MODEL-SUBSTITUTION RISK CONTROL: VERIFIED REPAIR OF VISUAL CLAIMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models can generate captions and answers that contain details not supported by the image. Removing such details reduces hallucination, but it can also discard information that could be recovered through a correct replacement. We introduce Model-Substitution Risk Control (MSRC), a training-free method that verifies visual claims and attempts to repair unsupported object claims before deleting them. MSRC retains supported claims and uses the frozen model to propose replacements for failed claims. Each replacement must pass fixed admissibility checks and independent visual verification before being accepted; otherwise, the unsupported claim is removed. The method follows a common claim-level repair process across image captioning and visual question answering, with task-specific handling of short answers. To balance hallucination reduction and information retention, MSRC selects its verification thresholds using held-out data and measures remaining errors relative to the original claim opportunities. Open-vocabulary replacements are assessed using benchmark annotations, lexical relations, and human-audited error estimates when automatic labels are unavailable. We evaluate MSRC across captioning, visual question answering, and hallucination benchmarks, reporting both claim-level risk and final-response quality.

1 INTRODUCTION

Large vision-language models (LVLMs) have made significant progress in image description, visual question answering, and multimodal generation. However, they often generate content that is not supported by the image. This visual hallucination remains a major challenge for reliable multimodal generation. In many cases, the model mentions objects that are not actually present, even though the response remains fluent and plausible. Such behavior was already observed in conventional image captioning (Rohrbach et al., 2018) and continues to affect modern LVLMs. POPE (Li et al., 2023b) further shows that object-presence judgments can be influenced by language priors and object co-occurrences rather than visual evidence alone. Hallucinations can also appear as incorrect attributes, unsupported details, false premises, or answers that cannot be justified by the available visual information (Jing et al., 2024; Wang et al., 2024b).

Existing methods reduce hallucination in different ways. Visual Contrastive Decoding suppresses predictions that are weakly grounded in the image (Leng et al., 2024). OPERA modifies decoding to reduce over-reliance on a small set of summary tokens (Huang et al., 2024), while HALC combines visual localization with contrastive decoding (Chen et al., 2024). Other approaches revise the generated response after detecting unreliable content. LURE performs hallucination-aware revision (Zhou et al., 2024), Woodpecker combines detection, verification, and correction (Yin et al., 2024), DeGF uses generative feedback for self-correction (Zhang et al., 2025), and REVERSE revisits unreliable generations through retrospective verification and resampling (Wu et al., 2025). These methods demonstrate that hallucinated content can often be detected or corrected rather than simply accepted. Identifying an unsupported claim raises a key question: *should it be deleted or corrected?* For example, if an image shows a cat on a sofa but the caption mentions a dog, deleting *dog* removes the hallucination but loses recoverable information. Replacing it with *cat* preserves that information, provided the correction is visually supported. This motivates our central principle: *re-grounding proposes a correction, while independent verification determines whether it is accepted.*

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.

We introduce *Model-Substitution Risk Control (MSRC)*, a training-free framework that selectively repairs object hallucinations through three actions: KEEP → SUBSTITUTE → DELETE. Each claim is verified using the frozen LVLM; supported claims are retained, while unsupported claims are re-grounded to propose replacements. A candidate is substituted only if it passes lexical and duplicate checks followed by independent visual re-verification. Otherwise, the original claim is deleted. Unlike verification-only filtering (VOF (Ours)), which deletes unsupported claims, and unconstrained self-correction, which may introduce new hallucinations, MSRC accepts only verified repairs. Since reducing hallucination through excessive deletion can also remove useful information, we measure risk over object-claim opportunities identified before repair, keeping the denominator fixed. We pair this with *true-claim yield* to measure supported information retained or recovered. For valid object names outside the benchmark vocabulary, MSRC uses a conservative human-audit mechanism rather than assuming they are correct or incorrect.

We formulate the complete keep–substitute–delete process as a risk-controlled policy, building on Conformal Risk Control (Angelopoulos et al., 2024) and Learn Then Test (Angelopoulos et al., 2025). While ConfLVLM (Li et al., 2025) controls factuality by filtering unsupported visual details, MSRC extends risk control to verified repair by explicitly accounting for newly generated replacements. Our risk formulation focuses on guard-eligible object claims in free-form generation. Additional evaluations examine the empirical performance of MSRC on visual question answering and discriminative hallucination benchmarks. A detailed discussion of related work is provided in Appendix A.

Our contributions are summarized as follows:

- We formulate LVLM object-hallucination mitigation as *verified selective substitution*: failed object claims are re-grounded, generated replacements are independently re-verified, and deletion is used only when no acceptable substitute is available.
- We develop a risk-control formulation for the complete keep–substitute–delete policy, including generated and open-vocabulary outputs, rather than certifying only retained original claims. The formulation explicitly couples hallucination risk with retained true-claim yield.
- We calibrate the two-dimensional repair policy using finite-grid threshold selection and independent held-out risk assessment. The operating point is selected to maximize retained true-claim yield subject to a conservative empirical risk criterion and is subsequently evaluated on a disjoint certification split.
- We evaluate MSRC against matched verification-only filtering (VOF (Ours)) and representative training-free, training-based, and self-correction baselines, reporting both hallucination and information-retention metrics to expose the risk–utility trade-off rather than rewarding trivial abstention.

2 METHODOLOGY

2.1 OVERVIEW

We consider a frozen large vision–language model (LVLM), g_θ, that generates a response $y = g_\theta(x, u)$ given an image x and prompt u. Following fine-grained hallucination analyses that decompose responses into individual visual facts (Jing et al., 2024; Li et al., 2025), MSRC operates at the claim level, preserving visually supported content while selectively repairing unsupported object claims. Each claim is first verified against the image. If verification fails, the LVLM proposes a visually grounded replacement, in the spirit of detect-then-correct pipelines (Yin et al., 2024; Zhou et al., 2024; Wu et al., 2025); unlike these pipelines, the replacement is accepted only after passing fixed admissibility guards and independent re-verification. Otherwise, the failed claim is deleted, as in verification-only filtering (Li et al., 2025). This yields the decision order KEEP → SUBSTITUTE → DELETE. MSRC separates repair generation from acceptance: re-grounding proposes an alternative, while verification determines whether it is sufficiently supported. The complete policy is governed by two thresholds, τ_0 for retaining original claims and τ_{sub} for accepting substitutes, which are calibrated jointly using conformal risk control and Learn Then Test. Figure 1 presents an overview of the MSRC framework.

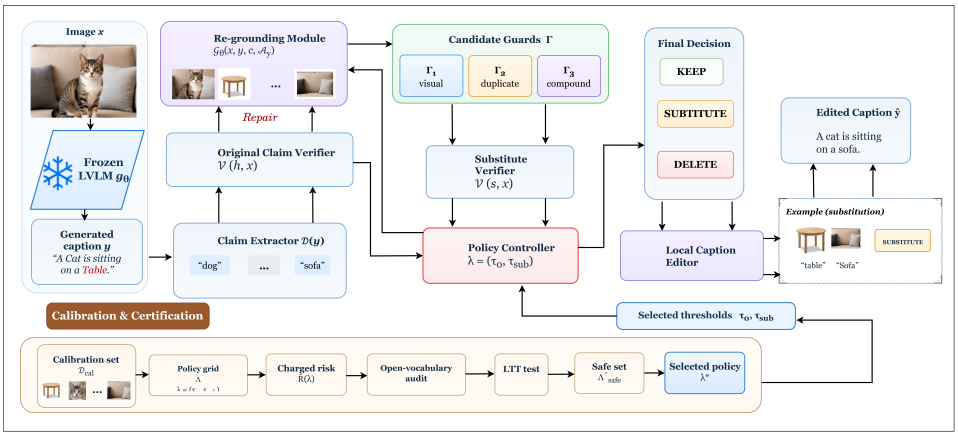

Figure 1: **Overview of MSRC.** *Top:* The frozen model's response is decomposed into object claims and verified against the image. Unsupported claims undergo re-grounding, admissibility checks, and independent re-verification. Calibrated thresholds determine whether each claim is kept, substituted, or deleted, with edits restricted to the affected text span (e.g., *table* → *sofa*). *Bottom:* Held-out calibration selects the policy thresholds, jointly accounting for substitution and deletion under a common risk budget.

2.2 Object-Claim Representation

A fixed deterministic extractor $\mathcal{D}$ decomposes the generated response into object claims, following the object-mention extraction used by CHAIR (Rohrbach et al., 2018) and claim-level decompositions of LVLM outputs.

$$\mathcal{D}(y) = \{c_j = (h_j, \sigma_j)\}_{j=1}^{m(x)}, \tag{1}$$

where h_j is the object expression associated with claim c_j, σ_j is its span in the generated response, and $m(x)$ is the number of eligible object-claim opportunities identified for image x. As in token-level hallucination localization (Park et al., 2025), the span σ_j in Eq. (1) connects the structured claim to its realization in the text. Consequently, MSRC can edit only the affected expression: a supported claim remains unchanged, an accepted substitute replaces the corresponding object span, and a rejected claim is removed without regenerating unrelated content. Not every noun phrase is treated as a visual object claim. A fixed predicate $\Gamma_{\text{vis}}(h) \in \{0, 1\}$ excludes non-visual expressions using the benchmark object vocabulary and its synonym mapping (Lin et al., 2014), WordNet lexical semantics, and fixed exclusion rules. Recognized compound object categories are handled as single semantic units. The exact construction is given in Eq. (15) of Appendix B.1. The formal calibration described below is restricted to the same guard-eligible object-claim population used by the implementation; original claims that cannot be automatically resolved by the benchmark vocabulary are not included in the calibrated risk set.

