EXPLORING AND MITIGATING OBJECT HALLUCINATION WITH NON-AUTOREGRESSIVE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-Language Models (VLMs) support a growing range of multimodal applications, yet their generated descriptions remain prone to object hallucination. A conservative caption can avoid unsupported mentions by saying less, but the resulting loss of coverage limits usefulness; this tension appears clearly across representative generation paradigms. Autoregressive VLMs such as LLaVA-1.5 describe more of the scene but hallucinate more objects, whereas non-autoregressive masked-diffusion VLMs such as LLaDA-V achieve much lower hallucination but leave more visible content unmentioned. We trace this contrast to how the two generation architectures resolve uncertainty: autoregressive decoding makes each emitted token an immediate, irreversible part of the context, whereas masked diffusion can revise low-confidence predictions before fixation. We formalize this distinction as *selective object commitment*, separating broad object discovery from the final decision to assert each candidate. Specifically, we decompose final commitment into proposal admission, diffusion realization, and omission recovery, and derive the stagewise selectivity needed to expand coverage without proportionally increasing hallucination risk. Guided by this formulation, we introduce COVER (Cross-view Object Verification and Evidence-guided Recovery), a training-free method that combines autoregressive and diffusion proposal views, builds an object plan from cross-view agreement and object-level real/null visual gain, applies the same contrast through Visual Residual Guided diffusion decoding (VRG) to guide token selection and fixation, and restores verified omissions through bounded recovery. On MSCOCO evaluations, COVER reduces sentence-level CHAIR by 28.1% while retaining 98.7% of the detailed proposer's object Recall. COVER establishes proposal–commitment separation as a practical cross-architecture principle for grounded multimodal generation. *The source code for our method will be made available upon publication.*

1 INTRODUCTION

Vision-language models (VLMs) are widely used to describe, search, and reason about visual content (Liu et al., 2023). Yet they confront a failure, dubbed *object hallucination*, in which they generate object mentions unsupported by the visual input (Rohrbach et al., 2018; Liu et al., 2024a). For instance, a caption may mention a dog beside a pedestrian even though no dog is visible. Because such errors can remain linguistically coherent, they corrupt visual facts, mislead users and downstream systems, and weaken trust in VLM-generated descriptions.

Existing work has proposed several mitigations that primarily focus on AutoRegressive (AR) based language models. VCD and AGLA alter visual evidence; DoLa, OPERA, and AdaIAT reshape token scores, layer evidence, or attention; OPA-DPO learns from corrected preferences; and LURE revises completed captions (Leng et al., 2024; An et al., 2025; Chuang et al., 2024; Huang et al., 2024; Zhong et al., 2026; Yang et al., 2025; Zhou et al., 2024). Existing methods can reduce hallucination by constraining VLMs to output only high-confidence objects while suppressing low-confidence ones. However, such improvements often come at the cost of reduced coverage, as potentially risky objects are deliberately excluded to minimize hallucination. Consequently, existing approaches struggle to achieve a desirable trade-off between low hallucination and high coverage, meaning that a low

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.

hallucination rate alone does not necessarily guarantee an informative and useful description. The problem is therefore two-sided: autoregressive VLMs introduce visually unsupported objects, while overly aggressive hallucination mitigation may erase supported ones.

Resolving this tension requires distinguishing visually unsupported assertions from visual grounded detail rather than suppressing both hallucination rate and recall rate. We begin by comparing autoregressive and non-autoregressive VLMs. Captions from the autoregressive LLaVA-1.5 model cover much of the annotated scene but hallucinate frequently, whereas the non-autoregressive masked-diffusion model LLaDA-V produces far fewer unsupported objects while leaving more visible content unmentioned (Liu et al., 2024b; You et al., 2025). Our comparison also shows that low hallucination is not unique to masked diffusion: restricting LLaVA-1.5 to only salient objects yields an even stronger conservative endpoint than direct LLaDA-V, but both omit much of the scene. The relevant distinction lies instead in how uncertainty is resolved: autoregressive decoding fixes each emitted token immediately, whereas masked diffusion can revise low-confidence positions before fixation. We therefore test whether richer hints can make LLaDA-V mention more visually supported objects without increasing unsupported mentions at the same rate. Our *Caption Augmentation* (CaptionAug) intervention first obtains an auxiliary LLaDA-V description and then conditions its caption on the following task-specific answer generation. Without changing the checkpoint, CaptionAug substantially expands grounded coverage and improves the local trade-off relative to a comparably conservative autoregressive setting, showing that the conservative diffusion model can move toward richer descriptions while retaining stronger discrimination between supported and unsupported objects.

Object-level diagnostics further show that CaptionAug increases LLaDA-V's ability to distinguish supported from unsupported objects. We call this behavior *selective object commitment*: as coverage expands, visually supported objects should enter the final caption more readily than unsupported ones. Unsupported LLaVA-1.5 objects are emitted with lower probability and higher entropy, whereas unsupported LLaDA-V predictions fix later and change more often while masked. These dynamics motivate a design principle: use autoregressive decoding to expose broad candidates and revisable non-autoregressive generation to control which candidates become final assertions.

Guided by this principle, we introduce COVER (Cross-view Object Verification and Evidence-guided Recovery), a training-free framework using two frozen VLMs. LLaVA-1.5 supplies broad proposals; direct and CaptionAug LLaDA-V captions supply conservative and expanded diffusion views. Source signatures and image-conditioned visual gain define an object plan, Visual Residual Guided decoding (VRG) realizes a selective scaffold through real/null logit contrasts, and bounded recovery restores verified planned omissions. COVER thereby realizes selective object commitment as a cross-architecture division of labor, preserving broad discovery while controlling which objects become final assertions.

Across MSCOCO, AMBER, Object HalBench, and OpenCHAIR, COVER establishes a favorable high-coverage operating point. On matched MSCOCO samples, it reduces $(\mathrm{CHAIR}_s, \mathrm{CHAIR}_i)$ from $(0.597, 0.170)$ to $(0.429, 0.129)$ while retaining 0.786 object Recall versus the proposer's 0.797. Stage analyses show complementary proposal evidence from agreement and gain, selective suppression by VRG, and controlled restoration by recovery. These results support "proposal–commitment" separation as a practical cross-architecture principle for grounded generation.

Our contributions are threefold:

- **Bridging generation architecture and hallucination uncertainty.** We are the first to connect the hallucination–coverage trade-off in VLM captioning to when uncertain object hypotheses become irreversible: autoregressive decoding commits each emission immediately, whereas masked diffusion repeatedly revises low-confidence predictions before fixation. We formalize this cross-architecture phenomenon as selective object commitment.
- **COVER: cross-architecture selective commitment.** Guided by this finding, we introduce COVER, which reinterprets broad autoregressive captions as object proposals and operationalizes final commitment through cross-view proposal admission, visual-residual-guided masked-diffusion realization, and capped omission recovery, without additional training.
- **Multi-benchmark validation.** Three annotation-based benchmarks and an open-vocabulary transfer evaluation characterize the hallucination–coverage trade-off and isolate the roles of proposal admission, diffusion realization, and recovery.

2 Related Work

Object hallucination and generative evaluation. Object hallucination is the generation of object mentions unsupported by the visual input, a specific cross-modal faithfulness error within the broader taxonomy of VLM hallucination (Liu et al., 2024a). Existing evaluations can be organized along two axes: whether a model discriminates object presence or freely generates a description, and whether visual support is judged against a closed ontology or an open vocabulary. For open-ended captioning, generative protocols are the relevant case. CHAIR and Object HalBench are closed-vocabulary, annotation-based protocols that map mentions to the MSCOCO object ontology, enabling sentence-level error, mention-level error, and object recall to be measured together (Rohrbach et al., 2018; Lin et al., 2014; Yu et al., 2024). AMBER broadens annotation-based evaluation to objects, attributes, and relations and couples its generative task with a separate discriminative task (Wang et al., 2023). OpenCHAIR instead extracts concrete nouns from free-form captions and judges their visual support in an open vocabulary (Ben-Kish et al., 2024). These protocols expose different failure modes and are not interchangeable: grounding gains under constrained evaluation need not transfer to open-ended caption generation (Geigle et al., 2024). A shared confound is output conservatism, since a model can reduce hallucination simply by mentioning less. Hallucination should therefore be paired with recall when exhaustive annotations permit and with object-density proxies otherwise.

Hallucination mitigation in autoregressive VLMs. Prior work mitigates autoregressive VLM hallucination through training-time alignment or inference-time intervention (Liu et al., 2024a). At inference, VCD and AGLA alter the input or visual view, whereas DoLa, OPERA, ICT, AdaIAT, and logit-lens-guided methods reshape token scores or internal representations (Leng et al., 2024; An et al., 2025; Chuang et al., 2024; Huang et al., 2024; Chen et al., 2025; Zhong et al., 2026; Wang et al., 2026); OPA-DPO instead learns from on-policy corrected preferences (Yang et al., 2025). Post-hoc systems revise completed captions: LURE trains a hallucination revisor, while Woodpecker invokes an external visual-knowledge correction pipeline (Zhou et al., 2024; Yin et al., 2024). These approaches intervene within an autoregressive trajectory or after it has produced a description. A complementary direction is to assign broad object proposal and grounded realization to two frozen generation paradigms, using cross-view evidence before the final caption without training a task-specific revisor or invoking an external tool chain. In all cases, removed hallucinations must be considered together with preserved grounded content.

Diffusion language and multimodal models. The important distinction between autoregressive and diffusion language models is not parallel prediction alone, but when token decisions become fixed. Mask-Predict established iterative parallel refinement with confidence-based remasking (Ghazvininejad et al., 2019). Discrete diffusion subsequently gave token corruption and denoising a probabilistic formulation (Austin et al., 2021); masked diffusion specializes this framework to an absorbing mask state and masked-token objectives (Sahoo et al., 2024), and LLaDA scales the formulation to large language modeling (Nie et al., 2025). Diffusion-LM represents a separate continuous-state route to controllable text generation (Li et al., 2022). In the multimodal setting, LaViDa and LLaDA-V condition masked-diffusion language models on visual encoders for instruction following (Li et al., 2025; You et al., 2025). Together, these works establish iterative denoising as a viable alternative to left-to-right generation, with unresolved positions remaining revisable before commitment. These properties suggest a complementary use of pretrained masked-diffusion VLMs: iterative denoising can provide a conservative caption scaffold and an image-conditioned verification process for object proposals generated by a high-coverage autoregressive model.

