closed-source models, without accessing or modifying their parameters.

4 Experiments

Our experiments are organized around four research questions that progressively assess performance, experience internalization, privileged supervision, and learned behavior of CogRef: ✤ **RQ1.** Can CogRef consistently mitigate multimodal hallucinations across different MLLMs and model scales without compromising general multimodal capability? ✤ **RQ2.** Does experience internalization improve the effectiveness of supervisory feedback? ✤ **RQ3.** Can CC-OPSD coordinate experience and evidence privileges under supervisory conflicts? ✤ **RQ4.** What capabilities does the learned circuit internalize to guide the final response?

Implementation details. We train the supervisory circuit for 100 optimizer steps using LoRA with rank 64. Following Zhao et al. (2026), we keep both privileged teachers frozen throughout training. We evaluate CogRef on two hallucination benchmarks, MHaluBench (Chen et al., 2024b)

Table 1: Main results on multimodal hallucination and general multimodal benchmarks. Each entry presents the results of the 4B/8B cognitive cores, respectively. The best results are shown in **bold**.

Method	MHaluBench			HallusionBench			MMStar	MMMU-Pro Vision
	Accuracy ↑	Recall ↑	F1 ↑	qAcc ↑	fAcc ↑	aAcc ↑	Accuracy ↑	Accuracy ↑
Cognitive core: Qwen3-VL-Instruct (4B/8B)								
Baseline	60.16/77.74	52.61/75.24	68.77/84.93	34.73/40.66	42.20/49.13	65.81/70.95	61.47/64.00	16.40/24.51
VCD	59.68/78.23	52.03/75.82	68.27/85.31	35.82/42.64	43.93/50.58	66.70/**71.57**	62.13/64.33	17.30/26.13
Self-Aug	59.52/77.90	51.84/75.63	68.11/85.09	34.07/42.20	43.35/50.58	65.46/71.39	62.20/64.47	15.70/25.72
DnR	51.77/76.45	42.36/73.11	59.43/83.81	34.95/38.68	42.20/47.98	66.43/70.15	61.40/63.13	16.20/24.80
CogRef	**76.94/83.39**	**74.66/83.37**	**84.37/89.33**	**42.20/43.74**	**51.16/51.16**	**71.48**/71.21	**65.47/66.00**	**25.49/26.65**
Cognitive core: InternVL3.5 (4B/8B)								
Baseline	54.68/61.77	46.42/55.32	63.07/70.70	36.04/38.68	43.06/47.40	67.85/**70.77**	53.07/57.80	27.86/29.54
VCD	56.61/61.45	48.55/54.74	65.11/70.31	36.04/38.24	44.22/47.40	68.11/70.59	51.33/59.13	28.73/30.52
Self-Aug	57.42/61.61	49.32/54.74	65.89/70.40	38.46/38.68	**45.38**/48.55	**69.09**/70.33	53.47/**59.60**	29.83/31.33
DnR	38.87/57.10	27.47/49.90	42.84/65.98	38.02/36.04	41.04/42.20	66.08/68.64	51.73/53.67	19.48/25.38
CogRef	**79.19/82.74**	**77.37/82.01**	**86.11/88.80**	**40.44/41.10**	45.09/**49.13**	68.64/69.44	**64.87**/59.40	**31.21/32.89**
Cognitive core: LLaVA-OneVision-1.5 (4B/8B)								
Baseline	31.45/44.03	17.79/33.08	30.21/49.64	32.75/38.24	40.46/45.95	65.19/67.94	64.33/64.93	26.82/28.09
VCD	35.00/45.48	22.44/34.82	36.54/51.58	34.29/39.12	40.17/45.66	65.10/68.47	63.13/63.53	26.76/28.32
Self-Aug	34.84/45.32	22.24/34.62	36.28/51.36	33.85/36.48	40.17/42.77	65.63/66.87	65.33/65.33	27.34/27.98
DnR	28.23/40.65	13.93/29.01	24.45/44.91	30.99/38.02	35.55/41.91	63.51/67.85	61.67/64.60	25.20/27.86
CogRef	**76.77/77.10**	**74.85/75.24**	**84.31/84.57**	**42.42/41.32**	**52.31/51.45**	**71.48/70.95**	**67.47/67.67**	**32.02/32.95**

and HallusionBench (Guan et al., 2024), and two general multimodal benchmarks, MMStar (Chen et al., 2024a) and MMMU-Pro (Vision) (Yue et al., 2025). This combination assesses hallucination mitigation together with the preservation of general multimodal capabilities. We report the official evaluation metrics of each benchmark. More details are provided in *Appendices* B and C.

Baselines. We compare with the baseline and three representative hallucination mitigation methods: VCD (Leng et al., 2024), Self-Aug (Im et al., 2026), and DnR (Jeong et al., 2026). The supervisory circuit is initialized from Qwen3-VL-4B and trained once using the 1,000 constructed experience samples. We employ three representative MLLMs, Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025), and LLaVA-OneVision-1.5 (An et al., 2025), as cognitive cores.