2.3 Visual Verification and Verified Repair

For an eligible object expression h, MSRC asks a fixed binary verification question in the object-existence format of POPE (Li et al., 2023b), such as "Is there a h in the image?" Rather than reading off the generated answer, the implementation aggregates the model probability assigned to several affirmative and negative first-token variants, analogous to probability-of-"True" self-evaluation in language models (Kadavath et al., 2022). Let $\mathcal{V}_Y$ and $\mathcal{V}_N$ denote the fixed sets of affirmative and negative token variants, respectively. The visual-support score is given as

$$V(h, x) = \frac{\sum_{k \in \mathcal{V}_Y} p_\theta(k \mid q(h), x)}{\sum_{k \in \mathcal{V}_Y} p_\theta(k \mid q(h), x) + \sum_{k \in \mathcal{V}_N} p_\theta(k \mid q(h), x)}, \tag{2}$$

where $q(h)$ denotes the fixed verification prompt. Because predictive confidence is only an imperfect indicator of hallucination (Xiao & Wang, 2021; Kadavath et al., 2022), the score in Eq. (2) is used

as relative evidence for visual support and is not assumed to be a calibrated probability; its operating thresholds are instead selected and assessed using held-out data, as described in Sec. 2.6. An original claim is kept when $V(h, x) \geq \tau_0$. If $V(h, x) < \tau_0$, MSRC asks the frozen LVLM to re-ground the failed object using the image and local response context, similar in role to the correction step of Woodpecker (Yin et al., 2024) and the resampling step of REVERSE (Wu et al., 2025). We write the resulting candidate compactly as $\hat{h} = R_\theta(x, y, h, \sigma)$. Greedy decoding is used, and the generated output is reduced to a single candidate object expression. Generation alone is not treated as evidence that the substitute is correct. Before acceptance, the candidate must satisfy fixed deployment guards. We write their combined decision as

$$\Gamma(\hat{h}, x, y) = \Gamma_{\text{vis}}(\hat{h})\Gamma_{\text{dup}}(\hat{h}, x, y), \tag{3}$$

where the first term in Eq. (3) checks object eligibility (Eq. (15)) and the second rejects duplicate candidates in the relevant sentence or running-caption context. Compound handling and the exact duplicate scope are implementation details described in Appendix B.2. Passing the guard does not establish visual correctness. The candidate is therefore verified independently using the same score V of Eq. (2). With $\lambda = (\tau_0, \tau_{\text{sub}})$, the claim-level policy is

$$\pi_\lambda(h, x) = \begin{cases} h, & V(h, x) \geq \tau_0, \\ \hat{h}, & V(h, x) < \tau_0,\ \Gamma(\hat{h}, x, y) = 1,\ V(\hat{h}, x) \geq \tau_{\text{sub}}, \\ \varnothing, & \text{otherwise.} \end{cases} \tag{4}$$

Equation (4) expresses the main MSRC intervention. The original claim is preserved whenever it is supported. A failed claim is replaced only by an admissible candidate that passes a second visual verification. Otherwise, the claim is deleted; this last branch corresponds to claim-level abstention in selective generation (Lee et al., 2024; Tayebati et al., 2025). The accepted action is applied only to the stored span σ.

2.4 Risk and Retained Information

MSRC is calibrated over the original eligible claim opportunities rather than over only the claims that remain after intervention. This distinction is important because deleting a failed claim should remove its error contribution but should not remove that opportunity from the denominator.

For image X_i, let m_i denote the number of calibrated object-claim opportunities before intervention. Let $O_{ij}(\lambda) \in \{h_{ij}, \hat{h}_{ij}, \varnothing\}$ denote the final outcome for claim j under Eq. (4), and let $T(w, X_i) \in \{0, 1\}$ indicate whether an emitted object claim w is unsupported by the image. The true claim-opportunity loss (Eq. (5)), the per-image loss, and its population risk (Eq. (6)) are then

$$\ell_{ij}(\lambda) = \begin{cases} 0, & O_{ij}(\lambda) = \varnothing, \\ T(O_{ij}(\lambda), X_i), & O_{ij}(\lambda) \neq \varnothing. \end{cases} \tag{5}$$

$$L_i(\lambda) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ell_{ij}(\lambda), \qquad \mathcal{R}(\lambda) = \mathbb{E}[L_i(\lambda)]\,. \tag{6}$$

The per-image loss $L_i(\lambda) \in [0, 1]$ is bounded, which is the setting required by conformal risk control. The denominator m_i is fixed before intervention. Therefore, deletion contributes zero final error but cannot artificially improve the risk by shrinking the denominator, in contrast to filtering-based factuality control, where risk is measured over retained claims only (Li et al., 2025). This is the quantity implemented by the calibration code. Risk alone does not measure how much useful content survives intervention. Let $K_{ij}(\lambda)$, defined in Eq. (7), indicate whether the policy emits a final claim for opportunity j. Over an index set $\mathcal{I}$, MSRC measures retained true-content yield as in Eq. (8), which plays a role analogous to the object-coverage metric of AMBER (Wang et al., 2024a):

$$K_{ij}(\lambda) = \mathbf{1}[O_{ij}(\lambda) \neq \varnothing] \tag{7}$$

$$\widehat{Y}_{\mathcal{I}}(\lambda) = \frac{\sum_{i \in \mathcal{I}} \sum_{j=1}^{m_i} K_{ij}(\lambda) \left[1 - \widetilde{\ell}_{ij}(\lambda)\right]}{\sum_{i \in \mathcal{I}} m_i}. \tag{8}$$

Here $\widetilde{\ell}_{ij}$ is the automatically available or conservatively charged loss defined in Eq. (10). Risk and

yield therefore play complementary roles: risk (Eq. (6)) penalizes unsupported content that survives intervention, whereas yield (Eq. (8)) rewards supported content that is retained or recovered.

2.5 Open-Vocabulary Risk Accounting

Generated substitutes are not limited to the closed benchmark vocabulary, while closed-vocabulary metrics such as CHAIR (Rohrbach et al., 2018) cannot evaluate object names outside this vocabulary (Ben-Kish et al., 2024). MSRC therefore distinguishes automatically resolvable substitutes from open-vocabulary repairs that remain unresolved.

Let $\mathcal{C}$ denote the benchmark object vocabulary (Lin et al., 2014), $G(x) \subseteq \mathcal{C}$ the annotated categories present in image x, and $\kappa(w)$ the fixed mapping from an object expression w to a benchmark category when such a mapping exists, following the CHAIR synonym lists (Rohrbach et al., 2018). In addition, let $A(w, x) \in \{0, 1\}$ indicate whether the implementation's fixed WordNet hypernym relation resolves w as a valid semantic ancestor of an annotated object in $G(x)$. The automatic substitute label is

$$Y(w, x) = \begin{cases} 0, & \kappa(w) \in G(x) \text{ or } A(w, x) = 1, \\ 1, & \kappa(w) \in \mathcal{C} \setminus G(x), \\ \perp, & \text{otherwise.} \end{cases} \tag{9}$$

Thus, under Eq. (9), a substitute is marked supported when it matches an annotated category directly or is resolved by the fixed semantic-ancestor rule. A mapped benchmark category that is absent from the image is marked unsupported. Only the remaining substitutes are assigned $Y(w, x) = \perp$ and enter the human-audit mechanism. The audit is score conditional because deployment accepts a substitute only when its verification score exceeds τ_{sub} (Eq. (4)). A fixed threshold grid $\Lambda = \{\tau^{(1)}, \ldots, \tau^{(G)}\}$ is specified before the audit labels are inspected. For each $\tau \in \Lambda$, let n_τ be the number of audited unresolved substitutes satisfying $V \geq \tau$, and let k_τ be the number judged visually supported. A one-sided Clopper–Pearson lower bound $\underline{p}_\tau$ is computed with a Bonferroni correction over the fixed grid (Eq. (16)). The corresponding conservative error charge is $\gamma(\tau) = 1 - \underline{p}_\tau$; if $n_\tau < n_{\min}$, the implementation uses the worst-case charge $\gamma(\tau) = 1$ (Eq. (17)). The exact construction is given in Appendix B.3.

For each original claim opportunity, the charged loss used by calibration replaces the unknown indicator T of Eq. (5) with the automatic label of Eq. (9) or the audit charge:

$$\widetilde{\ell}_{ij}(\lambda) = \begin{cases} 0, & O_{ij}(\lambda) = \varnothing, \\ Y(O_{ij}(\lambda), X_i), & Y(O_{ij}(\lambda), X_i) \in \{0, 1\}, \\ \gamma(\tau_{\text{sub}}), & Y(O_{ij}(\lambda), X_i) = \perp . \end{cases} \tag{10}$$

The resulting charged image loss is

$$\widetilde{L}_i(\lambda) = \frac{1}{m_i} \sum_{j=1}^{m_i} \widetilde{\ell}_{ij}(\lambda). \tag{11}$$

Because every claim-opportunity loss in Eq. (10) lies in $[0, 1]$, we also have $0 \leq \widetilde{L}_i(\lambda) \leq 1$ in Eq. (11), preserving the bounded-loss condition of conformal risk control. Importantly, unresolved accepted substitutes receive the threshold-specific audit charge, while deleted claims continue to contribute zero loss without changing the fixed denominator m_i.

2.6 Policy Selection and Held-Out Certification

The original-claim and substitute thresholds interact. Increasing τ_0 sends more original claims to the repair branch, whereas increasing τ_{sub} rejects more proposed substitutes and redirects them to deletion. MSRC therefore searches over the complete policy $\lambda = (\tau_0, \tau_{\text{sub}})$ rather than calibrating the two thresholds independently, treating calibration as selection over a finite grid of candidate procedures as in Learn Then Test.

The available calibration images are randomly partitioned, using a fixed seed, into a policy-selection set $\mathcal{I}_{\text{sel}}$ and an independent certification set $\mathcal{I}_{\text{cert}}$. For any index set $\mathcal{I}$, we define the empirical

charged risk and the corresponding conservative finite-sample criterion as

$$\widehat{\mathcal{R}}_{\mathcal{I}}(\lambda) = \frac{\sum_{i\in\mathcal{I}} \widetilde{L}_i(\lambda)}{|\mathcal{I}|}, \qquad \widehat{B}_{\mathcal{I}}(\lambda) = \frac{|\mathcal{I}|\widehat{\mathcal{R}}_{\mathcal{I}}(\lambda) + 1}{|\mathcal{I}| + 1}. \tag{12}$$

The criterion $\widehat{B}_{\mathcal{I}}$ in Eq. (12) is a conservative empirical risk criterion motivated by conformal risk control, with $B = 1$ for the charged loss in Eq. (11).