3 From Generation Architecture to Selective Object Commitment

Low hallucination may reflect accurate grounding or simply fewer object mentions. We therefore begin by asking whether an autoregressive VLM can suppress hallucination through shorter, more conservative captions without sacrificing visible content.

3.1 Autoregressive Generation: Higher Coverage at the Cost of Hallucination

Shorter captions may discourage unsupported objects but also erase grounded content. We test this by varying the same LLaVA model between *strict* and *detailed* captioning. Table 1 shows that the

Table 1: Matched MSCOCO comparison of conservative and expanded generation.

Setting	$\mathrm{CHAIR}_s \downarrow$	$\mathrm{CHAIR}_i \downarrow$	Recall↑	Avg. objects
LLaVA-1.5, strict caption	0.080	0.0579	0.5163	1.936
LLaVA-1.5, main objects	0.146	0.0724	0.5826	2.292
LLaVA-1.5, detailed (64 tokens)	0.242	0.0828	0.6126	4.422
LLaVA-1.5, detailed (96 tokens)	0.528	0.1514	0.7290	6.660
LLaVA-1.5, detailed (128 tokens)	0.642	0.1860	0.7816	8.138
LLaDA-V, direct	0.122	0.0627	0.4590	3.064
LLaDA-V, caption-augmented	0.202	0.0677	0.6390	4.020

detailed policy raises Recall by 51.4%, but increases CHAIR_s more than eightfold. This steep coupling is consistent with autoregressive factorization:

$$p_{\mathrm{AR}}(y \mid I) = \prod_{t=1}^{T} p(y_t \mid y_{<t}, I). \tag{1}$$

Once an object is emitted, it becomes fixed context for later predictions and cannot be revised. This raises the next question: if irreversible left-to-right generation contributes to the trade-off, can a revisable non-autoregressive architecture reach a different regime?

3.2 Non-Autoregressive Masked Diffusion: Revisable but Incomplete

We turn to LLaDA-V, a non-autoregressive masked-diffusion VLM that revises unresolved tokens before fixation. Direct generation yields low hallucination but only 0.4590 Recall, indicating an overly conservative strategy. Moreover, Table 1 shows that strict LLaVA-1.5 achieves both lower hallucination and higher Recall, so direct performance alone does not establish an architectural advantage. What masked diffusion offers is a revisable trajectory in which uncertain positions remain editable before fixation. We therefore test whether LLaDA-V can move toward a more descriptive operating point without losing its ability to reject unsupported objects. We leverage *Caption Augmentation* (*CaptionAug*) for masked diffusion, adapting description-conditioned generation (Ghosh et al., 2025). The model first produces an auxiliary description $d = f_\theta(I; p_{\mathrm{desc}})$ and then generates $c_{\mathrm{AUG}} = f_\theta(I, d; p_{\mathrm{cap}})$, encouraging visible but uncertain objects to enter the second pass.

CaptionAug raises LLaDA-V Recall by 39.2%. At comparable coverage, it attains slightly higher Recall than 64-token detailed LLaVA-1.5 with 16.5% lower CHAIR_s, showing that richer conditioning improves the hallucination–coverage trade-off.

Figure 1 connects this improvement to uncertainty and fixation. Direct LLaDA-V has a supported versus unsupported commitment ratio $3.6\times$ that of LLaVA-1.5 (4.07 versus 1.12), and CaptionAug raises it by a further 19.7% to 4.87. Unsupported autoregressive emissions are less confident but irreversible; masked diffusion instead fixes unsupported objects later and revises them more often. These results show asymmetric expansion: supported commitment rises more strongly than

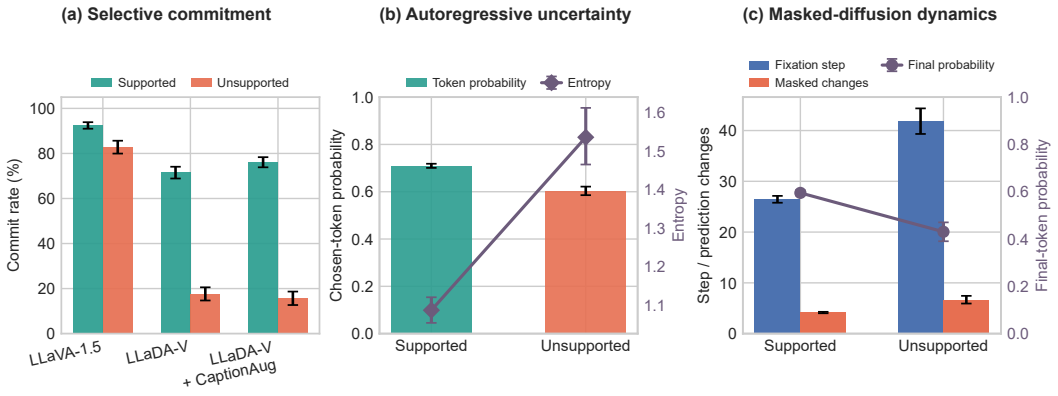

Figure 1: Object commitment and uncertainty across autoregressive and masked-diffusion generation.

unsupported commitment. We refer to this property as *selective object commitment*, in which coverage expands without weakening the distinction between visually supported and unsupported objects.

CaptionAug nevertheless remains below the broad autoregressive endpoint in Recall. This leaves the central question: how can broad object discovery and selective verification be combined without inheriting hallucinations or omissions?

3.3 Selective Commitment Separates Discovery from Assertion

To answer this question and operationalize selective object commitment, we combine the two architectures while assigning them different roles: an autoregressive model proposes a broad object set, while a masked-diffusion model verifies and realizes it. For an image and object pair with support label Z and $\pi = P(Z = 1)$, let Q_m denote candidate admission, S_m initial realization, and H_m recovery of an admitted omission under policy m. With $S_m \le Q_m$, $H_m \le Q_m(1 - S_m)$, and final commitment $C_m = S_m + H_m$, the corresponding rates $q_{m,z}$, $r_{m,z}$, and $h_{m,z}$ give

$$c_{m,z} = q_{m,z}\big[r_{m,z} + (1 - r_{m,z})h_{m,z}\big], \qquad R_m = \left[1 + \frac{\pi}{1-\pi}\frac{c_{m,1}}{c_{m,0}}\right]^{-1}, \tag{2}$$

where $c_{m,z} = P(C_m = 1 \mid Z = z)$ and $R_m = P(Z = 0 \mid C_m = 1)$ is the hallucination risk of a committed candidate. High coverage with low risk therefore requires more supported commitments without reducing their ratio to unsupported commitments.

Proposition 1 (Expansion and selectivity)**.** *For two policies on the same pair distribution with positive class-specific rates, $R_{m'} \le R_m$ and $c_{m',1} \ge c_{m,1}$ hold if and only if*

$$\frac{c_{m',1}}{c_{m',0}} \ge \frac{c_{m,1}}{c_{m,0}}, \qquad c_{m',1} \ge c_{m,1}. \tag{3}$$

Writing $v_{m,z} = r_{m,z} + (1 - r_{m,z})h_{m,z}$, we obtain

$$\frac{v_{m',1}}{v_{m',0}} \ge \frac{v_{m,1}}{v_{m,0}}\frac{q_{m,1}/q_{m,0}}{q_{m',1}/q_{m',0}}. \tag{4}$$

This condition inspires us to base every downstream decision on evidence: selectivity lost by expanding autoregressive proposals must be restored when LLaDA-V admits, realizes, or recovers them. Proofs and boundary cases appear in Appendices L–M.

3.4 Evidence-Gated Admission and Recovery Control Risk

We translate Equation 4 into two design principles. First, we combine source agreement with image-conditioned evidence so that broader proposals preferentially admit supported objects. For source tier s and evidence-pass event E_t, Bayes' rule gives

$$\mathrm{Odds}(Z = 1 \mid s, E_t) = \mathrm{Odds}(Z = 1 \mid s)\frac{P(E_t \mid Z = 1, s)}{P(E_t \mid Z = 0, s)}. \tag{5}$$

We therefore admit a broad autoregressive candidate only when cross-view agreement or a visual test supplies sufficient evidence.

Second, selective realization can still omit supported candidates, so we recover only verified omissions. Let M_0, F_0 be the total and unsupported scaffold mentions; if recovery adds k_j objects to image j, including f_j unsupported additions, with $k = \sum_j k_j$ and $f = \sum_j f_j$, then

$$\Delta\mathrm{CHAIR}_i = \frac{k}{M_0 + k}\left(\frac{f}{k} - \frac{F_0}{M_0}\right), \qquad \Delta\mathrm{Recall}_j = \frac{k_j - f_j}{|G(I_j)|}. \tag{6}$$

The full recovery identities are stated in Appendix L.3. They expose a one-way trade-off: supported additions raise Recall, while unsupported additions govern both mention- and sentence-level risk. We therefore recover only verified plan omissions and jointly constrain their evidence and volume.

Overall, our formulation converts captioning into three controlled decisions: expand proposals, verify and realize them selectively, and recover verified omissions under a bounded risk budget. Section 4 instantiates these principles as COVER.

