VISUAL-EVIDENCE-GUIDED LATENT CORRECTION FOR HALLUCINATION MITIGATION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) remain prone to generating hallucinated content inconsistent with the input image during autoregressive generation. Recent online self-correction paradigms integrate verification and correction into the generation process, enabling models to roll back and regenerate upon detecting hallucinations. This addresses the limitations of conventional generation adjustment, which struggles to revise existing errors, and post-hoc verification, which relies on post-generation processing. However, existing online correction mechanisms mainly rely on sampling and text-conditioned adjustments after rollback, while lacking an explicit mechanism to incorporate visual evidence into the correction process after rollback. To address this issue, we propose Visual-Evidence-Guided Latent Correction (VE-LC). After hallucination detection and rollback, but before text regeneration, VE-LC performs Latent Visual Correction (LVC), which conducts reasoning in continuous latent space to re-integrate visual information before resuming standard autoregressive generation. We further introduce Local Hallucination Correction Supervision (LHCS), which strengthens local hallucination correction through edited-token reweighting and a visual candidate contrastive constraint. Experiments on six visual hallucination and general multimodal benchmarks and three LVLM architectures show that VE-LC can effectively reduce visual hallucinations while maintaining content coverage and general multimodal capabilities.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have demonstrated strong capabilities in visual understanding and language generation across tasks such as image captioning, visual question answering, multimodal reasoning, and visual dialogue Liu et al. (2023); Li et al. (2023a); Bai et al. (2023); Chen et al. (2024). Given visual inputs and user prompts, LVLMs can generate fluent and semantically coherent natural-language responses. However, these models remain prone to generating content inconsistent with the input image, a phenomenon known as visual hallucination, such as fabricating nonexistent objects or incorrectly predicting object attributes, counts, spatial relationships, and other visual details Guan et al. (2024); Park et al. (2025); Jiang et al. (2025). Since such errors are often expressed in fluent and coherent language, they can be difficult to identify from the text alone, substantially undermining the reliability of LVLMs. Therefore, mitigating visual hallucinations has become an important problem in LVLM research.

Existing studies mitigate visual hallucinations at different stages of generation. One line of work reduces visually unsupported hallucination content through model training, prediction distribution adjustment, or search strategies Leng et al. (2024); Wang et al. (2024b); Sarkar et al. (2025); Yu et al. (2024; 2025); Liu et al. (2024a), while another revises existing errors after generation through visual verification, rewriting, or post-processing Yin et al. (2024); Zhou et al. (2024b). Nevertheless, these approaches still face a fundamental limitation: generation, verification, and correction are typically performed in separate stages or independent modules, making it difficult to rectify errors once they arise during generation, while post-generation verification and revision often rely on additional models and complex pipelines. Zhong et al. (2024).

Recently, REVERSE Wu et al. (2025) introduced the first framework that unifies generation, verification, and correction within a single LVLM. As shown in Fig. 1(a), when unreliable content

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.

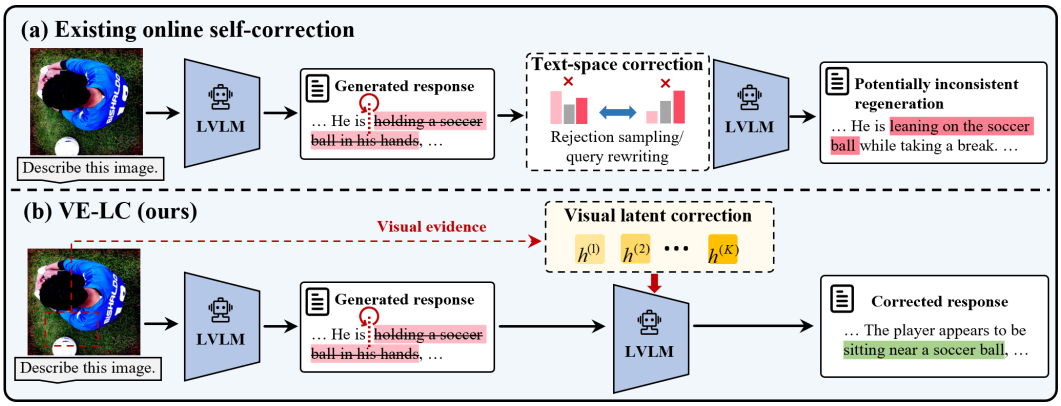


Figure 1: (a) Existing online self-correction performs rollback and regenerates the continuation mainly based on textual information. (b) VE-LC explicitly re-integrates visual evidence through latent correction before resuming autoregressive generation.

is detected, the model rolls back to a previous reliable position and revises subsequent generation through rejection sampling and query rewriting, enabling online and iterative self-correction during autoregressive generation. This design enables LVLMs to jointly perform generation, verification, and correction, providing a more integrated solution for visual hallucination mitigation. However, the correction process of REVERSE mainly focuses on regenerating text after rollback, with limited further exploitation of visual evidence from the input image. Since visual hallucination fundamentally arises from the inconsistency between generated content and visual evidence from the input image, a natural question arises: after hallucination detection and rollback, how can visual evidence be explicitly leveraged to guide the subsequent correction process?