Let Θ_0 be the candidate grid for τ_0 and let Λ be the same fixed substitute-threshold grid used by the score-conditional audit (Sec. 2.5). Given target risk α, the selection split is searched for a high-yield policy satisfying

$$\hat{\lambda} = \underset{\lambda\in\Theta_0\times\Lambda}{\arg\max}\, \widehat{Y}_{\mathrm{sel}}(\lambda) \quad \text{subject to} \quad \widehat{B}_{\mathrm{sel}}(\lambda) \le (1-\eta)\alpha, \tag{13}$$

where $\widehat{Y}_{\mathrm{sel}}$ is the yield of Eq. (8) on $\mathcal{I}_{\mathrm{sel}}$ and $\eta \in (0,1)$ is a fixed selection margin. In the implementation, an additional substitute-precision safeguard is applied during the search and is relaxed only if no feasible policy satisfies it; this safeguard is described in Appendix B.4 rather than treated as part of the core statistical definition. After selection, $\hat{\lambda}$ is frozen. No threshold, prompt, guard, audit charge, or extraction rule is changed using the certification split. The selected policy meets the held-out empirical criterion if

$$\widehat{B}_{\mathrm{cert}}(\hat{\lambda}) \le \alpha. \tag{14}$$

This is a held-out empirical criterion, not a formal population-risk guarantee. In particular, MSRC does not introduce a separate Hoeffding radius (Bentkus, 2004) in the implemented calibration procedure; the same finite-sample criterion in Eq. (12) is used for selection (Eq. (13)) and certification (Eq. (14)). For comparison, the matched verification-only filtering baseline (VOF (Ours)), which corresponds to filtering-based factuality control (Li et al., 2025), disables the substitution branch of Eq. (4) while retaining the same object population, verifier, original-claim threshold family, loss definition, and calibration procedure. A failed original claim is therefore deleted directly. This comparison isolates the contribution of verified repair from the effect of verification and removal alone.

3 Experiments

We evaluate MSRC on four frozen LVLM backbones: LLaVA-v1.5-7B, InstructBLIP-7B, Qwen2.5-VL-3B, and BLIP-2. Our evaluation spans image captioning, open-ended visual question answering, and discriminative hallucination benchmarks, including MS-COCO CHAIR, AMBER, MMHal-Bench, HaloQuest, POPE, and MME. MSRC requires no additional training, with verification and repair thresholds calibrated separately for each backbone. Further details on implementation (Appendix C.1), evaluation protocols (Appendix C.2), datasets (Appendix C.3), and hyperparameters (Appendix C.4) are provided in the respective appendices.

4 Results

We compare MSRC with the frozen backbone (Vanilla) and Verification-Only Filtering (VOF (Ours)), which deletes unsupported claims without attempting repair. Both methods share the same extraction, verification, guards, and calibration procedure, allowing us to isolate the contribution of verified substitution. Unless stated otherwise, thresholds are calibrated separately for each backbone and benchmark at $\alpha = 0.05$ using held-out data. We report hallucination reduction, answer accuracy, and information retention, followed by comparisons with existing methods and qualitative analysis. Additional results are provided in the appendix: discriminative hallucination benchmarks (Appendix G); calibration, risk–coverage behaviour, and stability (Appendix D); open-vocabulary evaluation on BLIP-2 (Appendix E); and comparison with CEBC (Mishra et al., 2026) (Appendix F).

Captioning Results. As shown in Tables 1 and 2, MSRC substantially reduces hallucination across all three backbones. On COCO-CHAIR, instance-level hallucination (C_I) decreases by 69.7%, 43.3%, and 73.6% relative to Vanilla on LLaVA, Qwen2.5, and InstructBLIP, respectively. Compared with VOF (Ours), MSRC further reduces C_I by 1.02 and 0.97 points on LLaVA and

Table 1: **COCO-CHAIR evaluation results on 500 images.** Lower CHAIR scores indicate fewer hallucinations.

Method	$C_S \downarrow$	$C_I \downarrow$
LLaVA-1.5-7B		
Vanilla	51.70	14.33
OPERA Huang et al. (2024)	47.80	14.60
DoLA Chuang et al. (2024)	51.60	14.10
VCD Leng et al. (2024)	48.60	14.90
AGLA An et al. (2025)	43.00	14.10
MemVR Zou et al. (2025)	46.60	13.00
HALVA Sarkar et al. (2025)	41.40	11.70
Woodpecker Yin et al. (2024)	45.80	14.80
REVERSE (τ=0.003) Wu et al. (2025)	37.00	10.30
REVERSE (τ=0.0003) Wu et al. (2025)	13.60	6.10
VOF (Ours)	16.73	5.36
MSRC (Ours)	**13.20**	**4.34**
Qwen2.5-VL-3B		
Vanilla	39.27	8.47
DoLA Chuang et al. (2024)	47.60	14.00
REVERSE (τ=0.01) Wu et al. (2025)	39.40	10.50
VOF (Ours)	15.20	**3.44**
MSRC (Ours)	**14.97**	4.80
InstructBLIP-7B		
Vanilla	51.90	14.90
VOF (Ours)	15.03	4.90
MSRC (Ours)	**14.87**	**3.93**

Table 2: **AMBER-Generative evaluation results.**

Method	CHAIR↓	Cover↑	Hal↓	Cog↓
LLaVA-1.5-7B				
Vanilla	7.4	51.3	34.5	4.0
OPERA Huang et al. (2024)	7.3	49.6	32.0	3.5
DoLA Chuang et al. (2024)	7.6	51.6	36.0	4.0
VCD Leng et al. (2024)	6.7	46.5	27.8	2.0
AvisC Woo et al. (2025)	6.3	46.6	25.6	2.0
HALVA Sarkar et al. (2025)	6.6	**53.0**	32.2	3.4
Woodpecker Yin et al. (2024)	6.9	48.9	30.4	3.6
REVERSE (τ=0.003) Wu et al. (2025)	6.0	52.2	30.4	3.0
REVERSE (τ=0.0003) Wu et al. (2025)	4.0	26.9	**10.2**	0.9
VOF (Ours)	4.7	50.8	22.7	1.7
MSRC (Ours)	**3.7**	48.5	15.9	**0.7**
Qwen2.5-VL-3B				
Vanilla	6.1	**53.9**	30.1	2.6
DoLA† Chuang et al. (2024)	9.7	48.1	31.4	1.9
REVERSE† (τ=0.01) Wu et al. (2025)	7.5	51.5	34.4	3.6
VOF (Ours)	4.7	53.6	23.6	1.2
MSRC (Ours)	**3.6**	51.0	**15.8**	**0.7**
InstructBLIP-7B				
Vanilla	13.3	**53.4**	52.4	6.3
VCD Leng et al. (2024)	7.6	47.7	29.9	2.2
AvisC Woo et al. (2025)	6.7	46.7	28.0	2.6
Octopus Suo et al. (2025)	6.1	48.5	**22.2**	**1.3**
VOF (Ours)	5.5	51.9	25.0	1.6
MSRC (Ours)	**5.4**	52.5	24.6	1.5

InstructBLIP, but increases it by 1.36 points on Qwen2.5, while improving sentence-level CHAIR across all backbones. On AMBER, CHAIR decreases from 7.4 to 3.7, 6.1 to 3.6, and 13.3 to 5.4, with coverage of 48.5%, 51.0%, and 52.5%, respectively. Although MSRC achieves lower CHAIR than VOF (Ours) throughout, coverage decreases on LLaVA and Qwen2.5. In contrast, strict REVERSE achieves 26.9% coverage on LLaVA Wu et al. (2025), highlighting the importance of balancing hallucination reduction with information retention.

Question Answering Results. As shown in Tables 3 and 4, MSRC consistently improves answer accuracy while reducing hallucination. On HaloQuest, average accuracy increases from 23.19% to 68.75% on LLaVA, 57.89% to 76.48% on Qwen2.5, and 22.86% to 65.79% on InstructBLIP. Compared with VOF (Ours), MSRC further improves accuracy by 7.73, 4.77, and 23.19 percentage points, respectively. Notably, visually challenging accuracy improves over VOF (Ours) across all backbones, indicating that verified repair can preserve supported answers otherwise lost through deletion. On MMHal-Bench, MSRC increases the average score from 2.594 to 3.396, 4.438 to 4.615, and 2.542 to 3.188, while reducing hallucination rates from 53.1% to 34.4%, 18.8% to 13.5%, and 51.0% to 37.5%, respectively. These results demonstrate the value of verified repair in maintaining answer informativeness alongside hallucination reduction. Additional qualitative examples are provided in Appendix H.

4.1 Ablation Study

Component contributions. Table 5 examines the contribution of individual components. Removing mention coverage increases within-run C_I by 1.16, 1.43, and 1.14 points on LLaVA, Qwen2.5,

Table 3: **HaloQuest evaluation results.** Accuracy (%) measures the proportion of correct, visually grounded answers. Higher values indicate better performance.

Method	Avg.↑	FP↑	VC↑	IC↑
LLaVA-1.5-7B				
Vanilla	23.19	17.11	40.98	11.57
DoLA Chuang et al. (2024)	22.90	17.20	40.10	11.60
HALVA Sarkar et al. (2025)	23.90	21.10	37.40	10.70
REVERSE (τ=0.003) Wu et al. (2025)	30.70	31.80	31.50	26.90
REVERSE (τ=0.0003) Wu et al. (2025)	32.30	29.40	18.70	58.80
PostAlign Wu et al. (2026)	34.44	33.90	37.20	32.20
FINER-Tuning Xiao et al. (2026)	38.80	–	–	–
VOF (Ours)	61.02	71.38	39.34	67.77
MSRC (Ours)	**68.75**	**81.91**	**41.53**	**76.86**
Qwen2.5-VL-3B				
Vanilla	57.89	54.61	**61.20**	61.16
DoLA Chuang et al. (2024)	27.40	16.50	51.10	19.00
CCoT Mitra et al. (2024)	41.06	–	–	–
VRGA Peng et al. (2026)	59.03	–	–	–
REVERSE (τ=0.01) Wu et al. (2025)	45.10	42.90	41.80	55.50
VOF (Ours)	71.71	77.63	57.38	78.51
MSRC (Ours)	**76.48**	**83.55**	**61.20**	**81.82**
InstructBLIP-7B				
Vanilla	22.86	25.00	27.87	9.92
VOF (Ours)	42.60	52.30	16.94	57.02
MSRC (Ours)	**65.79**	**82.89**	**28.42**	**79.34**

Table 4: **MMHal-Bench evaluation results.** Performance measures visual grounding and hallucination reduction.