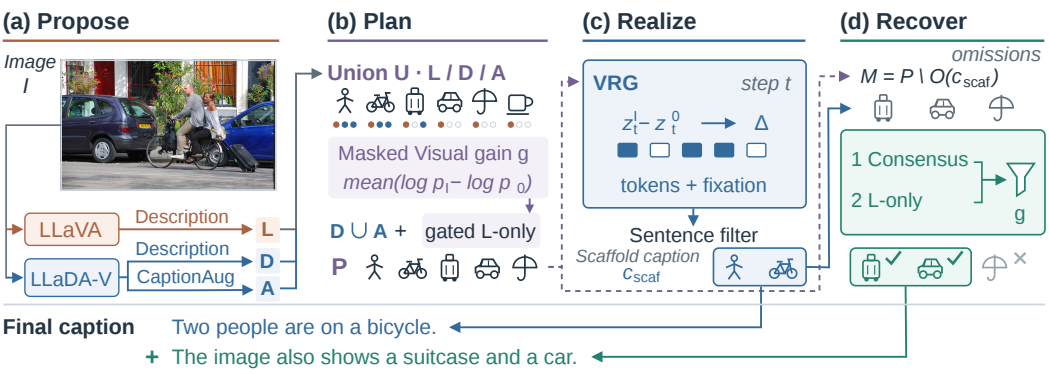

Figure 2: Overview of COVER.

4 THE COVER FRAMEWORK

We propose COVER (Cross-view Object Verification and Evidence-guided Recovery), a training-free framework that converts a one-shot autoregressive description into an evidence-gated, revisable commitment process. Rather than accepting the autoregressive caption as final, COVER reinterprets its objects as proposals and delegates final commitment to a masked-diffusion VLM. Figure 2 summarizes four stages: Propose, Plan, Realize, and Recover.

4.1 FILTERING BROAD PROPOSALS THROUGH CROSS-VIEW PLANNING

Because LLaDA-V is selective but incomplete and LLaVA-1.5 is broad but noisy, we use LLaVA as the high coverage proposer (view L) and LLaDA-V as the verifier, producing direct view D and CaptionAug view A. From their canonical object sets we form $\mathcal{U} = O^{\mathrm{L}} \cup O^{\mathrm{D}} \cup O^{\mathrm{AUG}}$ and retain each candidate's source signature. The plan begins with the diffusion union $P_0 = O^{\mathrm{D}} \cup O^{\mathrm{AUG}}$; an L-only candidate enters only if it passes a visual test conditioned on the image.

We compute that evidence by masking object phrase $o_{1:m}$ in context u and comparing its token probabilities under real image I and zero-image tensor I_0:

$$g(o; I, u) = \frac{1}{m} \sum_{j=1}^{m} \log \frac{p_{\theta,j}(o_j \mid u, \widetilde{o}, I)}{p_{\theta,j}(o_j \mid u, \widetilde{o}, I_0)}. \tag{7}$$

We admit at most K_P L-only objects above threshold τ_P, ranked by gain. Agreement supplies a reliability prior, while gain prevents the broad proposer from transferring unsupported objects unchecked. Together, they implement the evidence based admission required by Equation 5.

4.2 REALIZING THE PLAN THROUGH VISUAL RESIDUAL GUIDANCE

Admission alone cannot prevent a visually weak token from appearing during realization. We therefore condition LLaDA-V on plan P and guide every denoising step with a real/null residual. For paired logits z_t^I, z_t^0, we clip $r_t = z_t^I - z_t^0$ and form positive top-k boost $b_t = \mathrm{TopK}([r_t - m]_+, k)$:

$$\widetilde{z}_t = z_t^I + \alpha b_t. \tag{8}$$

VRG also raises the fixation priority of selected tokens with positive residual evidence, letting the image affect both token choice and commitment time. We then remove sentences containing out-of-plan objects to obtain the filtered diffusion scaffold c_{scaf}. Together, residual guidance and plan filtering suppress weak commitments before they enter the final caption. Full guidance equations appear in Appendix I.

4.3 RESTORING VERIFIED OMISSIONS WITH BOUNDED RECOVERY

The conservative scaffold may still omit supported plan objects, so we recover only $M = P \setminus O(c_{\mathrm{scaf}})$ rather than reopening generation over the full vocabulary. Consensus omissions receive priority,

Algorithm 1 COVER inference

Input: Image I, null image I_0, frozen proposer f_{L} and verifier f_{D};
gates τ_P, τ_G, image quantile ρ_I, caps K_P, K_R, K_L, decoding settings
Output: Final caption c

STEP 1 · PROPOSE
1: Generate detailed c^{L}, direct c^{D}, and description-conditioned c^{AUG}
2: Extract ordered $O^{\mathrm{L}}, O^{\mathrm{D}}, O^{\mathrm{AUG}}$ and source signatures $s(o)$

STEP 2 · VERIFY + PLAN
3: $P_0 \leftarrow O^{\mathrm{D}} \cup O^{\mathrm{AUG}}$; $L_{\mathrm{only}} \leftarrow O^{\mathrm{L}} \setminus P_0$
4: Compute $g(o; I, u)$ for $o \in L_{\mathrm{only}}$ using Equation 7
5: $P \leftarrow P_0 \cup \mathrm{Top}^{g}_{K_P}\{o \in L_{\mathrm{only}} : g(o) \geq \tau_P\}$

STEP 3 · REALIZE
6: $c_{\mathrm{raw}} \leftarrow \mathrm{VRGDecode}(I, I_0, P)$ using iterative real/null guidance
7: $c_{\mathrm{scaf}} \leftarrow \mathrm{PlanFilter}(c_{\mathrm{raw}}, P)$

STEP 4 · RECOVER
8: $M \leftarrow P \setminus O(c_{\mathrm{scaf}})$; **if** $M = \varnothing$ **then return** c_{scaf}
9: Score omissions and set $\tau_I \leftarrow Q_{\rho_I}(\{g(o) : o \in M,\ g(o) \text{ available}\})$
10: Form $\widehat{R}, \widehat{L}$ using agreement, image-level, and global gain gates
11: $H_R \leftarrow \mathrm{First}_{K_R}(\widehat{R})$; $H_L \leftarrow \mathrm{First}_{\min(K_L, K_R - |H_R|)}(\widehat{L})$
12: **return** c_{scaf} with fixed-template additions for H_R then H_L

followed by visually gated L-only omissions. Image-level and global gain gates control eligibility, while a total cap and an L-only sub-cap bound the added content. Selected objects are appended with fixed templates. Recovery restores verified coverage while its gates and caps control the added risk predicted by Equation 6. Implementation details appear in Appendices B and I.

Together, the four stages transform a one-shot autoregressive description into a revisable, evidence-gated commitment process. Cross-view planning filters noisy proposals, VRG grounds their realization, and bounded recovery restores verified omissions, preserving broad object coverage while reducing hallucination in the final caption.

5 EXPERIMENTS

We evaluate whether COVER retains high canonical-object coverage while reducing object hallucination, then audit which stages produce the observed operating point. All runs use a single NVIDIA RTX A6000. Full numerical tables, historical results, and diagnostic protocols appear in the appendices.

5.1 EXPERIMENTAL SETUP

Datasets. We evaluate on MSCOCO (Lin et al., 2014; Rohrbach et al., 2018), AMBER (Wang et al., 2023), Object HalBench (Yu et al., 2024), and OpenCHAIR (Ben-Kish et al., 2024). Following AGLA (An et al., 2025), each MSCOCO run uses 500 randomly sampled val2014 images; we report the mean across runs. AMBER uses all 1,004 generative images, Object HalBench uses 300 images, and OpenCHAIR uses its released evaluation subset.

Models and methods. We instantiate the autoregressive proposer with LLaVA-1.5-7B and the non-autoregressive masked-diffusion verifier/realizer with LLaDA-V (Liu et al., 2024b; You et al., 2025). Comparisons include the training-free inference methods VCD, DoLa, OPERA, AGLA, and AdaIAT, the preference-trained OPA-DPO, and the learned post-hoc revisor LURE.

Evaluation metrics. For MSCOCO and Object HalBench, sentence-level CHAIR (CHAIR_S, C_S) is the fraction of captions containing at least one unsupported object, while instance-level CHAIR (CHAIR_I, C_I) is the fraction of generated object mentions that are unsupported. Recall is the fraction of annotated image–object pairs recovered by the caption. AMBER reports its generative CHAIR and Cover as percentages: CHAIR measures the fraction of distinct extracted response objects outside the AMBER annotations, while Cover measures the fraction of annotated objects recovered. OpenCHAIR reports the resolved open-vocabulary hallucination rate (OCH) and supported object mentions per image; the latter is a density measure rather than exhaustive Recall.

Table 2: Object hallucination and coverage across four benchmarks.

	MSCOCO			AMBER (%)		Object HalBench		OpenCHAIR	
Method	$C_S \downarrow$	$C_I \downarrow$	Recall↑	CHAIR↓	Cover↑	$C_S \downarrow$	Recall↑	OCH↓	Sup./img↑
LLaVA-1.5	0.597	0.170	0.797	10.3	52.4	0.603	**0.798**	0.141	**7.800**
VCD	0.605	0.172	0.788	10.8	52.1	0.587	0.773	0.142	7.442
DoLa	0.617	0.180	**0.799**	11.5	52.2	0.583	0.770	0.127	7.492
OPERA	0.531	0.152	0.779	9.2	50.3	0.573	0.774	0.130	7.212
AGLA	0.591	0.171	0.782	10.0	52.3	0.593	0.777	0.141	7.436
OPA-DPO	0.127	0.040	0.575	**2.5**	47.7	0.140	0.564	**0.090**	5.780
AdaIAT	0.395	0.119	0.751	6.6	48.4	0.430	0.740	0.111	7.268
LURE	0.332	0.101	0.650	7.9	47.8	0.360	0.666	0.237	5.756
LLaDA-V	**0.103**	**0.047**	0.473	4.5	36.4	**0.100**	0.430	0.099	4.392
CaptionAug	0.202	0.066	0.654	5.6	54.3	0.203	0.588	0.115	6.038
SourcePlan	0.391	0.121	0.751	6.8	55.0	0.367	0.721	0.127	7.046
COVER	**0.429**	**0.129**	**0.786**	**6.8**	**59.3**	**0.410**	**0.769**	**0.131**	**7.042**

5.2 MAIN RESULTS

On MSCOCO, COVER establishes the strongest high-recall operating point in Table 2: it retains 98.7% of the detailed proposer's object Recall while reducing C_S by 28.1% and C_I by 23.8%. COVER achieves the two lowest hallucination rates among methods whose Recall is greater than 0.78; methods with still lower hallucination sacrifice substantially more grounded coverage.