4.1 Overall Effectiveness across MLLMs (RQ1)

Hallucination Mitigation across MLLMs. Table 1 summarizes the results across three MLLM families and two model scales. CogRef consistently improves the corresponding base models on both hallucination benchmarks. On MHaluBench, accuracy improves by $\mathbf{16.78}/\mathbf{5.65}$, $\mathbf{24.51}/\mathbf{20.97}$, and $\mathbf{45.32}/\mathbf{33.07}$ points for Qwen3-VL, InternVL3.5, and LLaVA-OV-1.5 (4B/8B), respectively. On HallusionBench, the corresponding qAcc gains are $\mathbf{7.47}/\mathbf{3.08}$, $\mathbf{4.40}/\mathbf{2.42}$, and $\mathbf{9.67}/\mathbf{3.08}$ points. CogRef achieves the best result in $\mathbf{32}/\mathbf{36}$ evaluated settings, demonstrating consistent effectiveness across heterogeneous cognitive cores.

General Multimodal Capability Preservation. CogRef also improves general multimodal capability on MMStar and MMMU-Pro Vision. Compared with the corresponding baselines, MMStar improves by $\mathbf{4.00}/\mathbf{2.00}$, $\mathbf{11.80}/\mathbf{1.60}$, and $\mathbf{3.14}/\mathbf{2.74}$ points for Qwen3-VL, InternVL3.5, and LLaVA-OneVision-1.5, while MMMU-Pro Vision improves by $\mathbf{9.09}/\mathbf{2.14}$, $\mathbf{3.35}/\mathbf{3.35}$, and $\mathbf{5.20}/\mathbf{4.86}$ points. In contrast, each competing method shows performance degradation in at least one general multimodal setting, highlighting the stronger capability preservation of CogRef.

Together, these results provide a positive answer to **RQ1**: CogRef generalizes across MLLMs and model scales while retaining their general multimodal capabilities.

Table 2: Controlled analysis of multimodal experience internalization on MHaluBench using Qwen3-VL-4B as the cognitive core.

Setting	Performance		Prediction Transitions		
	Acc. ↑	F1 ↑	Corr. ↑	Pres. ↑	Net ↑
Base cognitive core	60.16	68.77	–	–	–
DnR	51.77	59.43	4.45	83.11	−8.39
Vanilla circuit	64.19	72.86	13.36	**97.86**	4.03
Ours	**76.94**	**84.37**	**47.37**	96.51	**16.77**

Table 3: Comparison of privileged supervision strategies. All variants use the same training data and optimization settings.

Training Strategy	Distillation Target	Acc. ↑	Rec. ↑	F1 ↑
Vanilla	–	64.19	57.64	72.86
Experience-only	p^{exp}	75.00	73.89	83.13
Evidence-only	p^{evi}	74.84	72.34	82.74
Joint-privilege	p^{joint}	72.58	70.02	80.98
CC-OPSD	p^{exp}	**76.94**	**74.66**	**84.37**

4.2 Effect of Multimodal Experience Internalization (RQ2)

Ablation of the Supervisory Circuit. We first examine whether the gain of CogRef can be attributed simply to introducing an additional feedback stage. As shown in Table 2, DnR decreases accuracy $\mathbf{60.16}\% \rightarrow \mathbf{51.77}\%$, suggesting that additional revision alone does not necessarily improve reliability. In contrast, introducing a vanilla supervisory circuit improves accuracy and F1 to $\mathbf{64.19}\%$ and $\mathbf{72.86}\%$, respectively. After experience internalization, both metrics improve substantially, with accuracy $\mathbf{64.19}\% \rightarrow \mathbf{76.94}\%$ and F1 $\mathbf{72.86}\% \rightarrow \mathbf{84.37}\%$, corresponding to gains of $\mathbf{12.75}$ and $\mathbf{11.51}$ points over the vanilla variant. Since the vanilla and internalized circuits share the same supervisory architecture and inference pipeline, this improvement isolates the contribution of the learned supervisory capability.

Effectiveness of Supervisory Circuit Feedback. We further examine whether the learned feedback leads to more reliable response-level interventions. Let $N_{\text{W}\to\text{C}}$ denote initially incorrect responses that become correct after feedback, and $N_{\text{C}\to\text{W}}$ denote initially correct responses that become incorrect. We report the correction rate $\text{Corr.} = N_{\text{W}\to\text{C}}/N_{\text{W}}$, the preservation rate $\text{Pres.} = N_{\text{C}\to\text{C}}/N_{\text{C}}$, and the normalized net improvement $\text{Net} = (N_{\text{W}\to\text{C}} - N_{\text{C}\to\text{W}})/N$. As shown in Table 2, experience internalization increases the correction rate from $\mathbf{13.36}\%$ to $\mathbf{47.37}\%$, while largely retaining the high preservation rate of the vanilla circuit. Consequently, the net improvement rises substantially from $\mathbf{4.03}$ to $\mathbf{16.77}$ points. In contrast, DnR achieves only a $\mathbf{4.45}\%$ correction rate and reduces preservation to $\mathbf{83.11}\%$, resulting in a negative net improvement of $\mathbf{-8.39}$ points. These results show that experience internalization substantially strengthens corrective feedback without inducing excessive changes to already grounded responses, yielding more effective supervisory intervention overall.