Method	Score↑	Hall.↓
LLaVA-1.5-7B		
Vanilla	2.594	53.1%
DoLA Chuang et al. (2024)	2.330	56.0%
HALVA Sarkar et al. (2025)	2.250	54.0%
REVERSE (τ=0.003) Wu et al. (2025)	2.560	47.0%
REVERSE (τ=0.0003) Wu et al. (2025)	3.280	**30.0%**
VCD Leng et al. (2024)	2.250	63.5%
Self-Aug Im et al. (2026)	2.550	59.0%
EviAnchor Jia et al. (2026)	2.780	49.0%
VOF (Ours)	2.906	43.8%
MSRC (Ours)	**3.396**	34.4%
Qwen2.5-VL-3B		
Vanilla	4.438	18.8%
DoLA Chuang et al. (2024)	2.720	46.0%
REVERSE (τ=0.01) Wu et al. (2025)	3.150	29.0%
VOF (Ours)	4.375	**13.5%**
MSRC (Ours)	**4.615**	**13.5%**
InstructBLIP-7B		
Self-Aug Im et al. (2026)	2.160	64.0%
EviAnchor Jia et al. (2026)	2.340	56.2%
Vanilla	2.542	51.0%
VOF (Ours)	2.542	46.9%
MSRC (Ours)	**3.188**	**37.5%**

Table 5: **Component ablations** on 500 MS-COCO images at $\alpha = 0.05$. The first row reports absolute values; subsequent rows report differences from the original within-run MSRC reference. Positive differences favour MSRC. #sub denotes accepted substitutions.

Variant	LLaVA-1.5 7B			Qwen2.5-VL 3B			InstructBLIP 7B		
	ΔC_I	Δded	#sub	ΔC_I	Δded	#sub	ΔC_I	Δded	#sub
MSRC (Ours)	4.34	7.82	32	4.80	7.76	24	3.93	7.60	26
w/o mention coverage	+1.16	+2.60	30	+1.43	+4.03	33	+1.14	+3.03	27
w/o precision floor	+0.00	+0.00	32	+0.00	+0.00	24	+0.04	+0.04	41
strict floor (CP-LCB)	-0.02	-0.06	0	-0.12	-0.46	0	+0.03	+0.03	0
w/o semantic gate	-0.02	+0.06	91	-0.07	-0.17	82	+0.14	+0.40	128
open-vocabulary substitutes	-0.04	-0.06	19	-0.12	-0.46	12	+0.01	+0.01	20
k-candidate (k=3)	-0.01	+0.01	42	-0.09	-0.32	25	+0.00	+0.01	25
w/ entailment ladder	-0.24	-0.24	19	+0.01	-0.17	4	+0.59	+1.18	10
caption-level duplicate scope	-0.02	-0.06	0	-0.07	-0.31	4	+0.03	+0.01	3
w/o duplicate guard	+0.02	+0.00	84	-0.10	-0.31	38	-0.07	-0.01	81

and InstructBLIP, respectively, highlighting its role in preventing excessive deletion. Removing the precision floor has negligible effect, whereas the stricter Clopper–Pearson variant suppresses all substitutions. Disabling the semantic gate substantially increases accepted substitutions without consistently improving hallucination. Open-vocabulary repair, multiple candidates, the entailment ladder, and duplicate-guard variants exhibit mixed, backbone-dependent effects, indicating that greater repair flexibility does not necessarily improve reliability.

Repair and calibration behaviour. Figure 2 isolates the contribution of verified repair under matched risk budgets. MSRC retains more content than VOF (Ours) at 13 of 15 operating points,

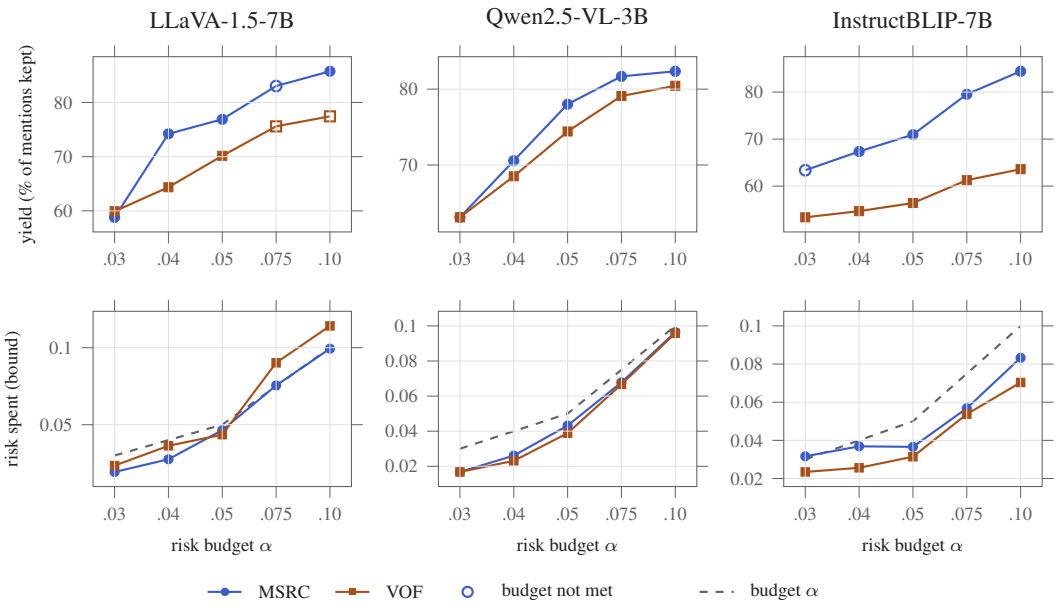

Figure 2: **Repair recovers content at matched risk budgets.** Comparison of MSRC (Ours) and VOF (Ours) under shared verification and calibration, differing only in the substitution branch.

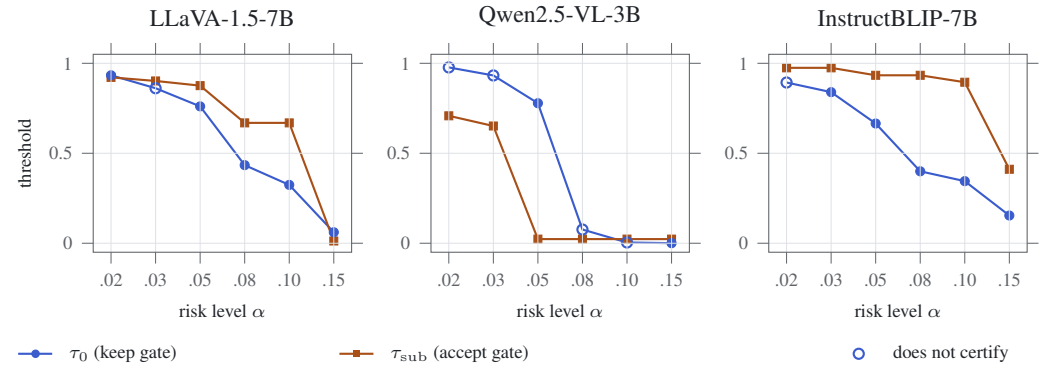

Figure 3: **Calibrated thresholds across risk levels.** Thresholds vary across backbones, supporting model-specific calibration.

with gains of 6.7–9.8 points on LLaVA, 1.9–3.6 on Qwen2.5, and 10.0–20.8 on InstructBLIP. The exceptions occur at $\alpha = 0.03$, where Qwen2.5 admits no substitution and LLaVA retains 1.1 points less content. Figure 3 further shows that the calibrated thresholds vary across backbones and risk budgets, motivating model-specific calibration. Additional risk–coverage, calibration, and stability analyses are presented in Appendix D.

5 Conclusion

We presented MSRC, a training-free framework for reducing visual hallucination through verified claim repair. MSRC proposes replacements for unsupported object claims using the frozen vision–language model, accepting them only after semantic checks and independent visual verification; otherwise, the claims are removed. Calibrated thresholds balance hallucination reduction and information preservation. Experiments on COCO-CHAIR, AMBER, HaloQuest, and MMHal-Bench show substantial hallucination reduction. Compared with verification-only filtering, MSRC improves question-answering accuracy and generally retains more information under matched risk budgets. However, substitution does not always improve instance-level CHAIR, while stricter hallucination control may reduce coverage. MSRC depends on extraction and verification quality and cannot recover unmentioned objects. These findings highlight the importance of preserving correct information while removing unsupported claims.

AI USE STATEMENT

Generative AI tools were used to help with manuscript writing, language editing, formatting, and other supporting tasks. The authors were responsible for the ideation, model design, experimental decisions, data processing, results, and their interpretation. All AI-assisted content was carefully reviewed and checked by the authors before being included in the manuscript.

REFERENCES

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

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. *arXiv preprint arXiv:2312.03631*, 2024.

Vidmantas Bentkus. On hoeffding's inequalities. *The Annals of Probability*, 32(2):1650–1673, 2004. doi: 10.1214/009117904000000360.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 7824–7846. PMLR, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0.

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. *arXiv preprint arXiv:2306.13394*, 2023.

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

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*, 2026. URL https://arxiv.org/abs/2510.13315.

Sihang Jia, Shuliang Liu, Songbo Yang, and Xuming Hu. EviAnchor: Mitigating hallucinations in large vision-language models via regional visual evidence compensation, 2026. URL https://arxiv.org/abs/2608.29092.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, 2024. doi: 10.18653/v1/2024.findings-emnlp.290.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1600.

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

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne 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, 2023b. doi: 10.18653/v1/2023.emnlp-main.20.

Zhuohang Li, Chao Yan, Nicholas J. Jackson, Wendi Cui, Bo Li, Jiaxin Zhang, and Bradley A. Malin. Towards statistical factuality guarantee for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11435–11456, 2025. doi: 10.18653/v1/2025.emnlp-main.576.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

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

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.557.