To this end, we propose the Visual-Evidence-Guided Latent Correction framework (VE-LC). As shown in Fig. 1(b), after detecting and rolling back visual hallucinations, VE-LC reintegrates visual evidence from the input image through Latent Visual Correction (LVC) and performs multi-step latent-space reasoning to adjust the continuous representations supporting subsequent generation. The model then resumes standard autoregressive generation. We further introduce Local Hallucination Correction Supervision (LHCS), which strengthens local correction through edited-token reweighting and a visual candidate contrastive constraint. Extensive experiments on six visual hallucination and general multimodal benchmarks, together with evaluations across three LVLM architectures, demonstrate that VE-LC effectively reduces visual hallucinations while preserving content coverage and general multimodal capabilities.

Our main contributions are as follows: (1) identifying insufficient re-integration of visual evidence after rollback as an overlooked limitation of online self-correction; (2) proposing VE-LC, which combines LVC for visual evidence re-integration with LHCS for local hallucination correction; and (3) demonstrating its effectiveness across six benchmarks and its generalizability across three LVLM architectures, while preserving content coverage and general multimodal capabilities.

2 Related Work

2.1 Visual Hallucination Mitigation in Large Vision-Language Models

Existing studies on visual hallucinations in LVLMs mainly focus on suppressing hallucinations before or during generation and correcting errors after generation. The former includes training with hallucination-aware data or preference signals Zhou et al. (2024a); Zhao et al. (2024), as well as intervening in the generation process by adjusting token distributions, search strategies, or internal representations Leng et al. (2024); Wang et al. (2024b); Huang et al. (2024); Lin et al. (2026). The latter revises generated hallucinations through visual verification, rewriting, or post-processing Yin et al. (2024); Zhou et al. (2024b). Recently, REVERSE Wu et al. (2025) further integrates verification, rollback, and regeneration into a single LVLM, enabling online self-correction during the generation process. Building upon this unified correction framework, we further explore how to

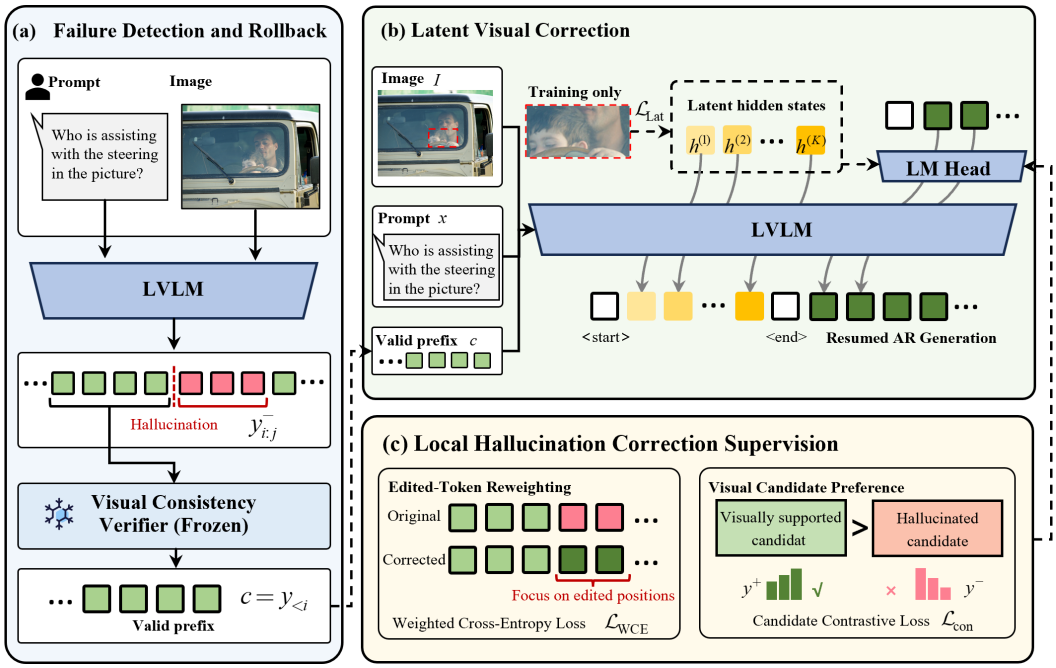


Figure 2: Overview of the proposed VE-LC framework.

incorporate visual evidence into the post-rollback correction process to alleviate the inconsistency between generated content and the input image.

2.2 Latent-Space Reasoning in Multimodal Models

Latent-space reasoning performs intermediate computation over continuous hidden representations, reducing reliance on explicit textual reasoning traces Deng et al. (2023); Shen et al. (2025); Yue et al. (2025). Recently, this idea has been extended to multimodal models. LVR extends latent reasoning to the visual modality to enhance multimodal reasoning capability Li et al. (2026). Subsequently, ILVR and Monet explore dynamic visual latent representations and continuous visual reasoning to enhance visual perception and multi-step reasoning Dong et al. (2026); Wang et al. (2026). In this work, we further investigate the use of latent-space computation for visual hallucination correction.

3 Method

We propose Visual-Evidence-Guided Latent Correction (VE-LC) for online visual hallucination correction during generation. As shown in Fig. 2, when the visual consistency verifier detects visually unsupported content, VE-LC first rolls back to the valid text prefix before the hallucination and truncates the corresponding KV cache. Latent Visual Correction (LVC) then performs continuous latent-space computation while preserving the complete multimodal context, and uses the resulting logit residual to adjust token prediction distribution. In addition, we introduce Local Hallucination Correction Supervision (LHCS), which strengthens the training signals at actual edit positions and distinguishes visually supported from unsupported candidates to further improve local correction.