4.3 Privilege Interaction and Concordance Calibration (RQ3)

Effect of Privilege Integration. We first compare different ways of incorporating privileged information into OPSD. Experience-only and evidence-only OPSD use the corresponding privileged view independently, while joint-privilege OPSD conditions a single teacher on both. As shown in Table 3, joint conditioning achieves $\mathbf{72.58}\%$ in accuracy, compared with $\mathbf{75.00}\%$ for experience-only supervision. Thus, simply exposing the teacher to both privileges does not necessarily yield more effective supervision.

Joint Supervisory Conflict. We next examine why direct privilege integration can be suboptimal. For each token, we measure the cosine similarity between the supervision directions induced by the experience and evidence teachers relative to the student. Figure 3 summarizes this interaction at two levels: the horizontal axis reports the mean token-level alignment within each sample, while the vertical axis measures the fraction of negatively aligned tokens. Samples with negative mean alignment exhibit global conflict, whereas a positive mean can still contain substantial local conflict. The figure shows that both forms occur in practice: some samples are conflicting on average, while many others contain a considerable proportion of conflicting tokens despite positive mean alignment. This fine-grained inconsistency

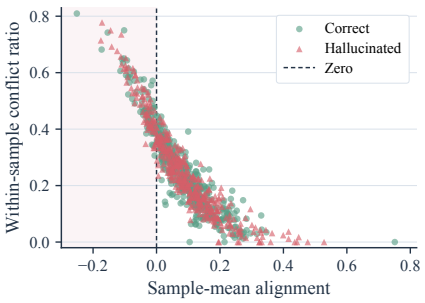


Figure 3: Global and local conflicts.



Question.
According to the positive sequence of the images, is there a dog in this video? Answer in one word.

Ground Truth: `no`

Candidate Response: `yes`

Before CC-OPSD Training.

Supervisory Feedback.

The draft answer is supported by the visual evidence. The image contains **a person with a dog's head (replacing their face)**, which directly indicates the presence of a dog in the sequence. The answer "yes" aligns with this observation. **Keep the draft answer unchanged.**

Final Output: `yes`

After CC-OPSD Training.

Supervisory Feedback.

The draft answer is not supported by the visual evidence. The image contains **a cat (not a dog)** in the sequence, so the correct answer should be "no". **Revise the draft answer to "no".**

Final Output: `no`

Figure 4: Qualitative results of COGREF. More results are provided in *Appendix* E.

helps explain why directly combining the two privileges can yield unreliable supervision and motivates token-level concordance calibration.

Concordance Calibration. CC-OPSD addresses these conflicts by retaining the experience teacher as the distillation target and using evidence agreement only to calibrate token-level supervision strength. As shown in Table 3, CC-OPSD achieves $\mathbf{76.94}\%$ in accuracy, improving over joint-privilege OPSD by $\mathbf{4.36}$ points. These results indicate that evidence privilege is more effective when it reinforces concordant supervision without directly redirecting the distillation target, providing a positive answer to **RQ3**.

4.4 ANALYSIS OF INTERNALIZED SUPERVISORY CAPABILITY (RQ4)

To understand what the learned supervisory circuit internalizes, we compare its feedback before and after CC-OPSD training (Figure 4). Before training, the circuit incorrectly identifies a "dog's head," and consequently provides incorrect feedback. After training, it correctly grounds the visible animal as a cat, and directs the cognitive core to revise its answer. This comparison suggests that CC-OPSD internalizes two complementary supervisory capabilities. ① **Reliable visual grounding.** The circuit grounds its judgment in the actual visual content, reducing hallucination within the supervisory feedback itself. ② **Accurate feedback.** The circuit translates the grounded judgment into an appropriate revision instruction, guiding the cognitive core to correct its erroneous response. These capabilities enable the supervisory circuit to determine when intervention is needed and when the cognitive core's original response should be preserved.

5 CONCLUSION

In this paper, we propose COGREF, a cognitive reflection framework that separates hallucination supervision from multimodal task execution. COGREF keeps the cognitive core frozen and internalizes correction and preservation experience into a dedicated supervisory circuit, enabling hallucination-aware feedback without external verification tools at inference. We further showed that heterogeneous privileged signals can introduce local conflicts during self-distillation, and proposed CC-OPSD to coordinate experience and evidence supervision through token-level concordance calibration. Experiments across different MLLMs and multimodal benchmarks demonstrate effective hallucination mitigation while preserving general multimodal capability. Limitations and future work are discussed in *Appendix* G.

AI USAGE STATEMENT

In this work, we used AI to assist with the construction and refinement of model-generated supervisory experience, language editing to improve grammar, sentence structure, and readability, and limited coding assistance. Generative AI tools were not used to develop the core scientific content of this work, including research ideas, methodology, experimental design, result analysis, or scientific conclusions. All AI-assisted content was manually reviewed and verified by the authors. The authors take responsibility for the accuracy and integrity of all content in this work, including materials produced with generative AI assistance.