Ashish Mishra, Tarun Kumar, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. CEBC: Conformal evidence-bounded control for low-hallucination vision–language generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46193–46206, 2026.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14420–14431, 2024. doi: 10.1109/CVPR52733.2024.01367. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mitra_Compositional_Chain-of-Thought_Prompting_for_Large_Multimodal_Models_CVPR_2024_paper.html.

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*, pp. 29893–29903, 2025.

Ruiying Peng, Xueyu Wu, Jing Lei, Lu Hou, Yuanzheng Ma, and Xiao-Hui Li. Deeper thought, weaker aim: Understanding and mitigating perceptual impairment during reasoning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12064–12073, 2026. URL https://arxiv.org/abs/2603.14184.

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, 2018. doi: 10.18653/v1/D18-1437.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Arik, and Tomas Pfister. Mitigating object hallucination in MLLMs via data-augmented phrase-level alignment. In *International Conference on Learning Representations*, 2025.

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

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29904–29914, 2025.

Sina Tayebati, Divake Kumar, Nastaran Darabi, Dinithi Jayasuriya, Theja Tulabandhula, Ranganath Krishnan, and Amit Ranjan Trivedi. CAP: Conformalized abstention policies for context-adaptive risk management for LLMs and VLMs. In *Proceedings of the 17th Asian Conference on Machine Learning*, volume 304 of *Proceedings of Machine Learning Research*, pp. 926–941. PMLR, 2025.

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. *arXiv preprint arXiv:2311.07397*, 2024a.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V. Le, Thang Luong, and Golnaz Ghiasi. HaloQuest: A visual hallucination dataset for advancing multimodal reasoning. In *Computer Vision – ECCV 2024*. Springer, 2024b.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.99. URL https://aclanthology.org/2025.findings-acl.99/.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2206.

Yixuan Wu, Yang Zhang, Jian Wu, Philip Torr, and Jindong Gu. PostAlign: Multimodal grounding as a corrective lens for MLLMs. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=zJnPyb2xrp.

Rui Xiao, Sanghwan Kim, Yongqin Xian, Zeynep Akata, and Stephan Alaniz. FINER: MLLMs hallucinate under fine-grained negative queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 36235–36244, 2026. URL https://arxiv.org/abs/2603.17662.

Yijun Xiao and William Yang Wang. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 2734–2744. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.236.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via DPO: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 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*, 67:220105, 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*, pp. 13807–13816, 2024.

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In *International Conference on Learning Representations*, 2025.

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*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80873–80899. PMLR, 2025. URL https://proceedings.mlr.press/v267/zou25e.html.

A Related Work

Hallucination in vision-language models. Hallucination in image-conditioned generation has been studied since early image-captioning systems. Rohrbach et al. (2018) showed that captioning models can mention objects that are absent from the image and introduced object-level measures for this behavior. The problem remains prominent in modern large vision–language models (LVLMs), where strong language priors can produce fluent but visually unsupported responses. POPE (Li et al., 2023b) evaluates object hallucination through object-presence questions, while FaithScore (Jing et al., 2024) decomposes generated responses into fine-grained visual facts. HaloQuest (Wang et al., 2024b) further exposes hallucinations in more difficult visual questions that require careful image grounding. These studies motivate treating a generated visual claim, rather than the response as a whole, as the unit on which hallucination should be detected and corrected.

Training-free mitigation and post-generation correction. Training-free methods improve visual grounding through contrastive decoding (VCD (Leng et al., 2024), HALC (Chen et al., 2024), Octopus (Suo et al., 2025)), attention-based mechanisms (OPERA (Huang et al., 2024), AGLA (An et al., 2025), AvisC (Woo et al., 2025)), and adaptive visual evidence (Self-Aug (Im et al., 2026), EviAnchor (Jia et al., 2026)). Post-generation methods instead revise unreliable content through learned revision (LURE (Zhou et al., 2024)), visual validation (Woodpecker (Yin et al., 2024)), generative feedback (DeGF (Zhang et al., 2025)), retrospective resampling (REVERSE (Wu et al., 2025)), or improved grounding (PostAlign (Wu et al., 2026)). Training-based approaches such as RLHF-V (Yu et al., 2024), HALVA (Sarkar et al., 2025), and OnDPO (Yang et al., 2025) improve faithfulness through alignment. Unlike these methods, MSRC uses a frozen LVLM to propose repairs and independently verifies each replacement before acceptance, accounting for the risk of newly generated content.

Uncertainty, selective prediction, and risk control. Uncertainty provides useful signals for detecting unreliable generation. Xiao and Wang (Xiao & Wang, 2021) connected predictive uncertainty with hallucination, SelfCheckGPT (Manakul et al., 2023) uses disagreement across sampled responses, and semantic-uncertainty methods (Kuhn et al., 2023; Farquhar et al., 2024) assess uncertainty at the level of meaning rather than individual tokens. While these approaches help identify unreliable content, they do not determine how a failed visual claim should be repaired. Statistical methods address a related problem by controlling prediction risk: Conformal Risk Control (CRC) (Angelopoulos et al., 2024) provides a framework for bounded losses, while Learn Then Test (LTT) (Angelopoulos et al., 2025) supports risk-aware selection among candidate procedures. In LVLMs, ConfLVLM (Li et al., 2025) uses conformal prediction to filter unsupported visual details, while selective-generation and conformal-abstention methods (Lee et al., 2024; Tayebati et al., 2025) reduce risk by withholding uncertain outputs. MSRC builds on this direction by allowing verified repair alongside deletion, accounting for newly generated claims while preserving supported visual information. Unlike filtering or abstention, MSRC can retain, repair, or remove a claim based on visual support. Since generated replacements may fall outside the annotation vocabulary, risk is evaluated over the *final emitted claim*, including substitutions. This couples hallucination control with retained true-claim yield, reducing reliance on excessive deletion.

B Additional Method Details

B.1 Object Eligibility

Let $\mathcal{F}$ and $\mathcal{S}$ denote fixed filler and scene/mass exclusion sets, respectively, and let $\nu(h)$ denote the selected WordNet noun lexname of expression h, when available. The visual lexname set is

$$\mathcal{L}_{\text{vis}} = \left\{ \begin{array}{c} \texttt{noun.animal}, \texttt{noun.artifact}, \texttt{noun.body}, \texttt{noun.food}, \\ \texttt{noun.plant}, \texttt{noun.object}, \texttt{noun.person} \end{array} \right\}.$$

The eligibility predicate can be written as

$$\Gamma_{\text{vis}}(h) = \mathbb{1}[h \notin \mathcal{F} \cup \mathcal{S} \wedge (\kappa(h) \in \mathcal{C} \vee \nu(h) \in \mathcal{L}_{\text{vis}})]\,. \tag{15}$$

The lexical resources and exclusion sets are fixed before calibration. The calibration script additionally restricts original risk-accounting claims to heads that can be mapped through the benchmark vocabulary, matching the deployed calibration population.

B.2 Repair Guard Details

The deployment implementation applies the same fixed object guard before a substitute can be credited during calibration. Let $\rho(x, y)$ denote the relevant generated-text context available to the repair procedure. The duplicate check can be summarized as

$$\Gamma_{\mathrm{dup}}(\hat{h}, x, y) = \mathbb{1}\left[\hat{h} \notin \mathrm{words}(\rho(x, y))\right].$$

The combined guard is

$$\Gamma(\hat{h}, x, y) = \Gamma_{\mathrm{vis}}(\hat{h})\Gamma_{\mathrm{dup}}(\hat{h}, x, y).$$

The implementation also handles recognized compound categories before re-grounding and applies the same duplicate scope during calibration and deployment.

B.3 Open-Vocabulary Audit Construction

The score-conditional audit uses the same fixed substitute-threshold grid Λ that is later searched during policy calibration. For each $\tau \in \Lambda$, let n_τ denote the number of audited unresolved substitutes with $V \geq \tau$, and let k_τ denote the number judged visually supported.

With audit failure level δ_a and a Bonferroni correction over $G = |\Lambda|$ fixed thresholds, the one-sided Clopper–Pearson lower bound is

$$\underline{p}_\tau = \begin{cases} \mathrm{Beta}^{-1}_{\delta_a/G}(k_\tau, n_\tau - k_\tau + 1), & 0 < k_\tau < n_\tau, \\ (\delta_a/G)^{1/n_\tau}, & k_\tau = n_\tau > 0, \\ 0, & k_\tau = 0. \end{cases} \tag{16}$$

The second branch is the closed-form lower endpoint used by the implementation when all audited items in the stratum are supported. The final score-conditional error charge is

$$\gamma(\tau) = \begin{cases} 1 - \underline{p}_\tau, & n_\tau \geq n_{\min}, \\ 1, & n_\tau < n_{\min}. \end{cases} \tag{17}$$

Thus, a threshold with too little audited evidence receives the maximally conservative charge.

B.4 Implementation-Matched Policy Search

The implementation searches a data-derived grid for τ_0 and the fixed audited grid for τ_{sub}. For each candidate substitute threshold, the corresponding score-conditional charge $\gamma(\tau_{\mathrm{sub}})$ is read from the audit table. Candidate policies are evaluated using the same guards and claim population used at deployment.

During selection, the implementation first requires accepted substitutes to have at least the precision of the claims retained by the matched VOF (Ours) comparator. If no policy satisfies both this safeguard and the risk criterion, the precision safeguard is relaxed and the risk-constrained search is repeated. Among feasible candidates, the implementation retains the policy with the highest true-content yield. This safeguard affects implementation-level policy search but does not alter the loss, audit charge, or certification criterion defined in the main method.

B.5 Deployment

At deployment, the complete pipeline is frozen. The LVLM generates the response incrementally, the extractor identifies object claims, and each eligible claim is processed by the selected policy $\pi_{\hat{\lambda}}$. Supported originals are retained. Failed originals may be re-grounded; a proposed substitute must pass the fixed guards and independent verification before replacing the original span. Claims without an admissible verified repair are deleted. No benchmark labels, audit labels, or calibration data are accessed during deployment.