3.1 Latent Visual Correction

Given the hallucination onset position i identified by the visual consistency verifier, we first roll back the generated sequence to the valid text prefix before the hallucination, $c = y_{<i}$. Meanwhile, the model truncates the KV cache corresponding to the hallucinated tokens and retains only the multimodal prefix cache associated with the image input and valid text prefix, denoted as $\mathcal{K}_c$. This removes both the hallucinated content and its cached states while preserving the visual and textual

context required for subsequent correction. Subsequently, LVC performs continuous latent-space computation over $\mathcal{K}_c$.

Given the visual representations of the input image, $V = \{v_1, v_2, \ldots, v_{T_v}\}$, we denote the hidden state at the current position after rollback as $h^{(0)}$. LVC feeds the continuous hidden state back into the model and performs K recursive latent-space updates: $h^{(k)} = \mathcal{F}_\theta\left(h^{(k-1)}; \mathcal{K}_c\right)$, $k = 1, \ldots, K$, where $\mathcal{F}_\theta$ denotes the continuous hidden-state update of the LVLM under the current multimodal context. The resulting latent-state sequence is denoted as $H_{\text{Lat}} = \left\{h^{(1)}, h^{(2)}, \ldots, h^{(K)}\right\}$.

Unlike standard autoregressive decoding, these continuous states are not written into the generation history as actual text tokens, but are used only as intermediate representations for latent-space computation. Thus, before text regeneration, LVC can adjust the current prediction using the visual information retained in the multimodal prefix cache.

To encourage latent-space computation to learn correction directions aligned with visual evidence, we further introduce visual alignment constraint during training. For correction samples with local region information, we use the corresponding ROI patches to construct the visual target; otherwise, we use the full-image patches. During training, visual targets are obtained from available region annotations or full-image visual features. The target visual representations are denoted as $V^{\text{tar}} = \left\{v_1^{\text{tar}}, v_2^{\text{tar}}, \ldots, v_{T_r}^{\text{tar}}\right\}$. We map the latent states into the target visual representation space: $\widetilde{V} = \mathcal{G}\left(H_{\text{Lat}}\right) = \{\widetilde{v}_1, \widetilde{v}_2, \ldots, \widetilde{v}_{T_r}\}$, and define the latent visual alignment loss as

$$\mathcal{L}_{\text{Lat}} = \frac{1}{T_r} \sum_{j=1}^{T_r} \left\| \widetilde{v}_j - v_j^{\text{tar}} \right\|_2^2 . \tag{1}$$

This objective encourages the latent representations to preserve correction-relevant visual information. The visual targets are used only during training; no additional ROI or bounding box is required at inference time.

After K latent-space updates, we use the prediction change before and after latent correction to guide subsequent text generation. Let $z_t^{(0)}$ and $z_t^{(K)}$ denote the prediction logits corresponding to the initial state $h^{(0)}$ and final state $h^{(K)}$, respectively. Their difference defines the latent correction residual:

$$r_t = \mathcal{N}\left(z_t^{(K)} - z_t^{(0)}\right), \tag{2}$$

where $\mathcal{N}(\cdot)$ denotes a normalization operation. We then add this residual to the base model logits: $\widetilde{z}_t = z_t^{\text{base}} + \alpha r_t$, where α controls the strength of the latent correction signal. The final generation distribution is

$$p_\theta^{\text{corr}}\left(y_t \mid V, x, y_{<t}\right) = \text{Softmax}\left(\widetilde{z}_t\right). \tag{3}$$

The actual text token sampled from this distribution is written into the main KV cache following standard autoregressive decoding, while the temporary states produced by latent-space computation are excluded from the generation history. After the current correction is completed, the model resumes standard autoregressive generation. If the visual consistency verifier detects visually unsupported content again, the rollback and latent correction procedure is triggered again.

3.2 Local Supervision for Visual Hallucination Correction

Visual hallucinations are often concentrated in a small number of object, attribute, or relation tokens within the generated sequence. However, standard sequence-level supervision assigns equal optimization weight to all positions, causing local error signals to be diluted by the large number of unchanged tokens. Moreover, simply increasing the generation probability of the correct correction does not directly reduce the relative preference for visually unsupported candidates. Therefore, we strengthen local hallucination correction from two aspects: edited tokens and candidate preference.

Edited-Token Reweighting. For hallucination correction samples, we first perform token-level alignment between the original output and the corrected target to identify the set of actually modified positions, denoted as $\mathcal{S}_{\text{edit}}$. We define the supervision weight according to whether a token lies in the edited region:

$$w_t = \begin{cases} w_{\text{edit}}, & t \in \mathcal{S}_{\text{edit}}, \\ 1, & t \notin \mathcal{S}_{\text{edit}}. \end{cases} \tag{4}$$

Based on this, we define the edited-token weighted cross-entropy loss as

$$\mathcal{L}_{\mathrm{WCE}} = -\left(\sum_{t=1}^{T_y} w_t\right)^{-1} \sum_{t=1}^{T_y} w_t \log p_\theta(y_t^* \mid V, x, H_{\mathrm{Lat}}, y_{<t}^*)\,. \tag{5}$$

where y_t^* denotes the t-th token in the target sequence. This objective preserves full-sequence supervision while increasing the contribution of actual edit positions to optimization.