The same high-coverage advantage transfers to the other annotation-based benchmarks. OPA-DPO obtains the lowest AMBER CHAIR (2.5), confirming that preference training can strongly suppress hallucinated objects, but its Cover falls to 47.7. In contrast, COVER achieves the highest Cover of all compared methods (59.3), improving over LLaVA-1.5 by 6.9 points (13.2%) while still reducing CHAIR by 34.0%. On Object HalBench, it retains 96.5% of the proposer's Recall while lowering C_S by 32.0%; among methods above 0.76 Recall, it is the least hallucinatory.

On OpenCHAIR, COVER lowers the proposer's OCH by 7.6% while retaining 90.3% of its supported mentions per image. Preference-trained OPA-DPO attains the lowest OCH, but the training-free COVER reports 21.8% more supported mentions per image. DoLa and AdaIAT remain stronger high-density training-free operating points; COVER instead reaches higher AMBER Cover and higher Recall on both MSCOCO and Object HalBench without optimizing additional model parameters.

Across all four benchmarks, LLaDA-V consistently occupies a low-hallucination but low-coverage endpoint, whereas LLaVA-1.5 supplies substantially broader coverage at a higher hallucination rate. This recurring complementarity motivates our cross-architecture design, which assigns hallucination control to non-autoregressive verification and coverage preservation to autoregressive proposals.

5.3 ABLATION STUDY

The ablation results explain how COVER converts broad proposals into selective final commitments. Figure 3 and Table 3 isolate the three decisions identified by our formulation: admission controls which candidates enter the object plan, masked diffusion realizes that plan under visual guidance, and bounded recovery restores supported objects omitted from the initial scaffold.

The proposal-admission study shows how agreement and visual gain play complementary roles. For proposer-only candidates, the 0.25 gain threshold raises precision from 0.3583 to 0.4788 while retaining 79.3% of supported candidates, and supported candidates are 0.3108 more likely to pass than unsupported candidates (95% CI $[0.2769, 0.3442]$). Cross-view agreement provides an additional reliability signal: candidate precision rises from 0.3544 for one proposing view to 0.8730 for two views and 0.9874 for all three. Agreement therefore identifies high-confidence shared proposals, while visual gain selectively recovers grounded objects contributed only by the broad proposer. Exact source signatures and full admission statistics appear in Appendices D and E.

Under an identical object plan, combined logit/confidence VRG reduces unsupported realization by 0.0628 while changing supported realization by only 0.0121 (Figure 3b). This asymmetric effect

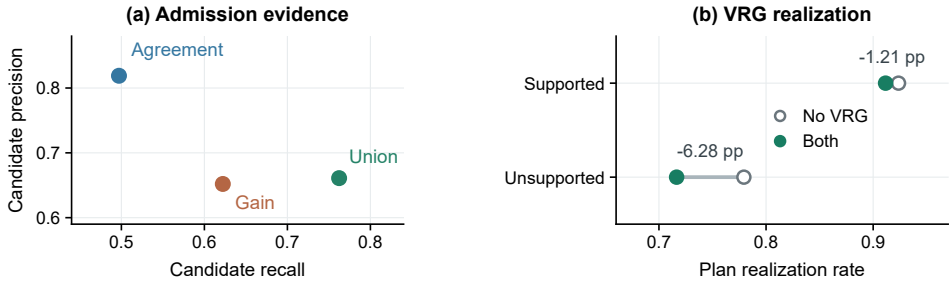

Figure 3: Stagewise evidence for COVER. (a) Cross-view agreement and visual gain provide complementary precision–coverage behavior during proposal admission. (b) Under a fixed object plan, VRG suppresses unsupported realization more strongly than supported realization.

Table 3: Ablation study.

(a) Recovery

Configuration	$C_S \downarrow$	$C_I \downarrow$	Recall↑
Plan + VRG scaffold	0.3913	0.1213	0.7508
Consensus-supported only	0.3977	0.1215	0.7728
All plan candidates, top 4	0.4593	0.1380	0.7943
LLaVA recovery, no global gate	0.4563	0.1367	0.7922
COVER	0.4293	0.1294	0.7863

(b) Proposal admission

Stage	Candidates	Supported	Precision↑	Coverage↑
Before threshold	2,986	1,070	0.3583	1.0000
Gain ≥ 0.25	1,771	848	0.4788	0.7925
Gain ≥ 0.25, top 3	1,690	810	0.4793	0.7570

shows that visual-residual guidance is not simply shortening the caption; it suppresses weak visual commitments while preserving most grounded plan content.

Recovery then determines the final coverage–risk operating point. Consensus-only recovery restores reliable omissions conservatively, while admitting every plan candidate or removing the global gate raises both Recall and hallucination. The reported configuration balances these directions at 0.7863 Recall and 0.4293 C_S, turning the low-hallucination scaffold into a high-coverage caption without reopening unrestricted generation. Detailed recovery-policy comparisons appear in Appendix J.2.

Taken together, these results align admission, realization, and recovery with Equation 2 and support a modular view: autoregressive generation explores broad candidates, masked diffusion provides revisable visual grounding, and the object plan connects them.

6 CONCLUSION

Object hallucination exposes a structural tension between broad coverage and reliable assertion. Our conservative controls show that low hallucination alone can result from omitting visible content, while uncertainty and fixation analyses reveal a key difference in how the two generation paradigms resolve uncertain objects: autoregressive decoding commits each emitted token irreversibly, whereas masked diffusion can revise predictions before fixation. CaptionAug increases supported commitment more strongly than unsupported commitment, motivating *selective object commitment* and its decomposition into proposal admission, diffusion realization, and omission recovery. Guided by this formulation, COVER combines broad autoregressive proposals with cross-view evidence, visual-residual-guided diffusion realization, and bounded recovery to control which objects enter the final caption. Across four benchmarks, COVER reduces hallucination while preserving coverage, including 28.1% lower sentence-level CHAIR at 98.7% proposer Recall on MSCOCO. These results establish proposal–commitment separation as a practical cross-architecture principle for grounded generation.

Ethics Statement

This work studies hallucination mitigation for image description, which can improve reliability in assistive and information-access applications. The method can still omit objects, misidentify people or sensitive attributes, and inherit biases from both foundation models. Its outputs should therefore be treated as assistive descriptions and verified when factual image content matters. Experiments use established public research datasets and do not introduce new human-subject data.

Reproducibility Statement

The appendices provide the expanded algorithm, final hyperparameters and prompts, per-pool results, proposal-source attribution, stage-specific diagnostics and ablations, the LURE protocol, formal statements, and proofs. Upon publication, we will release the source code, configuration files, and evaluation scripts needed to reproduce the main experiments.

AI Use Statement

In this work, generative AI tools were used to polish the language. All AI-assisted material was manually reviewed, verified, and revised by the authors. GPT-5.4 was used as an evaluator only for the explicitly labeled OpenCHAIR subset and a support-label sensitivity audit. We take responsibility for the final content of this work, including text with the aid of generative AI.

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 38th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926, 2025.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 17981–17993, 2021.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 22680–22698, 2024.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. ICT: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the 38th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4209–4221, 2025.

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 *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

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 (EMNLP)*, pp. 2728–2742, 2024.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. Mask-predict: Parallel decoding of conditional masked language models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6112–6121, 2019.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in LVLMs. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, 2025.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017.

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 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024.

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 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024.

Shufan Li, Konstantinos Kallidromitis, Hritik Bansal, Akash Gokul, Yusuke Kato, Kazuki Kozuka, Jason Kuen, Zhe Lin, Kai-Wei Chang, and Aditya Grover. LaViDa: A large diffusion language model for multimodal understanding. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. Diffusion-LM improves controllable text generation. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 4328–4343, 2022.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, volume 8693, pp. 740–755, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024b.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Proceedings of the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 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 (EMNLP)*, pp. 4035–4045, 2018.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander M. Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

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*, 2023.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate LVLM object hallucination. In *Proceedings of the 39th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25315–25325, 2026.

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 38th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 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(12):220105, 2024.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large language diffusion models with visual instruction tuning. *arXiv preprint*

arXiv:2505.16933, 2025.

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 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13807–13816, 2024.

Li'an Zhong, Ziqiang He, Jibin Zheng, Jin Li, Z. Jane Wang, and Xiangui Kang. AdaIAT: Adaptively increasing attention to generated text to alleviate hallucinations in LVLM. In *Proceedings of the 39th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*, 2024.

A Detailed Cover Algorithm

Algorithm 2 expands the main-text workflow into the individual scoring and decoding operations. Top-k selection returns at most k objects; plan-admission gain ties use source order. Ground-truth labels are loaded only by audit and evaluation scripts, never by generation, scoring, ranking, or recovery.