ETHICS STATEMENT

This work adheres to the ICLR Code of Ethics. No human subjects or animal experimentation were involved in this study. All datasets and pretrained models used in our experiments are publicly available and were used in accordance with their respective licenses and terms of use. No personally identifiable or sensitive information was collected or used.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed framework, experimental settings, and evaluation protocols used in this work. Implementation details, including model configurations and training settings, are provided in the paper. The pretrained multimodal large language models and evaluation benchmarks used in our experiments are publicly available.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Xiang An, Yin Xie, Kaicheng Yang, Wenkang Zhang, Xiuwei Zhao, Zheng Cheng, Yirui Wang, Songcen Xu, Changrui Chen, Didi Zhu, et al. Llava-onevision-1.5: Fully open framework for democratized multimodal training. *arXiv preprint arXiv:2509.23661*, 2025.

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*, 2025.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024a.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3235–3252, 2024b.

Google DeepMind. Gemini 3.1 pro model card, 2026. Model Card.

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

Eun Woo Im, Muhammad Kashif Ali, and Vivek Gupta. Self-aug: Query and entropy adaptive decoding for large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 136345–136368, 2026.

Sungheon Jeong, Ryozo Masukawa, Jihong Park, Sanggeon Yun, Wenjun Huang, Hanning Chen, Mahdi Imani, and Mohsen Imani. Draft and refine with visual experts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18816–18826, 2026.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Pengyu Li, Zhitao Gao, Lingling Zhang, Muye Huang, Yuanming Li, Fangzhi Xu, and Jun Liu. Visual-opsd: Cross-modal on-policy self-distillation for efficient unified multimodal reasoning. *arXiv preprint arXiv:2606.18974*, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Ryan Wei Heng Quek, Sanghyuk Lee, Alfred Wei Lun Leong, Arun Verma, Alok Prakash, Nancy F. Chen, Bryan Kian Hsiang Low, Daniela Rus, and Armando Solar-Lezama. Memo: Memory as a model. *arXiv preprint arXiv:2605.15156*, 2026.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Melissa J Sharpe, Hannah M Batchelor, Lauren E Mueller, Matthew PH Gardner, and Geoffrey Schoenbaum. Past experience shapes the neural circuits recruited for future learning. *Nature Neuroscience*, 24(3):391–400, 2021.

Weinan Sun, Madhu Advani, Nelson Spruston, Andrew Saxe, and James E Fitzgerald. Organizing memories for generalization in complementary learning systems. *Nature neuroscience*, 26(8): 1438–1448, 2023.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Changyao Tian, Danni Yang, Guanzhou Chen, Erfei Cui, Zhaokai Wang, Yuchen Duan, Penghao Yin, Sitao Chen, Ganlin Yang, Mingxin Liu, et al. Internvl-u: Democratizing unified multimodal models for understanding, reasoning, generation and editing. *arXiv preprint arXiv:2603.09877*, 2026.

Yuiga Wada, Kazuki Matsuda, Komei Sugiura, and Graham Neubig. Zina: Multimodal fine-grained hallucination detection and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32528–32538, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Lai Wei, Liangbo He, jun lan, Lingzhong Dong, Yutong Cai, Siyuan Li, Huijia Zhu, Weiqiang Wang, Linghe Kong, Yue Wang, Zhuosheng Zhang, and Weiran Huang. Zooming without zooming: Region-to-image distillation for fine-grained multimodal perception. In *Forty-third International Conference on Machine Learning*, 2026.

Tsung-Han Patrick Wu, Heekyung Lee, Jiaxin Ge, Joseph Gonzalez, Trevor Darrell, and David Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *Advances in Neural Information Processing Systems*, 38:65749–65777, 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.

Qianhao Yuan, Jie Lou, Xing Yu, Hongyu Lin, Le Sun, Xianpei Han, and Yaojie Lu. Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation. *arXiv preprint arXiv:2605.18740*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pp. 196–213. Springer, 2024.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *Forty-third International Conference on Machine Learning*, 2026.

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

A MULTIMODAL EXPERIENCE DETAILS

The main paper describes the construction of multimodal experience from raw response cases. Here, we provide the concrete form of the resulting experience data and a representative example used for supervisory circuit learning.

A.1 EXPERIENCE DATA FORMAT

After construction, the multimodal experience pool is represented as

$$\mathcal{D}_{\text{exp}} = \left\{ \left(I_i, Q_i, A_i, R_i, I_i^{\text{box}} \right) \right\}_{i=1}^{N}. \tag{9}$$

Here, I_i and Q_i denote the original image and question, respectively, and A_i is the candidate response produced by the cognitive core. R_i is the structured response-level experience synthesized from the claim-level grounding analysis, encoding both how individual claims should be handled and the transferable grounding principles behind these decisions. I_i^{box} is the localized evidence view obtained by highlighting question-relevant regions while retaining the complete image context.

More specifically, the information underlying R_i can be characterized by

$$\{u_{ij}, z_{ij}, g(z_{ij}), e_{ij}\}_{j=1}^{M_i}, \tag{10}$$

where u_{ij} denotes a consolidated claim unit, z_{ij} its grounding state, $g(z_{ij})$ the corresponding handling action, and e_{ij} the abstracted experience pattern. These claim-level signals are synthesized into R_i to describe how the candidate response should be handled as a whole. Accordingly, grounded content contributes preservation experience, whereas contradicted, absent, or insufficiently supported content contributes experience for appropriate intervention. Importantly, R_i is not a reference answer; it serves as training-time privileged experience for learning supervisory behavior.