Visual Candidate Contrastive Constraint. To further distinguish visually supported from visually unsupported generations, we construct a visually supported candidate y^+ and a visually unsupported candidate y^- under the same image, textual context, and correction position. The visually supported and unsupported candidates are scored by their average token log-likelihoods:

$$s^+ = \frac{1}{L^+}\sum_{t=1}^{L^+} \log p_\theta\big(y_t^+ \mid V, x, H_{\mathrm{Lat}}, y_{<t}^+\big)\,,\ s^- = \frac{1}{L^-}\sum_{t=1}^{L^-} \log p_\theta\big(y_t^- \mid V, x, H_{\mathrm{Lat}}, y_{<t}^-\big)\,, \tag{6}$$

where L^+ and L^- denote the token lengths of the positive and negative candidate segments, respectively.

We then adopt a margin-based pairwise ranking loss:

$$\mathcal{L}_{\mathrm{con}} = \mathrm{softplus}\left(\gamma_{\mathrm{con}} - \left(s^+ - s^-\right)\right), \tag{7}$$

where γ_{con} denotes the ranking margin. By enforcing a higher relative score for the visually supported candidate than for the hallucinated one, this objective not only increases the conditional likelihood of the correct correction segment but also explicitly reduces the relative preference for visually unsupported candidates.

Combining the above constraints, the final training objective is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{WCE}} + \lambda_{\mathrm{Lat}}\mathcal{L}_{\mathrm{Lat}} + \lambda_{\mathrm{con}}\mathcal{L}_{\mathrm{con}}. \tag{8}$$

where λ_{Lat} and λ_{con} control the contributions of latent visual alignment and the visual candidate contrastive constraint, respectively.

4 Experiments 4,500

We systematically evaluate the proposed method on various visual hallucination tasks. We first compare VE-LC with existing methods on multiple visual hallucination benchmarks and further examine its generalization across different LVLMs. We then conduct controlled ablations to analyze the individual contributions of latent visual correction and different training objectives, followed by further analyses of latent representation changes, inference efficiency, and qualitative examples.

4.1 Experimental Setup

Models and Implementation Details. We mainly adopt LLaVA-v1.5-7B Liu et al. (2024b) as the base model for the main experiments and ablation studies. To evaluate the cross-architecture applicability of VE-LC, we further conduct experiments on Qwen2.5-VL-7B Bai et al. (2025) and InternVL3-1B Zhu et al. (2025). The training data are constructed from a formatted Visual CoT corpus Shao et al. (2024), resulting in 4,948 samples, including 3,900 for training, 494 for validation, and the remaining 554 as a held-out test set. The splits are disjoint at both the image and request levels, and all external evaluation sets are excluded from training and model selection. By default, LVC performs 8 continuous latent-space updates. During training, we freeze the base model parameters and optimize only the LoRA adapters Hu et al. (2021), with a learning rate of 1×10^{-5}. All experiments are conducted on an NVIDIA A40 GPU with 48 GB memory. Additional details on data construction and implementation are provided in Appendix A.

Benchmarks and Evaluation Metrics. We evaluate the effectiveness of VE-LC from two aspects: visual hallucination mitigation and general multimodal capabilities. We adopt AMBER Wang et al. (2024a) and CHAIR Rohrbach et al. (2018) to evaluate generative visual hallucinations; adopt

Table 1: Generative object hallucination evaluation results.

Methods	AMBER				CHAIR	
	Chair ↓	Cover ↑	HalRate ↓	Cog ↓	Chair_s ↓	Chair_i ↓
LLaVA-v1.5-7B	11.2	49.9	47.9	4.7	49.8	15.5
+VCD	9.2	**51.7**	41.6	4.5	55.4	16.6
+ICD	7.7	<u>50.7</u>	35.1	3.7	49.2	15.4
+OPERA	6.3	48.8	26.7	2.9	48.6	14.5
+HA-DPO	10.2	49.5	43.2	3.5	39.8	11.7
+Woodpecker	12.3	47.9	43.2	4.1	41.2	<u>8.7</u>
+REVERSE	9.0	49.5	40.5	3.8	<u>34.0</u>	9.9
+Hulluedit	<u>4.3</u>	43.8	<u>20.9</u>	<u>1.6</u>	52.6	14.1
+**VE-LC (Ours)**	**4.1**	47.2	**16.8**	**1.0**	**29.4**	**8.4**

POPE Li et al. (2023b) to evaluate object-existence hallucinations, and adopt MME Fu et al. (2025) and MMBench Liu et al. (2024c) to evaluate general visual perception and multimodal reasoning capabilities; and adopt MMHal-Bench Sun et al. (2024) to evaluate hallucination issues in open-ended visual question answering. These benchmarks have been widely adopted by the community and cover multiple evaluation dimensions for visual hallucination mitigation and general multimodal understanding. Specifically, on AMBER, we report CHAIR (object hallucination), Cover (content coverage), HalRate (hallucination rate), and Cog (cognitive ability); on CHAIR, we report CHAIR_s (sentence-level hallucination) and CHAIR_i (instance-level hallucination); on POPE, we report Accuracy (accuracy), Precision (precision), and F1 (comprehensive evaluation metric); on MME, we report scores on Existence, Count, Position, and Color, as well as the Sum (overall score) metric; on MMBench, we report Accuracy; and on MMHal-Bench, we report Overall Score and Hallucination Rate.