Algorithm 2 Expanded Cover inference

Input: Image I, zero image I_0, frozen models $f_{\mathrm{L}}, f_{\mathrm{D}}$;
decoding settings, gates τ_P, τ_G, image quantile ρ_I, caps K_P, K_R, K_L
Output: Final caption c

$c^{\mathrm{L}} \leftarrow f_{\mathrm{L}}(I; p_{\mathrm{detail}})$; $c^{\mathrm{D}} \leftarrow f_{\mathrm{D}}(I; p_{\mathrm{cap}})$
$d \leftarrow f_{\mathrm{D}}(I; p_{\mathrm{desc}})$; $c^{\mathrm{AUG}} \leftarrow f_{\mathrm{D}}(I, d; p_{\mathrm{cap}})$
Extract ordered canonical objects $O^{\mathrm{L}}, O^{\mathrm{D}}, O^{\mathrm{AUG}}$ and signatures $s(o)$
$\mathcal{U} \leftarrow O^{\mathrm{L}} \cup O^{\mathrm{D}} \cup O^{\mathrm{AUG}}$; $P_0 \leftarrow O^{\mathrm{D}} \cup O^{\mathrm{AUG}}$
$L_{\mathrm{only}} \leftarrow O^{\mathrm{L}} \setminus P_0$; score $g(o)$ on L_{only} using Equation 21
$P \leftarrow P_0 \cup \mathrm{Top}^{g}_{K_P}\{o \in L_{\mathrm{only}} : g(o) \geq \tau_P\}$
Initialize a masked response state x conditioned on P
for each denoising step t **do**
    Compute z_t^I and z_t^0 with f_{D} at the same x, prompt, and plan
    Obtain r_t, b_t, and $\widetilde{z}_t$ using Equations 22–24
    Predict clean tokens $\widehat{x}_0$ from $\widetilde{z}_t$; compute $\widetilde{q}_t$ using Equation 25
    Update x by fixing scheduled unresolved positions ranked by $\widetilde{q}_t$
end for
$c_{\mathrm{scaf}} \leftarrow$ decoded caption after removing sentences with out-of-plan objects
$M \leftarrow P \setminus O(c_{\mathrm{scaf}})$ ▷ Only planned omissions can be recovered
if $M = \varnothing$ **then return** c_{scaf}
Score planned omissions; $M_g \leftarrow \{o \in M : g(o) \text{ is available}\}$
$\tau_I \leftarrow Q_{\rho_I}(\{g(o) : o \in M_g\})$
$\widehat{R} \leftarrow \{o \in M : \sum_j s_j(o) \geq 2,\ g(o) > 0,\ g(o) \geq \tau_I\}$
$\widehat{L} \leftarrow \{o \in M \cap L_{\mathrm{only}} : g(o) > 0,\ g(o) \geq \tau_G\}$
$H_R \leftarrow \mathrm{First}_{K_R}(\widehat{R})$ in plan order
$H_L \leftarrow \mathrm{First}_{\min(K_L,\, K_R - |H_R|)}(\widehat{L})$ in plan order
$c \leftarrow c_{\mathrm{scaf}}$ followed by one fixed-template sentence per object in H_R then H_L
return c

Conventions. First_k takes at most k objects in their deterministic object-plan order. The reported configuration uses $\rho_I = 0.5$ and supplies $\tau_G = Q^{\mathrm{global}}_{0.5}(L)$, the label-free global median of proposer-only gain scores. Numerical settings are listed in Appendix B.

Table 4: Final MSCOCO configuration.

Component	Setting
LLaVA proposal	LLaVA-1.5-7B, greedy, 128-token budget
Direct diffusion	LLaDA-V, 128 steps
Object-plan realization	96-token block, 128 steps
Plan LLaVA-only threshold/cap	0.25 / 3 objects
VRG mode	both
VRG weights	$\alpha = 0.15$, $\beta = 0.01$
VRG residual	margin 0, clip 5, positive-only top-128
Consensus candidate gate	positive gain + within-image median
LLaVA-only candidate gate	positive gain + label-free global median
Recovery caps	4 total, at most 2 LLaVA-only
Realization	deterministic template append

B Hyperparameters and Prompts

The object-plan prompt is: "Object plan: {objects}. Write one concise factual caption for the image. Mention the objects in the plan when they are clearly visible. Do not mention any object outside the plan. Do not list the objects; write a natural image caption." The masked visual-gain instruction is: "Complete the assistant sentence using only an object that is clearly visible in the image." The final deterministic addition is: "The image also shows {object phrase}."

C Per-Pool Matched Results

Table 5: Per-pool CHAIR results for the matched LLaVA-1.5, LLaDA-V, and COVER comparison across six MSCOCO samples.

Seed	Method	CHAIR_s	CHAIR_i	Recall	Avg. objects
42	LLaVA-1.5	0.610	0.1728	0.8036	8.378
	LLaDA-V	0.094	0.0488	0.4569	3.032
	COVER	0.408	0.1265	0.7882	5.548
86	LLaVA-1.5	0.610	0.1723	0.7909	8.416
	LLaDA-V	0.104	0.0442	0.4742	3.210
	COVER	0.448	0.1303	0.7822	5.724
123	LLaVA-1.5	0.584	0.1575	0.8101	7.988
	LLaDA-V	0.108	0.0424	0.4862	3.022
	COVER	0.414	0.1317	0.8034	5.466
314	LLaVA-1.5	0.620	0.1808	0.7963	8.354
	LLaDA-V	0.110	0.0527	0.4786	3.224
	COVER	0.448	0.1350	0.7812	5.676
456	LLaVA-1.5	0.558	0.1575	0.7940	8.332
	LLaDA-V	0.100	0.0423	0.4660	3.118
	COVER	0.426	0.1225	0.7793	5.504
7003	LLaVA-1.5	0.602	0.1779	0.7873	8.298
	LLaDA-V	0.100	0.0492	0.4769	3.130
	COVER	0.432	0.1305	0.7834	5.638

D Proposal-Source Attribution

For each image and canonical object, we record whether the candidate appears in the LLaVA-1.5 caption, the direct LLaDA-V caption, and the CaptionAug caption. The seven non-empty exact signatures in Table 6 are disjoint. Precision measures the fraction of candidates present in the annotated COCO object set, while GT coverage measures the fraction of annotated image objects contributed by that signature.

Table 6: Exact proposal-source signatures over 11,667 candidate records from the six matched MSCOCO pools. Intervals use 5,000 image-clustered bootstrap resamples.

Exact signature	N	Precision [95% CI]	GT cov. [95% CI]	Unsup./img
LLaVA-1.5 only	4,293	0.3445 $[0.3296, 0.3593]$	0.1559 $[0.1486, 0.1631]$	0.9380
Direct only	211	0.5024 $[0.4378, 0.5700]$	0.0112 $[0.0091, 0.0134]$	0.0350
CaptionAug only	586	0.3737 $[0.3328, 0.4147]$	0.0231 $[0.0200, 0.0264]$	0.1223
LLaVA-1.5 + direct	377	0.8143 $[0.7746, 0.8535]$	0.0324 $[0.0289, 0.0360]$	0.0233
LLaVA + CaptionAug	2,195	0.8966 $[0.8837, 0.9088]$	0.2074 $[0.1996, 0.2152]$	0.0757
Direct + CaptionAug	207	0.7295 $[0.6712, 0.7865]$	0.0159 $[0.0135, 0.0185]$	0.0187
All three	3,798	0.9874 $[0.9836, 0.9907]$	0.3952 $[0.3845, 0.4059]$	0.0160

The source marginals are $(\text{precision}, \text{coverage}) = (0.7037, 0.7908)$ for LLaVA-1.5, $(0.9393, 0.4546)$ for direct LLaDA-V, and $(0.8971, 0.6416)$ for CaptionAug. The precision gradient across signatures supports agreement as a reliability signal, while the source marginals show how the autoregressive and diffusion views contribute complementary coverage.

E Proposal-Stage Visual-Gain Audit

This analysis tests visual gain before object-plan construction. On each evaluation image, the candidate population contains every canonical object in $O^{\mathrm{L}} \setminus (O^{\mathrm{D}} \cup O^{\mathrm{AUG}})$ before thresholding or the top-3 cap, with support defined by membership in the official CHAIR ground-truth set.

The realized admission rule passes gain values at or above 0.25, ranks by descending gain, then by proposer source order and canonical object, and admits at most three candidates. All 2,986 candidates have saved gain values. Replaying this rule reproduces every saved proposer-only plan admission with zero mismatches.

Table 7: Detailed proposer-only admission flow. Threshold removals are measured from the full candidate population; cap-only removals are measured after thresholding.

Stage	Total	Supported	Unsupported	Precision	Supp. coverage
Before threshold	2,986	1,070	1,916	0.35834	1.00000
After threshold	1,771	848	923	0.47883	0.79252
After top-3 cap	1,690	810	880	0.47929	0.75701
Removed by threshold	1,215	222	993	–	–
Removed only by cap	81	38	43	–	–

The first co-primary endpoint tests threshold discrimination directly. If $A(o)$ denotes passage of $g(o) \geq 0.25$, the supported-minus-unsupported pass-rate difference is

$$\Delta_A = P(A = 1 \mid Z = 1) - P(A = 1 \mid Z = 0) = 0.31079, \tag{9}$$

with image-bootstrap 95% CI $[0.27687, 0.34417]$. Thus the threshold retains 79.25% of supported candidates and 48.17% of unsupported candidates, matching the likelihood-ratio direction required by Equation 5.

The second endpoint controls the number of admitted candidates within each image. For image j with n_j proposer-only candidates, G_j supported candidates, and k_j actual admissions, uniform sampling without replacement has expected supported count $k_j G_j / n_j$. The actual excess is

$$D_j^{\mathrm{plan}} = \sum_{i \in L_j} A_{ji} Z_{ji} - \frac{k_j G_j}{n_j}, \qquad \overline{D}^{\mathrm{plan}} = \frac{1}{N} \sum_{j=1}^{N} D_j^{\mathrm{plan}}. \tag{10}$$

The selected plans contain 810 supported candidates, compared with an exact expectation of 725.23, giving $\overline{D}^{\mathrm{plan}} = 0.04238$ supported objects/image with 95% CI $[0.03299, 0.05148]$. There are 484 images with nondegenerate choice, $0 < k_j < n_j$. Both intervals use 5,000 image resamples with seed 20260914; candidates are clustered with their image and are not treated as independent samples.

As secondary discrimination summaries, gain has AUROC 0.7036 and average precision 0.5354, compared with supported prevalence 0.3583. Supported candidates have mean/median gain $1.5668/1.4197$, while unsupported candidates have $0.2911/0.1798$. These values support visual gain as a proposer-only admission signal; the cap bounds the number of additional candidates admitted per image.

Candidate- and image-level resampling reproduces both reported point estimates and intervals.

F LLaVA-1.5 Prompt and Decoding Sweep

Table 8: LLaVA-1.5 prompt and decoding sweep on the common MSCOCO sample.