Table 6: **Summary of evaluation benchmarks.**

Benchmark	Task	Image source
COCO-CHAIR (Rohrbach et al., 2018)	Captioning	MS-COCO (Lin et al., 2014)
AMBER-G (Wang et al., 2024a)	Captioning	AMBER
OpenCHAIR (Ben-Kish et al., 2024)	Captioning	MS-COCO
HaloQuest (Wang et al., 2024b)	Open-ended VQA	Real and synthetic
MMHal-Bench (Sun et al., 2023)	Open-ended VQA	Open Images
POPE (Li et al., 2023b)	Yes/no	MS-COCO
MME (hallucination) (Fu et al., 2023)	Yes/no	MME

C Experimental Details

C.1 Implementation Details

We evaluate MSRC using four frozen LVLM backbones: LLaVA-v1.5-7B (Liu et al., 2023), Qwen2.5-VL-3B (Bai et al., 2025), InstructBLIP with Vicuna-7B (Dai et al., 2023), and MiniGPT-4 with Vicuna-7B (Zhu et al., 2023). MSRC is training-free and introduces no additional learned parameters. The same frozen backbone is used for response generation, visual verification, and re-grounding.Failed object claims are re-grounded using greedy decoding. A proposed replacement is accepted only after passing the fixed admissibility guards and independent visual re-verification. Because verifier-score distributions differ across backbones, the original-claim threshold τ_0 and substitute threshold τ_{sub} are calibrated separately for each model rather than transferred across models. Once selected, the complete policy is frozen and applied unchanged during evaluation.Open-vocabulary substitutes are first resolved using the benchmark vocabulary and WordNet. Accepted repairs that remain unresolved are handled using the score-conditional audit procedure described in Sec. 2.5. Additional calibration and audit settings are provided in Appendix B.

C.2 Evaluation Protocol

We evaluate MSRC across image captioning, open-ended visual question answering, and discriminative hallucination benchmarks. For free-form generation, we use CHAIR-MSCOCO (Rohrbach et al., 2018), AMBER (Wang et al., 2024a), MMHal-Bench (Sun et al., 2023), and HaloQuest (Wang et al., 2024b). CHAIR evaluates object hallucination in generated captions through CHAIR_i and CHAIR_s, while AMBER covers both generative and discriminative hallucination evaluation.

MMHal-Bench and HaloQuest assess open-ended visual responses, including unsupported claims and questions containing false premises. For these evaluations, we use Gemini 3.7 Flash as an independent LLM judge to assess generated answers against the corresponding questions and reference information under a fixed evaluation protocol.

We additionally evaluate POPE (Li et al., 2023b), which measures object-existence hallucination under Random, Popular, and Adversarial sampling, and MME (Fu et al., 2023), which evaluates broader multimodal perception and reasoning through discriminative questions. We follow the benchmark-specific metrics and report generative and discriminative results separately.

C.3 Dataset Details

We evaluate MSRC on seven benchmarks spanning image captioning, open-ended visual question answering (VQA), and discriminative hallucination evaluation. Table 6 summarizes the evaluation sets and metrics. All results use the official images, questions, and annotations of each benchmark, and calibration images are kept disjoint from evaluation images.

COCO-CHAIR. CHAIR (Rohrbach et al., 2018) measures object hallucination in image captions using the 80 object categories of MS-COCO (Lin et al., 2014), together with a fixed synonym list that maps caption words to categories. A mentioned object is counted as hallucinated when its category does not appear in the image's ground-truth segmentation or caption annotations. We report the instance-level metric C_I, the fraction of mentioned objects that are hallucinated, and the sentence-level metric C_S, the fraction of captions containing at least one hallucinated object. Following prior

work (Huang et al., 2024; Wu et al., 2025), the main results use 500 randomly sampled MS-COCO validation images. The calibration and risk–coverage analyses in Appendix D use a larger set of 3,000 evaluation images. The same COCO vocabulary also defines the benchmark vocabulary $\mathcal{C}$ and the category mapping κ used by MSRC

AMBER. AMBER (Wang et al., 2024a) is an LLM-free benchmark with dense manual annotations of objects, attributes, and relations. It covers both generative and discriminative evaluation. For the generative split (AMBER-G), the model describes each image, and four metrics are reported:

- CHAIR, the fraction of hallucinated objects among mentioned objects;
- Cover, the fraction of annotated objects that the response mentions;
- Hal, the fraction of responses containing any hallucination;
- Cog, which measures whether hallucinated objects resemble objects that humans tend to imagine.

Cover is particularly relevant to MSRC because it exposes content lost through deletion. We use 704 images from the generative split.

OpenCHAIR. Closed-vocabulary metrics such as CHAIR cannot judge objects outside the 80 COCO categories. OpenCHAIR (Ben-Kish et al., 2024) extends hallucination measurement to open-vocabulary object mentions. We use it with BLIP-2 (Li et al., 2023a) to check that MSRC also reduces hallucination beyond the fixed vocabulary (Appendix E). Following the original protocol, we additionally report BLEU-4, CIDEr, and the mean NLI contradiction probability $\bar{p}$ as measures of caption quality.

HaloQuest. HaloQuest (Wang et al., 2024b) is a VQA benchmark designed to elicit hallucination. It combines real images with synthetic images and groups questions into three categories:

- *false premise* (FP): the question assumes content that is not present in the image;
- *visually challenging* (VC): answering requires fine-grained visual inspection;
- *insufficient context* (IC): the image does not contain enough information to answer.

We evaluate on its 608-question evaluation set and report per-category and average accuracy. Because answers are free-form, correctness is judged by an LLM evaluator against the reference answers (Appendix C.2).

MMHal-Bench. MMHal-Bench (Sun et al., 2023) contains 96 image–question pairs built on Open Images. The questions span eight question types, such as attributes, counting, spatial relations, and adversarial objects, and twelve object topics. Each response is rated on a 0–6 scale against the image annotations and a reference answer. We report the average score and the hallucination rate, defined as the fraction of responses scored below 3. We keep the original rubric and replace the original GPT-based evaluator with the judge described in Appendix C.2.

POPE. POPE (Li et al., 2023b) reformulates object hallucination as binary questions of the form "Is there a *object* in the image?", posed about MS-COCO validation images. Questions about annotated objects form the positive set. Negative objects are drawn with three sampling strategies:

- *Random*: uniformly from the vocabulary;
- *Popular*: the most frequent objects in the dataset;
- *Adversarial*: objects that frequently co-occur with the ground-truth objects.

Each setting is balanced between "yes" and "no" answers. We report accuracy and F1 per setting and their average.

Table 7: **Fixed MSRC hyperparameters.** Only $(\tau_0, \tau_{\text{sub}})$ are calibrated; their backbone-specific values appear in Table 8.

Setting	Value
Generation and verification	
Response decoding	Greedy, beams $= 1$, `max_new_tokens` $= 512$
Repetition penalty	1.0 (backbone defaults disabled)
Re-grounding decoding	Greedy, ≤ 12 new tokens
Verifier score	$V = \sigma(\text{lse(yes)} - \text{lse(no)})$ at first answer position
Policy	
Claim scope	Benchmark (COCO) object vocabulary
Substitute vocabulary	Benchmark vocabulary; open vocabulary in ablation
Visual guard Γ_{vis}	On
Duplicate scope	Sentence-level; category-level
Semantic gate / mention coverage	Both on
Candidates per failed claim	($k = 3$ in ablation)
Precision floor	Point estimate; Clopper–Pearson bound in ablation
Calibration	
Risk level α	0.05 default; sweep $[0.01, 0.05, 0.10, , 0.15]$
Calibration / evaluation images	500 / 3000, disjoint
Selection margin η	0.15; budget $(1 - \eta)\alpha$
τ_0 search grid	51 quantiles of observed V
τ_{sub} search grid	Distinct substitute scores
Clopper–Pearson confidence	0.95, one-sided
Certification criterion	$\widehat{B}_{\text{cert}}(\hat{\lambda}) \leq \alpha$

MME. MME (Fu et al., 2023) evaluates perception and cognition through manually constructed yes/no questions, with two questions per image. Following the hallucination protocol of prior work (Leng et al., 2024; Huang et al., 2024), we use four perception subtasks: object-level *existence* and *count*, and attribute-level *position* and *color*. Each subtask score is the sum of question-level accuracy and image-level accuracy+, where accuracy+ requires both questions for an image to be answered correctly.

C.4 Hyperparameter Details

MSRC introduces no learned parameters. Only the thresholds $(\tau_0, \tau_{\text{sub}})$ are calibrated per backbone (Table 8); all other settings are fixed in advance, as summarized in Table 7. Decoding is greedy, and no hyperparameters are tuned on the evaluation split.

The selection margin η imposes a stricter budget $(1 - \eta)\alpha$ during selection, leaving headroom for certification at α on the held-out split. The thresholds are searched jointly because changing τ_0 alters the claims entering the repair branch and hence the population governed by τ_{sub}.

D Additional Ablation and Calibration Results

Calibration and repair behaviour. Table 8 reports the selected policy on 3,000 evaluation images. All three backbones satisfy the target risk level of $\alpha = 0.05$, with certified bounds of 0.038, 0.047, and 0.041, respectively. The selected substitution thresholds yield observed repair precision of 95.1%, 79.4%, and 96.6% on LLaVA, Qwen2.5, and InstructBLIP, with 204, 170, and 178 accepted substitutions, respectively. These results illustrate the backbone-dependent and selective nature of verified repair.

Risk–coverage trade-off. Table 9 evaluates the policy across three risk budgets. Increasing α from 0.02 to 0.10 raises coverage by 18.25, 12.43, and 13.08 percentage points on LLaVA, Qwen2.5, and InstructBLIP, respectively, while also increasing instance-level hallucination. Three of the nine operating points exceed their requested certification bounds and are explicitly marked.

Table 8: **Calibrated policy and repair behaviour**, $n = 3000$, $\alpha = 0.05$. Counts are over all claims in the evaluation captions.