Baselines. We compare VE-LC with representative visual hallucination mitigation methods. Specifically, VCD Leng et al. (2024), ICD Wang et al. (2024b), and OPERA Huang et al. (2024) mitigate hallucinations through decoding interventions at inference time; HA-DPO Zhao et al. (2024) improves visual alignment through training; Woodpecker Yin et al. (2024) corrects hallucinations through visual verification and post-generation revision; REVERSE Wu et al. (2025) combines online verification, rollback, and regeneration for self-correction during generation; and Hulluedit Lin et al. (2026) is included as a recent visual hallucination mitigation method.

4.2 Main Results

4.2.1 VE-LC Effectively Mitigates Visual Hallucinations while Preserving General Multimodal Capabilities

We first compare VE-LC with representative visual hallucination mitigation methods on LLaVA-v1.5-7B. To comprehensively evaluate its effectiveness, we consider generative object hallucination, discriminative visual hallucination, general multimodal capabilities, and open-ended visual question answering. Overall, VE-LC consistently improves performance across different hallucination evaluations while preserving or even improving the general multimodal capabilities of the base model, indicating that visual-evidence-guided correction reduces visually inconsistent generation without substantially compromising the model's original capabilities.

Generative Object Hallucination. Table 1 presents the results on AMBER and CHAIR. Overall, VE-LC demonstrates strong hallucination mitigation on both generative object hallucination benchmarks while maintaining competitive content coverage. Notably, existing methods do not always perform consistently across different generative evaluations: some methods effectively reduce hallucinations on one benchmark, but their advantages do not reliably transfer to another. In contrast, VE-LC achieves consistent improvements on both AMBER and CHAIR, indicating that its mitigation of object hallucinations is not tied to a specific evaluation setting. Furthermore, compared with REVERSE, the online self-correction method most closely related to our work, VE-LC achieves fur-

Table 2: Discriminative visual hallucination and general multimodal capability evaluation results.

Methods	POPE			MME					MMB
	Acc. ↑	Pre. ↑	Rec. ↑	Exi. ↑	Cou. ↑	Pos. ↑	Col. ↑	Sum ↑	↑
LLaVA-v1.5-7B	81.66	88.76	80.10	179.0	133.4	113.0	144.1	569.5	64.30
+VCD	84.06	90.25	82.99	190.8	115.5	111.5	135.8	553.6	72.61
+ICD	83.70	92.36	82.09	181.2	108.0	119.4	146.9	555.5	73.02
+OPERA	84.88	93.86	83.38	**195.1**	148.3	132.7	155.7	631.7	73.36
+HA-DPO	80.37	90.14	77.98	185.5	81.1	82.4	138.1	487.1	70.34
+Woodpecker	75.44	**95.69**	68.89	110.7	102.1	53.2	139.4	405.3	62.20
+REVERSE	86.80	93.02	**85.96**	194.5	138.3	113.4	158.5	604.7	68.04
+Hulluedit	86.75	93.73	85.78	190.8	156.3	134.2	**167.5**	648.7	73.68
+VE-LC (Ours)	**86.87**	93.93	85.87	191.1	**159.8**	**137.9**	163.8	**652.4**	**73.73**

Table 3: Generalization results of VE-LC across different LVLM architectures.

Methods	AMBER				CHAIR	
	Chair ↓	Cover ↑	HalRate ↓	Cog ↓	Chair_s ↓	Chair_i ↓
Qwen2.5-VL-7B	5.2	**64.3**	28.5	1.8	34.8	9.70
+VE-LC (Ours)	**4.7**	64.1	**27.2**	**1.6**	**29.0**	**7.79**
InternVL3-1B	7.4	**71.9**	56.6	5.0	33.0	9.5
+VE-LC (Ours)	**4.6**	62.6	**28.9**	**2.6**	**18.6**	**5.9**

ther improvements on both benchmarks. Since both methods can detect hallucinations and perform rollback during generation, this difference further suggests that simply searching for an alternative continuation is insufficient for effective correction. Re-integrating visual evidence from the input image after rollback provides stronger visual grounding for subsequent generation.

Discriminative Hallucination and General Multimodal Capabilities. Beyond free-form generation, we further evaluate VE-LC on discriminative visual hallucination and examine its impact on general multimodal capabilities. The results are shown in Table 2. On POPE, VE-LC demonstrates strong visual existence judgment, indicating that its hallucination mitigation generalizes from free-form generation to direct discrimination of image content. Detailed results under different POPE settings are provided in Appendix B. On the more comprehensive MME-Hallucination benchmark, the advantage of VE-LC further extends from object existence judgment to finer-grained visual attribute understanding, achieving the best overall performance. Meanwhile, VE-LC also maintains the best performance on MMBench. These results show that VE-LC mitigates different types of visual hallucinations while preserving general multimodal understanding, without compromising the model's original capabilities.

Open-Ended Visual Question Answering. We further evaluate VE-LC on open-ended visual question answering using MMHal-Bench. The results are shown in Fig. 3. Overall, VE-LC achieves the highest overall response score while maintaining a low hallucination rate. Although Hulluedit yields a slightly lower hallucination rate, VE-LC achieves better overall response quality, indicating a favorable balance between hallucination mitigation and response quality. These results further demonstrate that the effectiveness of VE-LC extends from object-level hallucination mitigation to more complex open-ended multimodal generation.