LLaVA-1.5 setting	CHAIR_s	CHAIR_i	Recall	Avg. objects
Detailed, 64 tokens, greedy	0.242	0.0828	0.6126	4.422
Detailed, 96 tokens, greedy	0.528	0.1514	0.7290	6.660
Detailed, 128 tokens, greedy	0.642	0.1860	0.7816	8.138
Conservative, 96 tokens, greedy	0.226	0.2257	0.6320	3.332
Conservative, 128 tokens, greedy	0.226	0.2461	0.6320	3.682
Strict caption, 96 tokens, greedy	0.080	0.0579	0.5163	1.936
Object list, 96 tokens, greedy	0.134	0.1228	0.3548	2.182
Main objects, 96 tokens, greedy	0.146	0.0724	0.5826	2.292
Conservative, 96 tokens, beam 3	0.148	0.1314	0.5857	2.374
Object list, 96 tokens, beam 3	0.180	0.1136	0.4393	4.630

The sweep shows how an autoregressive model can lower hallucination by shortening or constraining its output. As the prompt requests broader content, Recall rises together with both CHAIR measures, motivating a coverage-matched comparison rather than a lowest-hallucination comparison.

G Commitment and Trajectory Diagnostics

Figure 1 in Section 3 summarizes the object-level and trajectory evidence. This appendix specifies the metric units and the scope of those diagnostics.

Metric units. Let $G(I)$ be the annotated canonical object set, $O_{\mathrm{set}}(c)$ the set of extracted canonical objects, and $O_{\mathrm{mult}}(c)$ the mention multiset. For images with nonempty $G(I)$,

$$\mathrm{Recall}(c, I) = \frac{|O_{\mathrm{set}}(c) \cap G(I)|}{|G(I)|}, \tag{11}$$

$$\mathrm{CHAIR}_i = \frac{\sum_{(I,c)} \sum_{o \in O_{\mathrm{mult}}(c)} \mathbf{1}[o \notin G(I)]}{\sum_{(I,c)} |O_{\mathrm{mult}}(c)|}, \tag{12}$$

$$\mathrm{CHAIR}_s = \frac{1}{N} \sum_{(I,c)} \mathbf{1}[O_{\mathrm{set}}(c) \setminus G(I) \neq \varnothing]. \tag{13}$$

The evaluator resolves the CHAIR vocabulary and aliases before counting. Repeated mentions affect CHAIR_i but not set recall. The fixed candidate-union diagnostic excludes visible objects missed by every proposer, and consequently its supported commitment rate is not benchmark Recall.

Autoregressive commitment. The probability chain rule writes sequence likelihood as

$$p_\theta(y_{1:T} \mid I) = \prod_{t=1}^{T} p_\theta(y_t \mid y_{<t}, I). \tag{14}$$

Left-to-right inference makes emitted tokens fixed prefix context (Vaswani et al., 2017), so later distributions are conditioned on earlier object decisions. This irreversible ordering motivates comparing commitment uncertainty with the revisable fixation process of masked diffusion.

Uncertainty and fixation. For a token distribution $p_{i,m}^{(s)}$ at position i and denoising step s, predictive entropy is

$$h_{i,m}^{(s)} = -\sum_{v \in \mathcal{V}} p_{i,m}^{(s)}(v) \log p_{i,m}^{(s)}(v). \tag{15}$$

We compare uncertainty only within each model and decoding schedule (Shannon, 1948; Guo et al., 2017; Farquhar et al., 2024). Unsupported LLaVA-1.5 emissions have mean chosen-token probability 0.604 versus 0.710 for supported emissions and entropy 1.536 versus 1.087. Unsupported direct LLaDA-V objects fix at 41.87 versus 26.47 steps, change 6.64 versus 4.15 times while masked, and have final-token probability 0.431 versus 0.595. The two generators therefore expose complementary uncertainty signatures: autoregressive errors are committed with higher uncertainty, while masked-diffusion errors remain unresolved longer and undergo more revision.

H Caption Augmentation and Selective Expansion Details

Caption Augmentation uses the same frozen LLaDA-V checkpoint for two passes. The first pass produces an auxiliary description, and the second conditions caption generation on that description:

$$\begin{aligned} d &= f_\theta(I; p_{\text{desc}}), \\ c^{\text{D}} &\sim p_\theta(c \mid I; p_{\text{cap}}), \\ c^{\text{AUG}} &\sim p_\theta(c \mid I, d; p_{\text{cap}}). \end{aligned} \tag{16}$$

The description supplies semantic context that can expose objects omitted by direct captioning while preserving masked-diffusion revision. Together with the detailed autoregressive view, the two diffusion views form the proposal union

$$\mathcal{U}(I) = O^{\text{L}} \cup O^{\text{D}} \cup O^{\text{AUG}}. \tag{17}$$

The complete LLaVA-1.5 prompt and decoding sweep appears in Appendix F, and the corresponding uncertainty and fixation definitions appear in Appendix G.

Object-level selectivity. On the shared diagnostic sample, $1{,}974$ unique objects are proposed by at least one view. Let c_m^+ and c_m^- be mode m's commitment rates among supported and unsupported candidates in this fixed union. LLaVA-1.5 commits 92.5% and 82.9%, respectively, compared with direct LLaDA-V's 71.5% and 17.6%. Their selectivity ratios c_m^+/c_m^- are 1.12 and 4.07, while CaptionAug reaches 4.87. Its mode-by-support interaction is

$$\mathcal{I}_{\text{AUG}} = (c_{\text{AUG}}^+ - c_{\text{D}}^+) - (c_{\text{AUG}}^- - c_{\text{D}}^-) = 0.065, \tag{18}$$

with image-clustered 95% CI $[0.023, 0.108]$. Supported commitment rises more strongly than unsupported commitment, providing the object-level basis for selective proposal expansion and the stagewise formulation in Section 3.

I Detailed Scoring and Decoding Specification

I.1 Overview

COVER separates coverage, verification, and realization (Figure 2). It first collects a LLaVA-1.5 caption and two LLaDA-V generation views. Their proposal union $\mathcal{U}(I)$ supplies an object plan $P(I) \subseteq \mathcal{U}(I)$, corresponding to Q in Theorem 2. Cross-view signatures and masked visual gain guide plan construction and recovery, LLaDA-V realizes the scaffold (S) with VRG, and capped recovery revisits planned omissions (H).

Algorithm 1 summarizes the complete inference flow. Its two selection stages are distinct: planning determines which objects may be realized, while recovery considers only objects that remain absent from the scaffold.

I.2 Cross-View Object Proposals

For each image, we produce three captions:

$$c^{\mathrm{L}}, \qquad c^{\mathrm{D}}, \qquad c^{\mathrm{AUG}}. \tag{19}$$

Here L denotes LLaVA-1.5: c^{L} is its detailed caption, c^{D} is the direct LLaDA-V caption, and c^{AUG} is produced by the caption-augmentation operation introduced in Section 3. It gives LLaDA-V a second, more recall-oriented view without adding another model.

We extract ordered object sets O^{L}, O^{D}, and O^{AUG}. Each object o receives a three-bit support signature

$$s(o) = \left(\mathbf{1}[o \in O^{\mathrm{L}}], \mathbf{1}[o \in O^{\mathrm{D}}], \mathbf{1}[o \in O^{\mathrm{AUG}}]\right). \tag{20}$$

The signature induces agreement levels: all-three support, diffusion-view consensus, cross-model support, a single diffusion view, and LLaVA-only support. These are inference-time agreement categories, not dataset-level reliability estimates.

I.3 Masked Visual Gain

Agreement alone cannot distinguish shared visual evidence from shared language priors. We therefore score object tokens directly with LLaDA-V. Given a short grounding context u, an object phrase $o = (o_1, \ldots, o_m)$, real image I, and zero-image tensor I_0, let $\widetilde{o}$ denote the fixed input with the object span masked. At the corresponding token positions, we compute

$$g(o; I, u) = \frac{1}{m} \sum_{j=1}^{m} \left[\log p_{\theta,j}(o_j \mid u, \widetilde{o}, I) - \log p_{\theta,j}(o_j \mid u, \widetilde{o}, I_0)\right]. \tag{21}$$

The same prompt and token positions are used in both branches. Positive g means that the actual image increases the masked object's conditional probability relative to the counterfactual visual input. Length averaging places phrases on a per-token scale, producing a comparable image-dependence score for object admission.

I.4 Evidence-Constrained Object Planning

The initial plan includes all objects named by either diffusion view. LLaVA-only objects are sorted by visual gain, and at most three are added when $g(o) \geq 0.25$. The resulting plan P conditions a LLaDA-V prompt that requests one concise caption, forbids objects outside P, and drops generated sentences that violate the plan. This converts the LLaVA-1.5 caption from a textual scaffold into a proposal inventory: LLaDA-V controls the final language realization.

I.5 Visual Residual Guided Diffusion Decoding

VRG adjusts both token prediction and fixation priority within the existing LLaDA-V decoder. At each denoising step t, two forward passes share the current masked text x_t, object plan, prompt, and token positions, while differing only in the visual input I or I_0. Let $z_t^I, z_t^0 \in \mathbb{R}^{L \times |V|}$ be their token logits. We define a clipped residual and positive boost

$$r_t = \mathrm{clip}(z_t^I - z_t^0, -c, c), \tag{22}$$

$$b_t = \mathrm{TopK}\left([r_t - m]_+, k\right), \tag{23}$$

$$\widetilde{z}_t = z_t^I + \alpha b_t, \tag{24}$$

where m is a margin, c is a residual clip, and k restricts the boost to the strongest vocabulary entries at each position. The residual augments z_t^I, retaining the real-image scores as the baseline. Predicted clean tokens $\widehat{x}_0$ are sampled or selected from $\widetilde{z}_t$. Unlike the object-level gain in Section 4.2, this contrast acts on the evolving sequence state and its full vocabulary.

LLaDA-V must also choose which unresolved positions to fix under its decoding schedule. VRG augments the guided probability of the selected token with a positive residual bonus:

$$\widetilde{q}_{t,i} = \min\Big(1, p_{\widetilde{z}_t}(\widehat{x}_{0,i}) + \beta\,[r_{t,i}(\widehat{x}_{0,i}) - m]_+\Big). \tag{25}$$

This quantity ranks fixation priority. The logits branch uses the top-k boost b_t, whereas the confidence bonus reads the selected token's clipped residual r_t directly. Both operations reuse model outputs without introducing a learned confidence head.