Backbone	τ_0	τ_{sub}	Bound $\leq \alpha$	Kept	Deleted	Substituted	Sub. prec.
LLaVA-1.5 7B	0.761	0.876	0.038	18 978	4 510	204	95.1%
Qwen2.5-VL 3B	0.778	0.023	0.047	26 432	3 006	170	79.4%
InstructBLIP 7B	0.665	0.933	0.041	19 638	4 476	178	96.6%

Table 9: **Risk–coverage trade-off** on MS-COCO, $n = 3000$. † denotes operating points whose calibration bound exceeds the requested risk level.

Backbone	$\alpha = 0.02$			$\alpha = 0.05$			$\alpha = 0.10$		
	$C_S \downarrow$	$C_I \downarrow$	Cov.↑	$C_S \downarrow$	$C_I \downarrow$	Cov.↑	$C_S \downarrow$	$C_I \downarrow$	Cov.↑
LLaVA-1.5 7B	7.67	2.38	62.27	16.97	4.37	73.78	34.37	8.81	80.52
Qwen2.5-VL 3B	6.17†	1.57†	63.56	15.70	3.54	71.11	36.57†	7.99†	75.99
InstructBLIP 7B	5.77†	1.75†	63.65	15.10	3.90	71.55	30.53	7.81	76.73

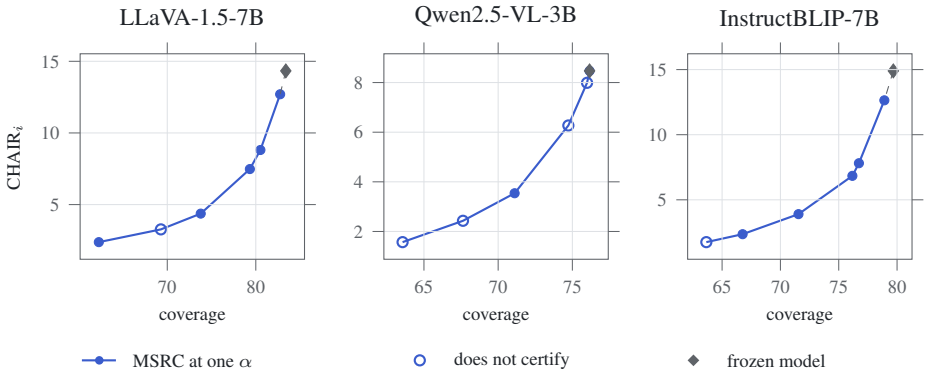

Figure 4: **Risk–coverage frontier on MS-COCO**, $n = 3000$. Filled markers indicate operating points certified on the held-out split.

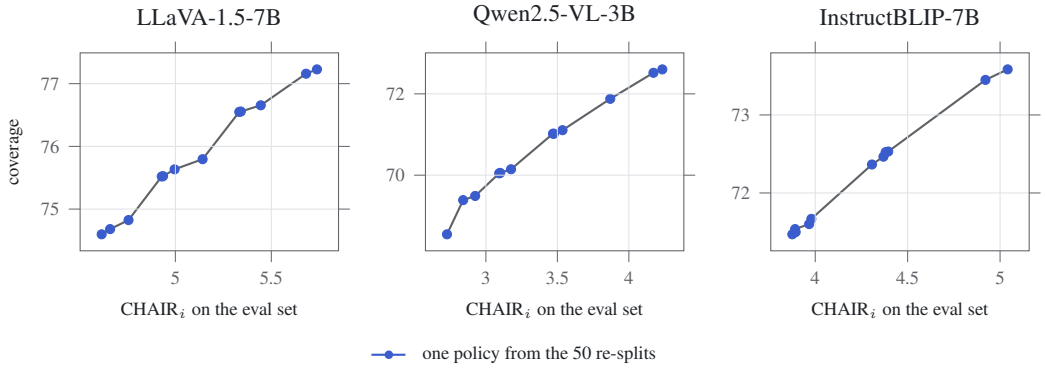

Figure 5: **Calibration stability.** Results from 50 re-splits of the calibration set, replayed on the evaluation set.

Risk–coverage and calibration stability. Figure 4 illustrates the trade-off between hallucination risk and retained coverage across operating budgets. Figure 5 examines 50 re-splits of the calibration set, with each resulting policy replayed on the evaluation set. Although the selected thresholds vary, the downstream results remain within a relatively narrow risk–coverage region, indicating that the overall operating behaviour is not determined solely by one calibration split.

Table 10: **BLIP-2 captioning and hallucination.** Lower is better for CH_i, CH_s, OCH, and $\bar{p}$. Compare OCH within, not across, the two evaluation groups.

Method	BLEU-4 ↑	CIDEr ↑	CH_i (%) ↓	CH_s (%) ↓	OCH (%) ↓	$\bar{p}$ ↓
BLIP-2	43.4	144.3	1.7	2.6	17.0	0.207
BLIP-2 + LURE	5.7	0.0	12.1	33.6	28.4	0.259
BLIP-2 + TLC-A	43.3	143.5	1.3	2.0	17.0	0.206
BLIP-2 + MOCHa	44.0	144.3	1.4	2.3	16.6	0.199
Our evaluation						
BLIP-2 + MSRC (Ours)	42.838	142.283	1.529	2.280	12.277	0.2113

Table 11: **Comparison with CEBC** on MS-COCO CHAIR, 500 evaluation images, `max_new_tokens` $= 20$. Bold indicates the better reported value between CEBC and MSRC on shared backbones.

VLM	Method	C_S ↓	C_I ↓	Recall↑	Prec.↑	F1↑
LLaVA-1.5 7B	Base	5.40	3.42	43.57	96.58	60.05
	CEBC	1.96	1.62	30.18	98.38	46.20
	MSRC (Ours), $\alpha = 0.02$	**1.80**	**1.40**	**39.06**	**98.60**	**55.95**
InstructBLIP 7B	Base	3.40	2.26	43.26	97.74	59.98
	CEBC	**1.57**	1.27	33.10	98.73	49.57
	MSRC (Ours), $\alpha = 0.02$	1.60	**1.08**	**40.85**	**98.92**	**57.83**
Qwen2-VL 7B/Qwen2.5-VL 3B	Base	1.80	1.44	28.18	98.56	43.83
	CEBC	1.80	1.55	**40.55**	98.45	**57.44**
	MSRC (Ours), $\alpha = 0.02$	**0.20**	**0.19**	24.29	**99.81**	39.07

E Open-Vocabulary Evaluation on BLIP-2

Results. OpenCHAIR evaluates hallucinated object mentions beyond the fixed MS-COCO vocabulary Ben-Kish et al. (2024). As shown in Table 10, MSRC achieves $\mathrm{CH}_i = 1.529\%$, $\mathrm{CH}_s = 2.280\%$, and $\mathrm{OCH} = 12.277\%$. Relative to the reported BLIP-2 baseline, the CHAIR scores are numerically lower by 0.171 and 0.320 percentage points, respectively, while BLEU-4 and CIDEr decrease by 0.562 and 2.017 points. The mean NLI contradiction probability is slightly higher (0.2113 versus 0.207). Since the published baselines and MSRC belong to separate evaluation groups, these differences are descriptive, and OCH scores should not be directly compared across groups. Moreover, because OpenCHAIR evaluates final captions, it cannot isolate the contribution of verified substitution from filtering alone.

F Comparison with Conformal Baseline

CEBC Mishra et al. (2026) uses an external DETR-ResNet50 detector to verify and revise generated captions. In contrast, MSRC obtains verification evidence from the generating VLM and calibrates claim-level verification and repair. For comparison, we rerun MSRC using greedy decoding and a 20-token generation budget. The published CEBC results use a different 500-image evaluation split; therefore, the comparison is not fully paired. *Note:* CEBC reports results on Qwen2-VL-7B, whereas our additional evaluation uses Qwen2.5-VL-3B. These are distinct backbones and should not be treated as a directly matched comparison.

Results. Table 11 shows that MSRC reports lower instance-level CHAIR than CEBC on all three shared backbones, with reductions of 0.22, 0.19, and 1.36 points on LLaVA, InstructBLIP, and Qwen2-VL, respectively. Object recall also increases by 8.88 and 7.75 points on LLaVA and InstructBLIP, with corresponding F1 gains of 9.75 and 8.26 points. However, Qwen2-VL recall decreases by 16.26 points relative to CEBC. Unlike detector-based methods that can introduce omitted objects, MSRC only repairs existing claims and cannot recover objects absent from the original caption. These results illustrate the hallucination–coverage trade-off while demonstrating detector-free verification under an aligned decoding budget.

Table 12: **POPE hallucination evaluation on MS-COCO.** Comparison of the proposed MSRC method with Vanilla decoding, VCD, and OPERA across three vision-language backbones: LLaVA-1.5-7B, Qwen2.5-VL-3B, and InstructBLIP-7B. Performance is evaluated using accuracy (Acc.) and F1-score under Random, Popular, and Adversarial negative sampling strategies. Average scores summarize performance across the three evaluation settings. Higher values indicate better object hallucination mitigation, and the best results for each backbone are highlighted in bold.

Backbone	Method	Random		Popular		Adversarial		Average	
		Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑	Acc. ↑	F1 ↑
LLaVA-1.5-7B	Vanilla	83.20	81.30	81.80	80.00	78.90	77.50	81.30	79.60
	VCD	87.70	87.10	85.30	85.00	80.80	81.30	84.60	84.40
	OPERA	89.10	89.00	86.00	86.30	79.10	80.90	84.70	85.40
	MSRC (ours)	**91.67**	**91.44**	**87.47**	**87.22**	**85.37**	**84.79**	**88.17**	**87.82**
Qwen2.5-VL-3B	Vanilla	84.70	82.60	84.10	82.00	82.20	80.30	83.60	81.60
	VCD	88.60	87.80	**87.10**	**86.40**	**84.20**	**83.90**	86.60	86.00
	OPERA	86.10	84.20	85.70	83.80	83.90	82.10	85.20	83.30
	MSRC (ours)	**91.57**	**91.55**	85.80	85.88	**84.20**	**83.90**	**87.19**	**87.11**
InstructBLIP-7B	Vanilla	80.70	80.40	78.20	78.30	75.80	76.50	78.20	78.40
	VCD	84.50	83.60	81.40	81.00	79.50	79.50	81.80	81.30
	OPERA	**89.80**	**89.60**	83.40	84.00	80.70	81.80	84.60	85.10
	MSRC (ours)	89.53	89.35	**86.63**	**86.74**	**84.13**	**84.28**	**86.76**	**86.79**

Table 13: **MME hallucination evaluation** across vision-language backbones. Results cover object-level (existence and count) and attribute-level (position and color) subtasks, each with a maximum score of 200 (total: 800). Bold indicates the highest score within each backbone among multiple evaluated methods.