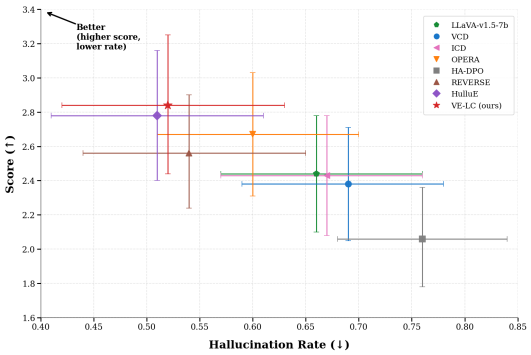


Figure 3: MMHal-Bench score versus hallucination rate.

Table 4: Controlled ablation results for Latent Visual Correction.

Method	Sem. Acc.↑	Sem./Part.↑	Token F1↑	Length MAE↓
Base Qwen	0.020	0.100	0.070	14.62
+LVC	**0.245**	**0.615**	**0.432**	**2.24**

Table 5: Incremental contributions of key components on AMBER.

Method	Chair↓	Cover↑	HalRate↓	Cog↓
LLaVA-v1.5-7B	11.2	49.9	47.9	4.7
+LVC	5.9	**51.4**	29.0	3.0
+VE-LC (Ours)	**4.1**	47.2	**16.8**	**1.0**

4.2.2 VE-LC Generalizes across Different Vision-Language Models

To examine the applicability of VE-LC across different model architectures, we further apply it to Qwen2.5-VL-7B and InternVL3-1B. The results are shown in Table 3. VE-LC consistently mitigates hallucinations on both additional LVLM backbones, with similar improvement trends on AMBER and CHAIR. On Qwen2.5-VL-7B, hallucinations are reduced while content coverage remains largely stable. On InternVL3-1B, VE-LC achieves more pronounced hallucination suppression, while the generated content becomes more conservative. Together with the results on LLaVA-v1.5-7B, these experiments show that VE-LC is applicable to different model families and parameter scales, demonstrating good cross-model generalization.

4.3 Ablation and Analysis

Feasibility of Latent-Space Visual Reasoning To verify whether visual evidence can effectively support local correction in latent space, we conduct a controlled evaluation on a fixed set of 200 stratified samples. The semantic metric measures the consistency between the corrected output and the reference answer, Token F1 measures token-level matching, and Length MAE measures the deviation of the local generation boundary from the reference; detailed definitions are provided in Appendix C. As shown in Table 4, introducing latent-space visual reasoning substantially improves both semantic consistency and token-level matching, while significantly reducing generation boundary deviation. These results indicate that visual evidence can provide effective guidance for local correction of erroneous segments in latent space, validating the feasibility of this mechanism for subsequent visual correction.

Incremental Contributions of Key Components To analyze the contributions of LVC and LHCS to overall hallucination mitigation, we conduct progressive ablations on LLaVA-v1.5-7B by introducing the two components sequentially. The results are shown in Table 5. Introducing LVC alone substantially reduces visual hallucinations while improving content coverage, indicating that post-rollback latent visual correction provides the primary hallucination mitigation benefit without relying on reduced generation. Adding LHCS further improves hallucination-related metrics, but slightly decreases content coverage. This suggests that LHCS further strengthens post-correction visual consistency on top of LVC, while introducing a trade-off between content coverage and visual reliability. Overall, LVC provides the main visual correction capability, while LHCS further strengthens hallucination control, with the two components showing clear incremental contributions to hallucination mitigation.

4.4 Efficiency and Qualitative Analysis

Inference Efficiency We further compare the end-to-end inference efficiency of different methods, with the detailed evaluation setup provided in Appendix D. As shown in Table 6, VE-LC achieves an average inference latency comparable to VCD and substantially lower than OPERA, Woodpecker, and REVERSE. This is consistent with the on-demand correction mechanism of VE-LC, where additional computation is triggered only when visual inconsistency is detected. Accordingly, VE-LC incurs a higher P95 latency, while its overall inference overhead remains low.

Table 6: Comparison of inference efficiency across different methods.

Method	Latency (s)↓	P95 (s)↓	Tokens/s↑	Tokens
LLaVA-v1.5-7B	3.80 (1.00×)	4.63	30.04	114.07
+VCD	16.65 (4.39×)	16.70	15.36	255.79
+ICD	7.82 (2.06×)	10.49	15.26	119.25
+OPERA	37.79 (9.95×)	64.30	2.86	108.03
+Woodpecker	115.17 (30.33×)	425.17	8.18	942.09
+REVERSE	58.90 (15.51×)	173.63	1.92	113.05
+VE-LC (Ours)	16.62 (4.38×)	34.57	6.93	115.24