During CC-OPSD, R_i serves as the *experience privilege*, whereas I_i^{box} provides the complementary *evidence privilege*. Both are used only during training and are removed at inference time.

A.2 REPRESENTATIVE EXPERIENCE CASE

The following example illustrates how the components of a single instance in $\mathcal{D}_{\text{exp}}$ are instantiated. The candidate response fabricates the text shown on a sign held by a teddy bear. The structured experience identifies the conflicting claim, grounds the correction in the visible text, and abstracts the corresponding response-handling principle. This example shows that $\mathcal{D}_{\text{exp}}$ provides more than a corrected target. The experience explicitly links a response claim to its visual grounding state and appropriate handling behavior, while abstracting the case into a reusable grounding principle. This structure allows the supervisory circuit to internalize both hallucination correction and the preservation of already grounded content.

B EVALUATION DETAILS

B.1 MODEL FAMILIES.

We evaluate COGREF with three representative open-source MLLM families, using the 4B and 8B variants of each family as cognitive cores. **Qwen3-VL** (Bai et al., 2025) is a general-purpose multimodal model family designed for strong visual perception and multimodal reasoning, with enhanced spatial–temporal modeling and multi-level visual feature integration. **InternVL3.5** (Wang et al., 2025) follows the ViT–MLP–LLM paradigm with dynamic high-resolution visual processing and further strengthens multimodal reasoning through its progressive training pipeline. **LLaVA-OneVision-1.5** (An et al., 2025) is a fully open vision-language model family trained with large-scale multimodal pretraining and instruction-tuning data, providing a distinct model architecture and training recipe from the other two families. Evaluating across these model families allows us to assess whether the learned supervisory circuit can provide consistent hallucination-aware supervision beyond a particular cognitive-core architecture or training paradigm.

Multimodal Experience Case



Original Image I_i



Localized Evidence I_i^{box}

Question Q_i**.** *What is the teddy bear saying? Offer a short answer.*

Candidate Response A_i**.** *I'm lost.*

Claim-Level Grounding.

- **Claim Unit** u_{ij}**:** "I'm lost."
- **Grounding State** z_{ij}**:** CONTRADICTED.
- **Handling Action** $g(z_{ij})$**:** CORRECT.
- **Visual Grounding:** The sign held by the teddy bear clearly reads "HELP!", which directly contradicts the textual content claimed in the candidate response.
- **Experience Pattern** e_{ij}**:** Textual claims about visible signs should be grounded in the characters that can be directly read from the image; fabricated or mismatched text should be corrected according to the visual evidence.

Final Experience R_i**.**
The candidate response contains a fabricated textual claim about the sign. The phrase "I'm lost." conflicts with the visible text "HELP!" and should therefore be corrected accordingly. Any other content that is visually supported should remain unchanged.

B.2 BENCHMARKS.

We evaluate hallucination mitigation on MHaluBench and HallusionBench, and use MMStar and MMMU-Pro to measure the preservation of general multimodal capabilities.

MHaluBench. MHaluBench (Chen et al., 2024b) is a multimodal hallucination detection benchmark covering both modality-conflicting and fact-conflicting hallucinations, including errors related to objects, attributes, scene text, and factual information. The benchmark provides hallucination annotations at multiple granularities. In our experiments, we exclusively use its *claim-level* evaluation, where model outputs are decomposed into atomic claims and each claim is classified according to whether it is supported or hallucinated. This setting provides a fine-grained measure of the model's ability to identify unsupported multimodal content without conflating multiple claims within the same response.

HallusionBench. HallusionBench (Guan et al., 2024) evaluates multimodal models under cases where visual evidence may conflict with linguistic priors or induce visual illusions. It organizes its questions into *Visual Dependent* (VD) and *Visual Supplement* (VS) categories. VD questions require visual evidence to determine the answer and include controlled variants in which image manipulation can alter the correct conclusion. VS questions can be partially interpreted from the textual

context, while the accompanying visual input provides additional evidence that should be correctly incorporated. The benchmark further constructs related questions over shared visual concepts and figures, enabling evaluation of both individual correctness and consistency across semantically related questions. This property makes HallusionBench particularly suitable for testing whether a model grounds its decisions in the provided visual evidence instead of relying on prior expectations.

MMStar. MMStar (Chen et al., 2024a) is a vision-indispensable multimodal benchmark designed to reduce text-only shortcuts and potential data leakage in multimodal evaluation. It covers six core capabilities: *Coarse Perception*, *Fine-grained Perception*, *Instance Reasoning*, *Logical Reasoning*, *Science & Technology*, and *Mathematics*, which are further decomposed into fine-grained evaluation axes. The benchmark therefore spans both visual perception and higher-level multimodal reasoning while explicitly emphasizing questions for which visual information is necessary. We use MMStar to examine whether hallucination mitigation preserves the general perception and reasoning capabilities of the cognitive core.