We evaluate logits-only ($\alpha > 0, \beta = 0$), confidence-only ($\alpha = 0, \beta > 0$), and both. The final scaffold uses $\alpha = 0.15$, $\beta = 0.01$, $m = 0$, $c = 5$, and $k = 128$. The zero-image branch is evaluated at every denoising step, so the residual changes with the partially generated caption. Proposition 4 characterizes this one-step reweighting, and the fixed-plan factorial measures its selective realization effect.

I.6 Cross-View Verification and Evidence-Guided Recovery

Object planning can expose supported objects that remain absent from the realized scaffold. Recovery revisits these planned omissions without reopening the unrestricted object space.

The final policy admits positive-gain consensus omissions at or above the within-image median, then uses globally gated proposer-only omissions for remaining slots. Eligible objects retain their deterministic object-plan order. Recovery uses at most four additions, of which two can be proposer-only, and appends each selected object with a fixed template.

For scored omissions $M_g = \{o \in P \setminus O(c_{\text{scaf}}) : g(o) \text{ is available}\}$, consensus omissions R_I, and proposer-only omissions L_I, the admissible sets are

$$\widehat{R}_I = \{o \in R_I : g(o) > 0,\ g(o) \geq Q_{0.5}(\{g(o') : o' \in M_g\})\}, \tag{26}$$

$$\widehat{L}_I = \{o \in L_I : g(o) > 0,\ g(o) \geq Q_{0.5}^{\text{global}}(L)\}. \tag{27}$$

The final list takes consensus candidates first, then globally gated proposer-only omissions under the remaining total and proposer-only caps. The proposer-only gate uses the label-free global median gain.

J Full Benchmark and Ablation Results

This appendix reports full-precision benchmark tables and the two stage-specific factorials summarized in Section 5.

J.1 Results

MSCOCO CHAIR. Table 9 gives the full-precision six-pool means, average object counts, and recovery operating points underlying the rounded main-text results.

Table 9: CHAIR aggregate on six matched MSCOCO samples.

Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall↑	AvgObj
LLaVA-1.5	0.5973	0.1698	0.7970	8.2943
VCD	0.6050	0.1724	0.7876	8.2750
DoLa	0.6170	0.1799	**0.7993**	8.2270
OPERA	0.5310	0.1523	0.7794	7.8943
AGLA	0.5910	0.1708	0.7824	8.1420
OPA-DPO	0.1273	**0.0397**	0.5746	4.5740
AdaIAT	0.3953	0.1190	0.7508	7.3483
LLaDA-V	**0.1027**	0.0466	0.4731	3.1227
Plan + VRG scaffold	0.3913	0.1213	0.7508	5.3867
Consensus-supported only	0.3977	0.1215	0.7728	5.4813
All plan candidates, top 4	0.4593	0.1380	0.7943	5.6803
LLaVA recovery, no global gate	0.4563	0.1367	0.7922	5.6640
Cover	0.4293	0.1294	**0.7863**	5.5927

AMBER. Table 10 adds the Hal and Cog components omitted from the compact main table and reports all reproduced baselines under the matched generative protocol.

Table 10: AMBER generative components on all 1,004 images. LURE uses the same detailed-prompt source captions as LLaVA-1.5.

Method	CHAIR↓	Cover↑	Hal↓	Cog↓
LLaVA-1.5	10.3	52.4	45.3	5.5
Beam Search	12.6	49.2	53.7	6.1
VCD	10.8	52.1	46.8	5.6
DoLa	11.5	52.2	49.8	6.2
OPERA	9.2	50.3	36.9	4.2
AGLA	10.0	52.3	43.7	5.8
OPA-DPO	**2.5**	47.7	**13.1**	1.1
AdaIAT	6.6	48.4	29.2	2.8
LURE	7.9	47.8	33.8	3.7
LLaDA-V	4.5	36.4	14.0	**0.5**
CaptionAug	5.6	54.3	26.3	2.0
SourcePlan	6.8	55.0	30.2	2.0
COVER	6.8	**59.3**	33.1	2.3

Object HalBench. Table 11 reports full metrics and the default, strict, and high-recall COVER variants on the 300-image evaluation.

Table 11: Object HalBench results on 300 images using CHAIR.

Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Recall↑	AvgObj
LLaVA-1.5	0.6033	0.1727	**0.7975**	7.7767
Beam Search	0.6300	0.1760	0.7880	8.0700
VCD	0.5867	0.1763	0.7733	7.6933
DoLa	0.5833	0.1824	0.7702	7.6567
OPERA	0.5733	0.1655	0.7744	7.5333
AGLA	0.5933	0.1747	0.7765	7.6700
OPA-DPO	0.1400	**0.0386**	0.5635	3.9700
AdaIAT	0.4300	0.1307	0.7398	7.0400
LURE	0.3600	0.1212	0.6663	5.7467
LLaDA-V	**0.1000**	0.0424	0.4302	2.9900
CaptionAug	0.2033	0.0699	0.5876	3.8600
SourcePlan	0.3667	0.1250	0.7209	5.0667
COVER (default)	0.4167	0.1347	0.7597	5.2967
COVER (strict)	0.4100	0.1372	0.7692	5.3467
COVER (high recall)	0.4467	0.1471	0.7775	5.4400

J.2 MECHANISM ANALYSIS

Commitment rates and token-trajectory diagnostics are reported in Appendices H and G. This section isolates the two mechanism interventions used by the complete pipeline.

Visual residual guidance. Table 12 holds object plans fixed and excludes recovery. Identical plan hashes across all cells isolate residual-logit and confidence guidance; their combination yields the lowest mean CHAIR values and the largest reduction in unsupported realization.

Table 12: VRG realization factorial across six matched MSCOCO pools. Entries are cross-pool means under identical object plans; GT/Hall Real. denote realization rates for supported/unsupported planned objects.

Mode	$\mathrm{CHAIR}_s \downarrow$	$\mathrm{CHAIR}_i \downarrow$	Recall↑	GT Real.↑	Hall Real.↓
No VRG	0.4103	0.1282	**0.7601**	0.9237	0.7793
Residual logits only	0.3960	0.1236	0.7533	0.9146	0.7455
Confidence only	0.4013	0.1260	0.7600	**0.9240**	0.7558
Both	**0.3913**	**0.1213**	0.7508	0.9116	**0.7165**

Candidate admission signals. Table 13 holds the candidate universe, source scaffold, recovery budget, and realization procedure fixed. Agreement supplies the highest candidate precision, visual gain broadens candidate recall, and their union reaches the largest caption and candidate Recall.

Table 13: Agreement and visual-gain admission factorial across six matched MSCOCO pools. Caption metrics are cross-pool means; Cand.P and Cand.R are the precision and recall of admitted candidates.

Admission evidence	$\mathrm{CHAIR}_s \downarrow$	$\mathrm{CHAIR}_i \downarrow$	Recall↑	Cand.P↑	Cand.R↑
Neither	0.3913	0.1213	0.7508	–	0.0000
Agreement only	0.4030	0.1225	0.7787	**0.8189**	0.4971
Visual gain only	0.4263	0.1291	0.7844	0.6520	0.6222
Agreement $\vee$ visual gain	0.4313	0.1302	**0.7927**	0.6610	**0.7623**

Proposal-source attribution. Appendix D reports the disjoint three-view source signatures and bootstrap intervals. Those exact signatures provide the full evidence behind the agreement categories summarized by the admission factorial.

Recovery policy. The component rows in Table 9 compare the complete policy with consensus-only recovery, unrestricted plan admission, and removal of the global proposer-only gate. The reported policy retains 0.7863 Recall at $(\mathrm{CHAIR}_s, \mathrm{CHAIR}_i) = (0.4293, 0.1294)$, while the alternatives trace the expected coverage–risk frontier. Together, the two factorials and recovery comparison identify a staged system: VRG filters weak realization, agreement and visual gain recover complementary evidence, and the final gates control how much expanded coverage is committed.

K LURE Correction Baseline

LURE is a learned revisor for completed descriptions (Zhou et al., 2024). We apply its released MiniGPT-4/Vicuna-13B correction checkpoint to the same LLaVA-1.5 source captions used by our comparisons. Visual teacher-forced replay supplies first-subtoken uncertainty, and the released masking and rewrite rules are retained. MSCOCO uses matched source captions, Object HalBench uses its 300-image LLaVA input, and AMBER uses all 1,004 images with the detailed prompt used for the main LLaVA row.

Table 14: LURE paired effects with 5,000 image-bootstrap resamples. Differences follow the stated left-minus-right order.

Contrast	$\Delta\mathrm{CHAIR}_s$ [95% CI]	$\Delta\mathrm{CHAIR}_i$ [95% CI]	ΔRecall [95% CI]
MSCOCO: LURE − source	-0.2600 $[-0.3040, -0.2140]$	-0.0692 $[-0.0846, -0.0541]$	-0.1370 $[-0.1564, -0.1170]$
MSCOCO: Cover − LURE	$+0.1000$ $[0.0500, 0.1500]$	$+0.0325$ $[0.0152, 0.0490]$	$+0.1275$ $[0.1031, 0.1516]$
HalBench: LURE − source	-0.2433 $[-0.3033, -0.1867]$	-0.0515 $[-0.0683, -0.0350]$	-0.1312 $[-0.1564, -0.1070]$

The paired results place LURE at a more conservative operating point, with lower hallucination and lower grounded-object Recall than its source. On AMBER, COVER improves over LURE on all four generative components; the full point estimates are reported in Table 10.

L FORMAL STATEMENTS FOR SELECTIVE COMMITMENT

We analyze three questions: when proposal expansion preserves selectivity, how conditional visual evidence changes proposal reliability, and how verified recovery changes caption metrics. The object of analysis is the final inclusion of a canonical object. The resulting factorization connects proposal, realization, and recovery to the hallucination–coverage trade-off, with proofs in Appendix M.