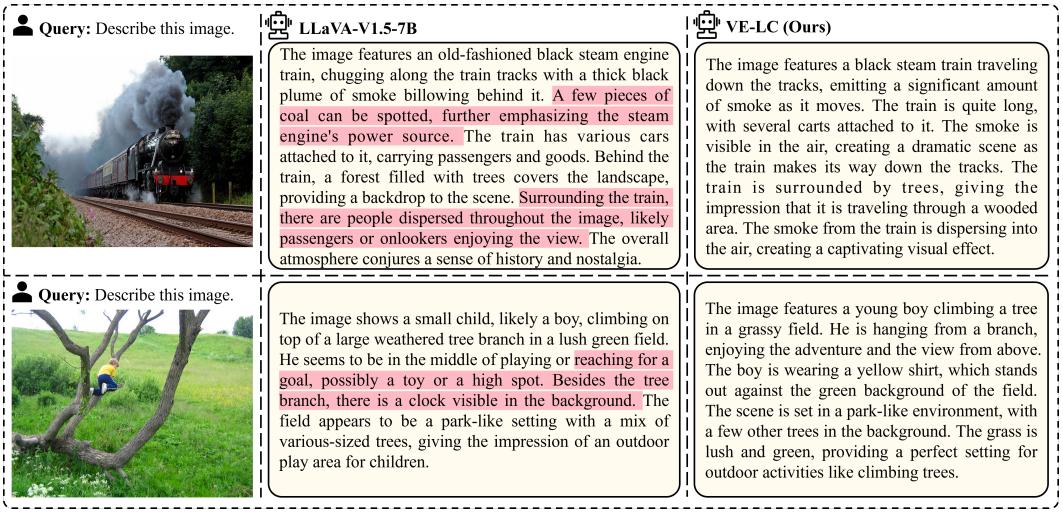


Figure 4: Qualitative Correction Examples

Qualitative Analysis. Figure 4 presents qualitative comparisons between VE-LC and the original LLaVA-V1.5-7B. The baseline model tends to introduce hallucinated content without sufficient visual support, whereas VE-LC effectively suppresses such hallucinations while preserving valid descriptions of the main objects, attributes, and scenes in the image. These examples further show that VE-LC can correct local visual errors without substantially weakening the remaining visually grounded content.

5 Conclusion and Limitations

This paper studies online visual hallucination correction in LVLMs and identifies insufficient reuse of visual evidence after rollback. We propose VE-LC, which combines LVC for latent-space visual re-integration before token regeneration with LHCS for stronger local correction. Experiments on six visual hallucination and general multimodal benchmarks across three LVLM architectures show that VE-LC reduces visual hallucinations while preserving competitive content coverage and general multimodal capabilities, demonstrating its effectiveness and cross-model generalization.

This work has two limitations. VE-LC relies on a visual consistency verifier, whose detection or localization errors may affect correction. In addition, frequent rollback and latent-space correction can reduce generation efficiency despite manageable overall overhead. Improving the reliability and efficiency of online correction remains future work.

AI use statement

In this work, we used generative AI tools for the retrieval and discovery of related work, and for research ideation or execution. We did not use generative AI tools for aiding or polishing the writing,

drafting sections of the paper, generating synthetic datasets, or proving mathematical claims. We reviewed all AI-assisted work. Specifically, we used generative AI tools to retrieve relevant reference URLs, and we manually searched for and curated the corresponding BibTeX entries for citation. We used generative AI tools to develop parts of the experimental code, and two authors verified and tested its correctness. We used AI-assisted procedures combined with manual verification to validate and evaluate part of the data. In the MMHal-Bench benchmark evaluation, we used the official settings to call APIs for model evaluation. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement

This work aims to ensure the reproducibility of the proposed method and experimental results. The overall framework, core algorithm, and training objectives of VE-LC are introduced in the method section, while the model, datasets, evaluation benchmarks, and experimental settings are described in the experimental section. The appendix further provides the data construction process, implementation details, and training configurations. In addition, we provide an anonymous code repository containing the implementation code and data processing pipeline to support the verification and reproduction of our results: `https://anonymous.4open.science/r/VE-LC-C8A0`.

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, June 2024.

Yuntian Deng, Kiran Prasad, Roland Fernandez, Paul Smolensky, Vishrav Chaudhary, and Stuart Shieber. Implicit chain of thought reasoning via knowledge distillation, 2023.

Shuai Dong, Siyuan Wang, Xingyu Liu, Chenglin Li, Haowen Hou, and Zhongyu Wei. Interleaved latent visual reasoning with selective perceptual modeling. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29316–29335, San Diego, California, United States, 2026. 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, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Does object grounding really reduce hallucination of large vision-language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2728–2742, Miami, Florida, USA, 2024. Association for Computational Linguistics.

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. 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)*, pp. 14375–14385, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

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

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014, June 2025.

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

Bangzheng Li, Ximeng Sun, Jiang Liu, Ze Wang, Jialian Wu, Xiaodong Yu, Emad Barsoum, Muhao Chen, and Zicheng Liu. Latent visual reasoning. In *International Conference on Learning Representations*, volume 2026, 2026.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 19730–19742. PMLR, 2023a.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. Vienna, Austria, 2025. Association for Computational Linguistics.

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, Singapore, 2023b. Association for Computational Linguistics.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. Hulluedit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11086–11095, June 2026.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 57689–57733, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024b.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024c.

Eunkyu Park, Minyeong Kim, and Gunhee Kim. Halloc: Token-level localization of hallucinations for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29893–29903, 2025.

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

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö. Arık, and Tomas Pfister. Mitigating object hallucination in mllms via data-augmented phrase-level alignment, 2025.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8612–8642. Curran Associates, Inc., 2024.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. Codi: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693, Suzhou, China, 2025. Association for Computational Linguistics.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024a.

Qixun Wang, Yang Shi, Yifei Wang, Yuanxing Zhang, Pengfei Wan, Kun Gai, Xianghua Ying, and Yisen Wang. Monet: Reasoning in latent visual space beyond image and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12030–12040, 2026.

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

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. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 65749–65777. Curran Associates, Inc., 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.

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

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19985–19995, 2025.

Zhenrui Yue, Bowen Jin, Huimin Zeng, Honglei Zhuang, Zhen Qin, Jinsung Yoon, Lanyu Shang, Jiawei Han, and Dong Wang. Hybrid latent reasoning via reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 5501–5530. Curran Associates, Inc., 2025.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2024.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011, Bangkok, Thailand, 2024. Association for Computational Linguistics.