MMMU-Pro. MMMU-Pro (Yue et al., 2025) extends MMMU (Yue et al., 2024) with a more stringent evaluation of expert-level multimodal understanding and reasoning. It removes questions that can be solved reliably without visual input, introduces more challenging answer candidates, and covers diverse disciplines including art and design, business, science, health and medicine, humanities and social science, and technology and engineering. We evaluate on the *Vision* setting, where the question and candidate options are embedded directly into a screenshot or photograph rather than provided as separate textual input. The model must therefore first perceive the visually rendered problem and then integrate the extracted information with domain knowledge and multimodal reasoning to determine the answer. This setting provides a challenging measure of whether the general visual understanding and reasoning capabilities of the cognitive core are retained after hallucination-oriented intervention.

Evaluation metrics. For MHaluBench, we report claim-level *Accuracy*, *Recall*, and *F1*. Taking hallucinated claims as the positive class, let TP denote hallucinated claims correctly identified as hallucinated, TN non-hallucinated claims correctly identified as grounded, FP grounded claims incorrectly classified as hallucinated, and FN hallucinated claims incorrectly classified as grounded. The metrics are computed as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN}, \tag{12}$$

and

$$\text{F1} = \frac{2TP}{2TP + FP + FN}. \tag{13}$$

HallusionBench reports three complementary metrics: answer accuracy (*aAcc*), question-pair accuracy (*qAcc*), and figure accuracy (*fAcc*). Let

$$c_j = \mathbb{1}[\hat{a}_j = a_j] \tag{14}$$

indicate whether the prediction for the j-th atomic question is correct. Answer accuracy evaluates each question independently:

$$\text{aAcc} = \frac{1}{N} \sum_{j=1}^{N} c_j. \tag{15}$$

For question-pair accuracy, let $\mathcal{P}$ denote the collection of groups sharing the same question under different visual contexts. A group is considered correct only when all of its associated instances are answered correctly:

$$\text{qAcc} = \frac{1}{|\mathcal{P}|} \sum_{P \in \mathcal{P}} \mathbb{1}\left[\forall j \in P,\ c_j = 1\right]. \tag{16}$$

Similarly, let $\mathcal{F}$ denote the collection of figures, with each $F \in \mathcal{F}$ containing all questions associated with the same figure. Figure accuracy requires all questions associated with a figure to be answered correctly:

$$\text{fAcc} = \frac{1}{|\mathcal{F}|} \sum_{F \in \mathcal{F}} \mathbb{1}\left[\forall j \in F,\ c_j = 1\right]. \tag{17}$$

Thus, aAcc measures instance-level correctness, whereas qAcc and fAcc impose stricter consistency requirements across related visual contexts and questions.

For MMStar and the Vision setting of MMMU-Pro, we report multiple-choice accuracy. Given the predicted option $\hat{a}_i$ and ground-truth option a_i for each evaluation instance, accuracy is computed as

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[\hat{a}_i = a_i\right]. \tag{18}$$

All metrics reported in the main results are expressed as percentages.

C More Implementation Details

For supervisory circuit learning, we construct a training pool of 1,000 samples. At each parameter update, we collect 8 valid on-policy rollouts from the current student. Training is performed for 100 update steps with a micro-batch size of 1, and the maximum sequence length is set to 1,024 tokens. Following the OPSD setting, the student operates with thinking disabled, while the privileged teacher uses thinking mode during distillation.

D Detailed Formulation of OPSD

We review the formulation of OPSD (Zhao et al., 2026), which forms the basis of our training objective. Given a training set $\mathcal{D} = \{(x_i, y_i^*)\}_{i=1}^{N}$, OPSD maintains a student policy p_S and a privileged teacher policy p_T. The student only observes the original input x_i, whereas the teacher is additionally conditioned on privileged information y_i^* available only during training.

For each sampled input x_i, OPSD first generates a trajectory from the current student policy:

$$\hat{y}_i = (\hat{y}_{i,1}, \ldots, \hat{y}_{i,T_i}) \sim p_S(\cdot \mid x_i). \tag{19}$$

The resulting student-generated trajectory defines the on-policy states on which distillation is performed. At each decoding step t, both the student and teacher are evaluated on the same student-generated prefix $\hat{y}_{i,<t}$, yielding

$$p_{i,t}^{S} = p_S\left(\cdot \mid x_i, \hat{y}_{i,<t}\right), \tag{20}$$

and

$$p_{i,t}^{T} = p_T\left(\cdot \mid x_i, y_i^*, \hat{y}_{i,<t}\right). \tag{21}$$

Importantly, the teacher does not generate a separate trajectory; instead, it provides privileged token-level supervision along the states visited by the current student.

The student is then optimized to match the privileged teacher distribution over the sampled trajectory:

$$\mathcal{L}_{\text{OPSD}} = \mathbb{E}_{(x_i, y_i^*)\sim\mathcal{D}}\mathbb{E}_{\hat{y}_i\sim p_S(\cdot|x_i)}\left[\frac{1}{T_i}\sum_{t=1}^{T_i} D_{\text{KL}}\left(p_{i,t}^{T} \| p_{i,t}^{S}\right)\right]. \tag{22}$$

By repeatedly sampling trajectories from the evolving student and distilling privileged supervision on these trajectories, OPSD provides dense token-level guidance on states actually encountered by the learner, while requiring no privileged information at inference time.