L.1 EXPANSION MUST BE MATCHED BY SELECTIVITY

Fix a distribution over image–object pairs before comparing policies, and let $Z \in \{0,1\}$ denote annotated support, with $\pi = P(Z=1) \in (0,1)$. Let Q_m indicate inclusion in policy m's object plan, S_m inclusion in its scaffold, and H_m recovery of a planned omission. Assume canonical-object containment: $S_m \le Q_m$ and $H_m \le Q_m(1-S_m)$. Final commitment is $C_m = S_m + H_m$. Define, for $z \in \{0,1\}$,

$$\begin{aligned} q_{m,z} &= P(Q_m = 1 \mid Z = z), \\ r_{m,z} &= P(S_m = 1 \mid Q_m = 1, Z = z), \\ h_{m,z} &= P(H_m = 1 \mid Q_m = 1, S_m = 0, Z = z), \\ v_{m,z} &= r_{m,z} + (1 - r_{m,z}) h_{m,z}. \end{aligned} \tag{28}$$

Rates on null conditioning events can be assigned arbitrarily when multiplied by zero. These are population rates, not independence assumptions or inference-time labels.

Theorem 2 (Stagewise selectivity)**.** *Under containment, supported commitment is $T_m = q_{m,1} v_{m,1}$. When $P(C_m = 1) > 0$, unsupported risk among committed pairs is*

$$R_m = \frac{(1-\pi) q_{m,0} v_{m,0}}{\pi q_{m,1} v_{m,1} + (1-\pi) q_{m,0} v_{m,0}}. \tag{29}$$

For two policies with positive class-specific commitment rates, $R_{m'} \le R_m$ and $T_{m'} \ge T_m$ hold if and only if

$$\frac{q_{m',1} v_{m',1}}{q_{m',0} v_{m',0}} \ge \frac{q_{m,1} v_{m,1}}{q_{m,0} v_{m,0}}, \qquad q_{m',1} v_{m',1} \ge q_{m,1} v_{m,1}. \tag{30}$$

No independence between views, candidates, or stages is required.

The original Selective Commitment Risk Decomposition is recovered by writing $c_m^+ = q_{m,1} v_{m,1}$, $c_m^- = q_{m,0} v_{m,0}$, and $\Lambda_m = c_m^+ / c_m^-$:

$$R_m = \left[1 + \frac{\pi}{1-\pi} \Lambda_m \right]^{-1}. \tag{31}$$

The stage factorization explains why more proposals alone cannot guarantee improvement. If expansion reduces $q_{m,1}/q_{m,0}$, verification and realization must increase $v_{m,1}/v_{m,0}$ by at least the reciprocal factor to preserve risk. Label-independent thinning leaves Λ_m unchanged while lowering T_m. CaptionAug, VRG, and recovery act on different factors, which motivates measuring their effects separately. T_m and R_m concern the fixed pair distribution; neither is automatically macro-averaged benchmark Recall or mention-weighted CHAIR_i.

L.2 CONDITIONAL VISUAL EVIDENCE

Condition on membership in a fixed proposal universe and let s denote a source-agreement tier. For an event $V_t = \{g \ge t\}$, Bayes' rule gives

$$\frac{P(Z=1 \mid s, V_t)}{P(Z=0 \mid s, V_t)} = \frac{P(Z=1 \mid s)}{P(Z=0 \mid s)} \frac{P(V_t \mid Z=1, s)}{P(V_t \mid Z=0, s)}. \tag{32}$$

When these probabilities are positive, the gate improves tier precision exactly when its last likelihood ratio exceeds one. This identity explains the complementary roles of agreement and visual gain: agreement defines a reliability tier, while gain discriminates candidates within that tier. Appendix E verifies the required pass-rate ordering for proposer-only candidates.

L.3 Recovery Has a Metric-Specific Cost

The final caption is not identical to its plan. Even a selective plan can lose supported objects during realization, and additive recovery can restore them while introducing errors. The next result connects these stages to the actual benchmark quantities without a calibration assumption.

Theorem 3 (Additive recovery identities)**.** *On a fixed N-image evaluation, suppose recovery preserves every scaffold mention and adds distinct canonical objects absent from that image's scaffold, each as one extracted mention. Let $M_0 > 0$ and F_0 be the total scaffold mention and unsupported-mention counts. Let k_j and f_j be the added and unsupported-added counts for image j, and $k = \sum_j k_j > 0$, $f = \sum_j f_j$. Then*

$$\mathrm{CHAIR}_i^{\mathrm{final}} - \mathrm{CHAIR}_i^{\mathrm{scaf}} = \frac{k}{M_0 + k}\left(\frac{f}{k} - \frac{F_0}{M_0}\right), \tag{33}$$

$$\mathrm{CHAIR}_s^{\mathrm{final}} - \mathrm{CHAIR}_s^{\mathrm{scaf}} = \frac{1}{N}\sum_j \mathbf{1}[F_{0,j} = 0]\mathbf{1}[f_j > 0] \geq 0, \tag{34}$$

$$\mathrm{Recall}_j^{\mathrm{final}} - \mathrm{Recall}_j^{\mathrm{scaf}} = \frac{k_j - f_j}{|G(I_j)|}, \qquad |G(I_j)| > 0. \tag{35}$$

Recovery lowers CHAIR_i when its added mentions are more precise than the scaffold's existing mentions, while supported additions raise Recall. It cannot lower CHAIR_s relative to that same scaffold, so recovery is designed as bounded coverage restoration within a full pipeline that has already reduced hallucination during planning and realization.

M Proofs for Selective Commitment

M.1 Stagewise Selectivity

All pair-level comparisons use one fixed reference law μ over image–canonical-object pairs. The law may represent micro-averaged object coverage or macro-averaged per-image coverage, provided the same law is used for every compared policy.

Proof of Theorem 2. Containment and disjointness imply $C = S + H$ and $C = 0$ on $Q = 0$. The conditional law of total probability gives, for each support class z,

$$P(C = 1 \mid Z = z) = q_z\{r_z + (1 - r_z)h_z\} = q_z v_z. \tag{36}$$

Bayes' rule yields Equation 29. For positive rates,

$$\frac{1 - R_m}{R_m} = \frac{\pi}{1 - \pi}\frac{q_{m,1}v_{m,1}}{q_{m,0}v_{m,0}}. \tag{37}$$

At fixed π, this odds ratio increases exactly when R_m decreases, while the second inequality in Equation 30 is $T_{m'} \geq T_m$, proving necessity and sufficiency. □

For positive $q_{m,z}$, the no-risk-increase condition can be written as

$$\frac{v_{m',1}}{v_{m',0}} \geq \frac{q_{m,1}/q_{m,0}}{q_{m',1}/q_{m',0}}\frac{v_{m,1}}{v_{m,0}}. \tag{38}$$

Equation 38 makes the design requirement explicit: any selectivity lost by broadening the plan must be recovered during realization and recovery. In contrast, label-independent thinning multiplies both class-specific commitment rates by the same factor, preserving risk while reducing coverage.

M.2 Conditional Evidence Identity

Within a source-agreement tier s, apply Bayes' rule to the visual-gain event $V_t = \{g \geq t\}$:

$$P(Z = z \mid s, V_t) = \frac{P(V_t \mid Z = z, s)P(Z = z \mid s)}{P(V_t \mid s)}. \tag{39}$$

Dividing the two support classes cancels the denominator and gives Equation 32. The accepted tier therefore has higher support odds exactly when $P(V_t \mid Z = 1, s) > P(V_t \mid Z = 0, s)$. This is the stage-specific criterion measured by the proposer-only admission analysis in Appendix E.

M.3 Recovery and Caption Metrics

Proof of Theorem 3. Append-only realization gives total final mention and false-mention counts $M_0 + k$ and $F_0 + f$. Subtracting the two mention fractions yields

$$\frac{F_0 + f}{M_0 + k} - \frac{F_0}{M_0} = \frac{fM_0 - F_0 k}{M_0(M_0 + k)} = \frac{k}{M_0 + k}\left(\frac{f}{k} - \frac{F_0}{M_0}\right). \tag{40}$$

An already hallucinated scaffold remains hallucinated, while a clean scaffold changes its sentence indicator exactly when it receives at least one unsupported addition. Summing these indicator changes proves Equation 34.

Each of the $k_j - f_j$ supported additions is a distinct canonical object absent from the scaffold, so the recalled set grows by exactly this count. Dividing by $|G(I_j)|$ proves Equation 35. □

These identities explain the role of the recovery gates. Added-object precision determines the change in CHAIR_i, clean-to-hallucinated transitions determine the change in CHAIR_s, and distinct supported additions determine the Recall gain. The recovery cap limits the volume of this final coverage expansion.

M.4 Visual Residual Guidance as a Logit Tilt

Proposition 4 (One-step residual tilt)**.** *At a fixed token position and masked state, let* $p = \mathrm{softmax}(z^I)$ *and* $q = \mathrm{softmax}(z^I + \alpha b)$*, with* $\alpha \geq 0$ *and* $0 \leq b_v \leq D$*. Then*

$$\log \frac{q_v}{q_w} - \log \frac{p_v}{p_w} = \alpha(b_v - b_w), \qquad \left|\log \frac{q_v/q_w}{p_v/p_w}\right| \leq \alpha D. \tag{41}$$

For any nonempty proper vocabulary subset $\mathcal{G}$*,*

$$\frac{q(\mathcal{G})}{1 - q(\mathcal{G})} = \frac{p(\mathcal{G})}{1 - p(\mathcal{G})} \frac{\mathbb{E}_p[e^{\alpha b_v} \mid v \in \mathcal{G}]}{\mathbb{E}_p[e^{\alpha b_v} \mid v \notin \mathcal{G}]}. \tag{42}$$

Proof. Softmax reweighting gives $q_v = p_v e^{\alpha b_v} / \sum_w p_w e^{\alpha b_w}$. Taking a ratio cancels the common normalization and proves the pairwise identity and bound. Summing the unnormalized masses over $\mathcal{G}$ and its complement gives Equation 42. □

For Equation 24, $D = [c - m]_+$ because clipping and positive top-k selection bound each boost. At the reported $(\alpha, c, m) = (0.15, 5, 0)$, the pairwise log-odds change is at most 0.75 within a denoising step. The fixed-plan factorial complements this bound by showing that the implemented tilt suppresses unsupported realization more strongly than supported realization.