LVLM	**Method**	**Object-level**		**Attribute-level**		
		Existence ↑	**Count** ↑	**Position** ↑	**Color** ↑	**Total** ↑
LLaVA-1.5-7B	Vanilla	175.6	124.6	114.0	151.0	565.3
	VCD	184.6	138.3	128.6	153.0	604.6
	OPERA	180.6	133.3	123.3	155.0	592.3
	MSRC	**196.0**	**155.0**	**143.3**	**187.6**	**681.9**
InstructBLIP-7B	Vanilla	141.0	75.3	66.6	97.3	380.3
	VCD	**168.3**	**92.3**	64.0	123.0	447.6
	OPERA	156.0	78.3	55.0	95.0	384.3
	MSRC	**168.3**	80.0	**83.3**	**129.7**	**461.3**
Qwen-VL	Vanilla	155.0	127.6	131.6	173.0	587.3
	VCD	156.0	131.0	128.0	**181.6**	596.6
	OPERA	**165.0**	**145.0**	**133.3**	180.0	**623.3**
Qwen2.5-VL-3B	MSRC	190.0	160.0	165.0	180.0	695.0

G DISCRIMINATIVE RESULTS

POPE Hallucination Evaluation. Table 12 presents the POPE results on MS-COCO across three vision-language backbones. MSRC achieves the highest average accuracy and F1-score for all three models. For LLaVA-1.5-7B, MSRC achieves 88.17% accuracy and 87.82% F1-score, improving over the strongest baselines by 3.47 and 2.42 percentage points, respectively. For Qwen2.5-VL-3B, MSRC achieves 87.19% accuracy and 87.11% F1-score, outperforming VCD by 0.59 and 1.11 percentage points. Similarly, on InstructBLIP-7B, MSRC reaches 86.76% accuracy and 86.79% F1-score, improving over OPERA by 2.16 and 1.69 percentage points. Although some baselines perform better in individual settings, MSRC maintains higher overall performance. These results show that MSRC effectively reduces object hallucinations across different backbones, particularly under challenging Popular and Adversarial sampling.

MME Hallucination Evaluation. Table 13 reports performance on four hallucination-related MME subtasks, covering object- and attribute-level grounding. On LLaVA-1.5-7B, MSRC achieves the highest scores across all four subtasks, reaching a total of 681.9, an improvement of 116.6 points over Vanilla and 77.3 points over VCD. On InstructBLIP-7B, MSRC obtains 461.3, improving over Vanilla by 81.0 points and achieving the highest total score among the evaluated methods. It also attains the highest position and color scores while matching VCD on existence. On Qwen2.5-VL-3B, MSRC achieves a total score of 695.0. Since the reported Qwen-VL baselines use a different backbone, these results are not directly comparable. Overall, the results support the effectiveness of MSRC in mitigating object- and attribute-level hallucinations across different vision-language backbones.

H Qualitative Examples

All three lanes are replayed from the same frozen verifier scores used for the quantitative tables, so Vanilla, VOF (Ours) and MSRC (Ours) differ only in the rule applied to a claim that fails verification. Each lane's *complete* output is printed, not an excerpt, so no part of a response is hidden by selective quotation.

How these examples were selected. They are chosen, and we state the rule so that it can be checked. From the 3000-image evaluation split we took every image on which the calibrated policy performed exactly one substitution and no deletion, the replaced object is *absent* from the COCO annotation, and the substitute is *present* in it. Fifty-two images satisfy this; eight are shown. The examples therefore illustrate the intended behaviour of verified repair rather than its average behaviour. The aggregate results, including the operating points at which MSRC (Ours) retains less content than VOF (Ours) and the budgets at which neither lane certifies, are reported in Section 4.1, Figure 2 and Appendix D.

H.1 Captioning

In each case the verifier scores the original object below the keep gate. VOF (Ours) can only remove the noun, which leaves the surrounding sentence ungrammatical; MSRC (Ours) re-grounds the claim, re-verifies the candidate, and substitutes it into the same span, so the sentence remains well formed and the recovered object is one the annotation contains. Every other sentence is produced identically by the three lanes, which the full text below makes visible.



LLaVA-1.5-7B, COCO 179322. Annotated objects: bowl, cup, microwave, person. `refrigerator` scores $V = 0.15$ and is absent from the annotation; the re-grounded replacement `microwave` re-verifies at $V = 0.96$ and is present in it.

Vanilla: The image features a man standing in a kitchen, holding a microwave oven door open. He is looking at the microwave, possibly examining its contents or checking its function. The microwave is placed on the left side of the scene, while the man is positioned more towards the center. In the background, there is a refrigerator on the right side of the kitchen. Additionally, there are two cups placed on the countertop, one near the center and the other towards the right side of the scene.

VOF (Ours): The image features a man standing in a kitchen, holding a microwave oven door open. He is looking at the microwave, possibly examining its contents or checking its function. The microwave is placed on the left side of the scene, while the man is positioned more towards the center. In the background, there is on the right side of the kitchen. Additionally, there are two cups placed on the countertop, one near the center and the other towards the right side of the scene.

MSRC (Ours): The image features a man standing in a kitchen, holding a microwave oven door open. He is looking at the microwave, possibly examining its contents or checking its function. The microwave is placed on the left side of the scene, while the man is positioned more towards the center. In the background, there is a microwave on the right side of the kitchen. Additionally, there are two cups placed on the countertop, one near the center and the other towards the right side of the scene.

LLaVA-1.5-7B, COCO 353505. Annotated objects: bear, bench, handbag, person, teddy bear. `chair` scores $V = 0.71$ and is absent from the annotation; the re-grounded replacement `bench` re-verifies at $V = 0.95$ and is present in it.



Vanilla: The image features a woman and a young girl standing outside, surrounded by three teddy bears. The woman is kneeling down, possibly helping the girl with the teddy bears. The teddy bears are placed on a bench, with one on the left side, one in the middle, and the third one on the right side. In the background, there is a chair located near the left side of the scene. The woman is wearing a hat, and the girl is holding a handbag. The scene appears to be a casual, outdoor gathering with the woman and the girl enjoying their time together.

VOF (Ours): The image features a woman and a young girl standing outside, surrounded by three teddy bears. The woman is kneeling down, possibly helping the girl with the teddy bears. The teddy bears are placed on a bench, with one on the left side, one in the middle, and the third one on the right side. In the background, there is located near the left side of the scene. The woman is wearing a hat, and the girl is holding a handbag. The scene appears to be a casual, outdoor gathering with the woman and the girl enjoying their time together.

MSRC (Ours): The image features a woman and a young girl standing outside, surrounded by three teddy bears. The woman is kneeling down, possibly helping the girl with the teddy bears. The teddy bears are placed on a bench, with one on the left side, one in the middle, and the third one on the right side. In the background, there is a bench located near the left side of the scene. The woman is wearing a hat, and the girl is holding a handbag. The scene appears to be a casual, outdoor gathering with the woman and the girl enjoying their time together.

InstructBLIP-7B, COCO 499388. Annotated objects: car, person. `truck` scores $V = 0.23$ and is absent from the annotation; the re-grounded replacement `car` re-verifies at $V = 0.99$ and is present in it.



Vanilla: The image features a residential street with a car parked on the side of the road. The car is parked near a street sign, indicating the location as Park Place. There are several houses visible in the background, with some of them located closer to the car and others further away. A truck is also be seen in the scene, parked on the other side of the street. The street is lined with lights, adding to the overall serene atmosphere of the neighborhood.

VOF (Ours): The image features a residential street with a car parked on the side of the road. The car is parked near a street sign, indicating the location as Park Place. There are several houses visible in the background, with some of them located closer to the car and others further away. truck also be seen in the scene, parked on the other side of the street. The street is lined with trees, adding to the overall serene atmosphere of the neighborhood.

MSRC (Ours): The image features a residential street with a car parked on the side of the road. The car is parked near a street sign, indicating the location as Park Place. There are several houses visible in the background, with some of them located closer to the car and others further away. A car can also be seen in the scene, parked on the other side of the street. The street is lined with trees, adding to the overall serene atmosphere of the neighborhood.

H.2 Question Answering

Both questions rest on a false premise. Vanilla accepts the premise and invents a detail to match it. VOF (Ours) can only delete: in the first case the answer is a single sentence, so deletion leaves the invented assertion standing unchanged, and in the second it returns a blanket refusal that the reference does not accept. MSRC (Ours) names the object that is actually absent.



HaloQuest 35, false premise. *"How many fish are swimming next to the cat?"*
Reference: There are no fish in this image.

× **Vanilla:** There are two fish swimming next to the cat.
× **VOF (Ours):** There are two fish swimming next to the cat.
✓ **MSRC (Ours):** There is no fish in the image.



HaloQuest 513, false premise. *"How many frogs are there?"*
Reference: 1. There are no frogs in this image.; 2. There is no frog that is visible on the photo.; 3. The frog doesn't exist.
× **Vanilla:** There are two frogs in the image.
× **VOF (Ours):** The requested detail is not supported by the image.
✓ **MSRC (Ours):** There is no frog in the image.

H.3 YES/NO QUESTION ANSWERING



LLaVA-1.5-7B, MMHal-Bench 68.
"Is the mouse connected to a computer?"
Reference: No, the mouse is not connected to a computer.
Vanilla: Yes, the mouse is connected to a computer, as it is plugged into a USB port.
MSRC (Ours): No, the mouse is not connected to a computer. It is sitting on a table, unattached to any device.



InstructBLIP-7B, MMHal-Bench 33.
"Is the jam on the bread made of strawberry?"
Reference: In fact, this photo only shows a cut and toasted bread slice. There is no jam on it.
Vanilla: No, it is not. The toast does not have any jam on it.
MSRC (Ours): No, it is not. The toast does not have any jam or food on it. It simply has a plain surface without any additional toppings or foods.