E Additional Qualitative Analysis

Qualitative Analysis of Supervisory Behavior. To further examine the behavior internalized by the supervisory circuit, we compare its feedback before and after CC-OPSD training on two representative cases. The first case highlights improved hallucination-aware visual grounding, while the second illustrates the transition from verbose and inconclusive reasoning to concise and actionable supervision. In both cases, the cognitive core initially produces an incorrect response, allowing us to directly assess whether the supervisory circuit can identify the underlying error and effectively guide its revision. For readability, `[...]` denotes omitted repetitive or intermediate content. All retained excerpts preserve the original wording and order.

Case 1: From Repeated Reconsideration to Stable Preservation



Question. *This monkey is removing the wheel. According to the positive sequence of the images, are they in the correct order? Answer in one word.*

Ground Truth: `yes` **Candidate Response:** `yes`

Before CC-OPSD Training.

Supervisory Feedback.

> Got it, let's check the images. The question is about whether the sequence of the monkey removing the wheel is correct. First, I need to look at the order of the images. [...]
>
> **Wait, maybe the sequence is correct. But let's think** [...]
>
> So the sequence shows the process of removing the wheel: starting with holding it, then moving it, then lifting, then taking it away. So the order is correct. Therefore, the draft answer "yes" is supported.
>
> **Wait, but wait. Let's check the images again.** The first image: the wheel is on the ground, and the monkey [...]

Final Output: `no`

After CC-OPSD Training.

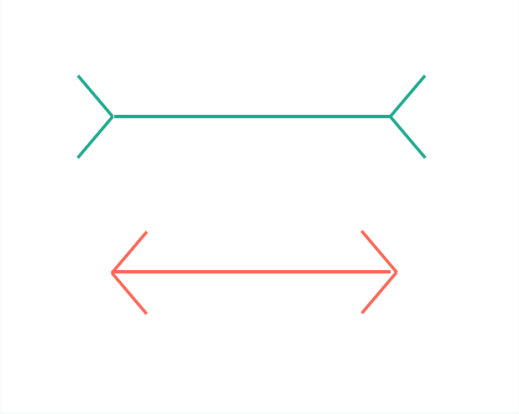
Supervisory Feedback.

> The draft answer is supported by the visual evidence. The sequence of images correctly shows the progression of the monkey removing the wheel (**from holding the wheel to it being detached and carried away**), so the answer "yes" is accurate. **Keep the draft answer unchanged.**

Final Output: `yes`

Analysis. Before training, the supervisory circuit already identifies the **correct temporal progression** and even reaches the correct conclusion that the candidate response is supported. However, it **repeatedly reopens the decision after reaching that conclusion**, preventing the feedback from converging to a stable supervisory action and ultimately yielding an incorrect output. After CC-OPSD training, the circuit compresses the relevant visual evidence into a clear action sequence, and follows it with the explicit instruction **"Keep the draft answer unchanged."** The correct candidate response is therefore preserved. This comparison further indicates that CC-OPSD improves the **stability and decisiveness of supervisory behavior**, particularly when the appropriate action is to preserve an already grounded response.

Case 2: From Visual Uncertainty to Grounded Feedback



Question. *Is the green line shorter than the red line in the image? Yes or No*

Ground Truth: `no` **Candidate Response:** `no`

Before CC-OPSD Training.

Supervisory Feedback.

> Got it, let's check the image. The green line is the top one, which has two short lines on each end (like a Y shape with the horizontal part). [...] Wait, maybe the green line's horizontal segment is shorter than the red's? **Wait, no, maybe they are the same length.** [...]
>
> **Wait, maybe the green line is shorter.** [...]
>
> **Wait, but maybe the green line is longer. Wait, this is tricky.** [...]

Final Output: `yes`

After CC-OPSD Training.

Supervisory Feedback.

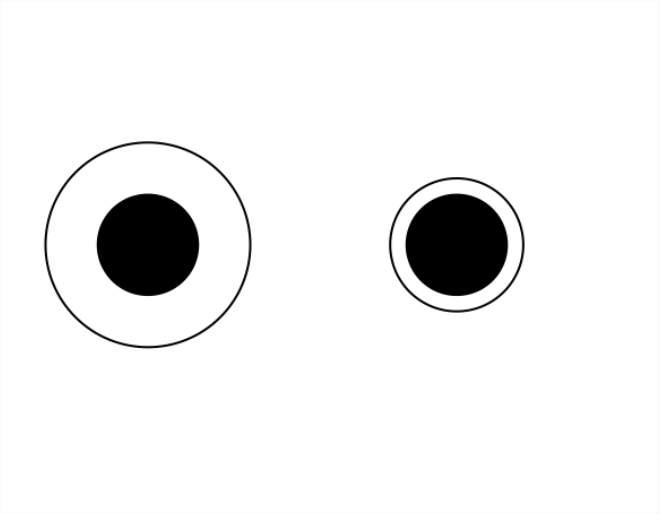
> The draft answer "no" is supported by the visual evidence. The green line and red line in the image have **equal horizontal lengths** (the main horizontal segments are identical in length), so the green line is **not shorter than the red line**. The draft answer correctly reflects this conclusion.
>
> **Keep the draft answer unchanged.**