Table 7: Detailed results on POPE under adversarial, random, and popular settings. ↑ indicates that higher values are better. The best and second-best results are highlighted in bold and underline, respectively.

Methods	Adversarial			Random			Popular			Average		
	Acc. ↑	Pre. ↑	F1 ↑	Acc. ↑	Pre. ↑	F1 ↑	Acc. ↑	Pre. ↑	F1 ↑	Acc. ↑	Pre. ↑	F1 ↑
LLaVA-v1.5-7B	79.12	83.28	77.71	83.45	92.85	82.10	82.40	90.16	80.50	81.66	88.76	80.10
+VCD	81.29	84.69	80.31	86.51	95.64	85.54	84.39	90.43	83.12	84.06	90.25	82.99
+ICD	81.55	87.66	79.91	85.44	96.38	84.09	84.11	93.03	82.26	83.70	92.36	82.09
+OPERA	83.28	<u>89.83</u>	81.77	86.02	<u>97.22</u>	84.71	85.35	**94.53**	83.66	84.88	<u>93.86</u>	83.38
+REVERSE	<u>84.78</u>	88.69	<u>83.97</u>	**88.27**	96.55	**87.57**	**87.35**	93.81	**86.33**	<u>86.80</u>	93.02	**85.96**
+VE-LC (Ours)	**85.10**	**89.90**	**84.20**	<u>88.20</u>	**97.40**	<u>87.30</u>	<u>87.30</u>	<u>94.50</u>	<u>86.10</u>	**86.87**	**93.93**	<u>85.87</u>

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

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025.

A Additional Experimental Details

Our task-specific training data are constructed from the formatted Visual CoT Shao et al. (2024) annotations. After filtering, the candidate corpus contains 359,126 samples covering 140,141 images. We sample 20,000 image–request pairs and use LLaVA-v1.5-7B to generate responses, which are further verified for visual consistency to construct correction and preservation samples. The resulting dataset contains 4,948 samples, including 3,082 correction samples and 1,866 preservation samples, and is split into 3,900 training, 494 validation, and 554 test samples with no image or request overlap across different splits.

For LLaVA-v1.5-7B, we freeze the base model and optimize LoRA adapters with rank 64 and scaling factor 128. We use a learning rate of 1×10^{-5} and an effective batch size of 16. Unless otherwise specified, LVC performs 8 latent update steps. The changed-token weight is set to 3, while λ_{lat} and λ_{con} are set to 0.05 and 0.25, respectively, with a contrastive margin of 0.5. For visual alignment, ROI features are used when region annotations are available; otherwise, full-image visual features are used.

B Detailed Results on POPE

Table 7 reports the detailed POPE results under adversarial, random, and popular settings. VE-LC consistently achieves strong performance across all three settings, indicating that its effectiveness is not restricted to a particular negative-sampling strategy. Although different methods perform best on individual metrics, VE-LC remains consistently competitive across different settings and achieves the best average accuracy and precision. These results further demonstrate the robustness of VE-LC in object-existence hallucination evaluation.

C Local Correction Evaluation Setup and Metrics

We sample 200 examples from an independent local correction set, covering errors in objects, attributes, colors, quantities, actions, relations, and positions. Each sample contains an image, an erroneous generation context, and a reference correction. The model generates a local correction from the image and erroneous context. All methods use identical samples and decoding settings.

Semantic Consistency. We use GPT-4.1-mini as a semantic judge. Given the original request, erroneous context, reference correction, and model output, it assigns one of three labels: *correct*, *partial*, or *incorrect*. Let y_i denote the label for sample i. Semantic Accuracy is

$$\mathrm{SemAcc} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(y_i = \mathrm{correct}\right). \tag{9}$$

Including partially correct outputs, Semantic/Partial Accuracy is

$$\mathrm{SemPart} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}\left(y_i \in \{\mathrm{correct}, \mathrm{partial}\}\right). \tag{10}$$

Token F1. Token F1 measures token-level overlap between the model output and reference correction. We lowercase text, remove punctuation, and collapse consecutive whitespace. Let p_i and r_i denote the prediction and reference, and $T(\cdot)$ the normalized token multiset. Precision and recall are

$$P_i = \frac{|T(p_i) \cap T(r_i)|}{|T(p_i)|}, \qquad R_i = \frac{|T(p_i) \cap T(r_i)|}{|T(r_i)|}. \tag{11}$$

The per-sample Token F1 is

$$F1_i = \frac{2P_iR_i}{P_i + R_i}, \tag{12}$$

and the final score is

$$\mathrm{TokenF1} = \frac{1}{N}\sum_{i=1}^{N} F1_i. \tag{13}$$

Length MAE. Length MAE measures the absolute token-length difference between the model output and reference correction:

$$\mathrm{LengthMAE} = \frac{1}{N}\sum_{i=1}^{N}\left||T(p_i)| - |T(r_i)|\right|. \tag{14}$$

D Inference Efficiency Evaluation Details

All methods are evaluated on a single NVIDIA A40 GPU with batch size $= 1$ using the same 100 test images. Before each run, we use five independent images for warm-up, and repeat the evaluation three times. The reported latency measures the complete inference process, including visual encoding, text generation, and method-specific verification, rollback, correction, or regeneration. Model loading, image I/O, and result writing are excluded from timing. We report the mean latency, P95 latency, generation throughput, and average output length.