Final Output: `no`

Analysis. Before training, the supervisory circuit **oscillates among mutually inconsistent visual judgments** and fails to converge to a stable supervisory decision. Although the candidate response is already correct, the incomplete feedback fails to preserve it, resulting in an incorrect final output of `yes`. After CC-OPSD training, the circuit **correctly identifies the two horizontal segments as equal in length** and explicitly concludes that the green line is not shorter. It then provides a **clear preservation instruction** to keep the correct candidate response unchanged, yielding the correct final output `no`. This comparison shows that CC-OPSD improves not only corrective supervision but also **the stability and selectivity of preservation behavior**, allowing the circuit to avoid unnecessary modification when the cognitive core is already correct.

Failure Cases. We further present two representative failure cases to characterize the remaining limitations of the trained supervisory circuit. The first arises from an illusion-inducing visual pattern that leads to an incorrect perceptual judgment, while the second reflects an overly strict interpretation of the semantic condition in the question. These examples reveal two failure modes in visual grounding and response interpretation.

Failure Case 1: Susceptibility to Visual Illusion



Question. *Is the black solid circle on the left the same size as the black solid circle on the right?*

Ground Truth: `yes` **Candidate Response:** `no`

Supervisory Feedback.

> The draft answer "no" is supported by the visual evidence. The black solid circle on the left is **visibly larger than the black solid circle on the right**, confirming they are not the same size. **Keep the draft answer unchanged.**

Final Output: `no`

Analysis. The supervisory circuit is itself **affected by the visual illusion induced by the surrounding circles**. It interprets the **apparent size difference as an actual difference** and therefore incorrectly preserves the erroneous candidate response. This failure indicates that the learned supervision remains constrained by **challenging perceptual cases in which the underlying visual representation is misleading**. While CC-OPSD improves hallucination-aware grounding, it cannot fully eliminate errors caused by difficult low-level visual perception.

Failure Case 2: Overly Strict Semantic Interpretation



Question. *According to the positive sequence images, does Homer Simpson disappear into the bushes?*

Ground Truth: `yes` **Candidate Response:** `yes`

Supervisory Feedback.

The draft answer is not supported by the visual evidence. The sequence shows Homer becoming **progressively less visible** (partially obscured, then mostly hidden), but the final frame reveals only a small portion of his head (**not his complete disappearance**). The visual sequence indicates that the draft answer incorrectly assumes **full disappearance**, while the visual evidence shows he is not fully vanished. **Replace "yes" with "no"** to accurately reflect the sequence.

Final Output: `no`

Analysis. The supervisory circuit **applies an overly strict interpretation of "disappear into the bushes,"** requiring the subject to become **completely invisible**. Under the intended semantics of the benchmark, **progressively entering the bushes and becoming almost fully occluded** is sufficient to satisfy this description. The circuit therefore **unnecessarily intervenes on an already correct candidate response** and changes `yes` to `no`. This case exposes a residual **semantic-boundary error**: although the relevant visual progression is correctly perceived, the supervision can still fail when the decision depends on how a natural-language condition is interpreted.

Table 4: Computational efficiency comparison on a single NVIDIA RTX 3090 GPU. Latency denotes the average inference time per sample, and memory denotes the peak allocated GPU memory.

Method	Latency ↓	Tokens/s ↑	Peak Memory (GiB) ↓
Vanilla	28.84	25.51	16.90
DnR	30.51	23.47	18.70
CogRef	28.73	25.40	16.89

F Computational Efficiency.

We further evaluate the inference efficiency on a single NVIDIA RTX 3090 GPU. As shown in Table 4, CogRef achieves an average latency of $\mathbf{28.73}$ s per sample and a throughput of $\mathbf{25.40}$ tokens/s, closely matching Vanilla. Its peak allocated memory also remains essentially unchanged, indicating that the learned hallucination-aware supervision introduces negligible additional memory overhead at inference. Compared with DnR, CogRef requires shorter per-sample inference time, while maintaining higher token-level throughput and lower peak GPU memory usage. Overall, these results show that the additional supervisory capability of CogRef can be incorporated without materially increasing the inference footprint over the vanilla supervisory design.

G LIMITATIONS AND FUTURE WORK

Despite the consistent improvements across model families and benchmarks, two aspects remain open for further study. First, the supervisory circuit is learned from a finite pool of model-specific experience, so its behavior is still affected by the coverage of the constructed hallucination and preservation cases. Our current setting deliberately uses a compact experience pool; extending it to more diverse error patterns and visual domains may further improve the coverage of supervisory behavior. Second, the learned circuit can still be challenged by intrinsically ambiguous visual evidence or subtle semantic boundaries. As illustrated by the failure cases in *Appendix* E, visual illusions may lead to incorrect grounding, while overly strict interpretations of natural-language conditions can occasionally trigger unnecessary intervention. Incorporating more diverse borderline cases may help improve robustness in such situations.