MASKED-STATE EVIDENCE FOR CORRECTING OBJECT HALLUCINATIONS IN DIFFUSION VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object hallucinations remain common in diffusion vision–language models (dVLMs). Unlike autoregressive VLMs, which expose a causal prefix that can be rolled back, dVLMs resolve a masked response canvas in parallel. Consequently, commitments need not follow a unique order, and completed bidirectional context can reconstruct an unsupported semantic state after it is reopened. We introduce *Evidence-Guided Commitment Correction* (EGCC), a training-free framework that treats factual revision as a controlled transition in this diffusion state space. Our key observation is that a dVLM can verify before it answers; at the still-masked answer state of a fixed object-existence probe, a continuous score from the dVLM's native token distribution better ranks objects by unsupportedness than the probe's hard decoded answer and remains informative even among commitments that receive the same decoded answer. Across all three dVLM backbones, unsupported object commitments also remain semantic attractors: simply reopening their spans reconstructs the same object in 80–86% of triggered cases. EGCC leverages masked-state evidence to detect these weak commitments and mediate a sparse transition to a locally reachable alternative. It preserves every non-target token by construction, constrains transition to be non-expansive over extracted object commitments, and requires each newly introduced object identified by the extractor to be better supported than the reopened source. Evidence estimation and revision use the same frozen dVLM, without an auxiliary verifier or full-response rollouts. With one numerical policy frozen across models and benchmarks, EGCC reduces measured object hallucination for LLaDA-V, MMaDA, and LaViDA on both COCO CHAIR and AMBER-Generative, with reductions of up to 3.8 CHAIR-S points and 11.7 AMBER Hal points. These results demonstrate masked-state evidence as an actionable interface for post-completion semantic correction in dVLMs.

1 INTRODUCTION

Vision–language models can fluently describe objects absent from the image: an unsupported phrase may be linguistically probable while visually false (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023; Kaul et al., 2024). Most mitigation methods target autoregressive (AR) VLMs, modifying next-token distributions or returning to an earlier causal prefix to regenerate the suffix (Leng et al., 2024; Huang et al., 2024; Favero et al., 2024; Chen et al., 2024; Yang et al., 2026b).

Diffusion VLMs instead resolve a masked canvas under bidirectional context, without a unique left-to-right commitment order or canonical rollback prefix (Nie et al., 2025; You et al., 2026; Yang et al., 2025; Li et al., 2025). They can revise completed responses by remasking selected positions while preserving the rest, but the surrounding context can reconstruct the original unsupported object.

Recent methods repair uncertain or context-brittle tokens through remasking (Huang et al., 2026; Zhai et al., 2026; Yao, 2026), or adjust dVLM commitment order using visual attention and denoising dynamics (Narnaware et al., 2026; Kim et al., 2026a;b). Yet an absent object can remain a confident continuation. Post-completion correction therefore requires *image-conditioned semantic evidence* to determine which commitments to reopen and which alternatives to retain.

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

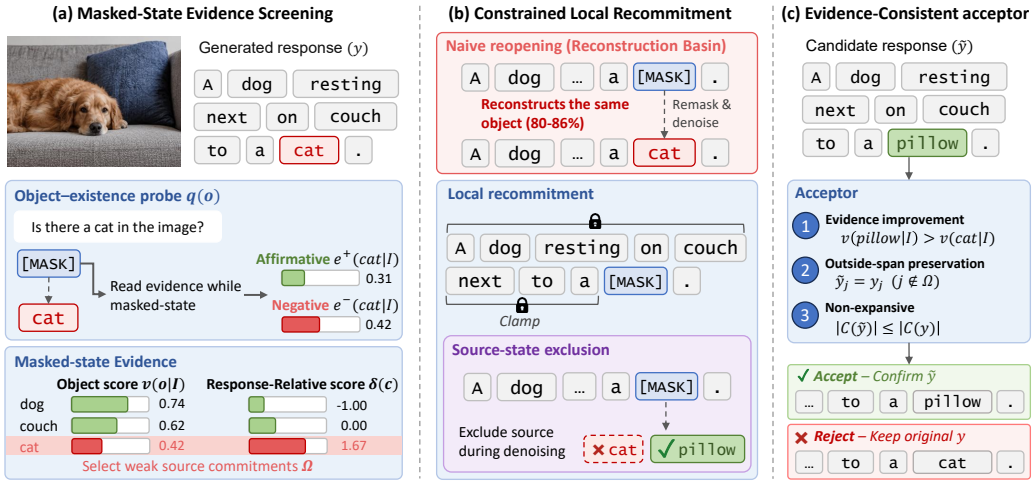

Figure 1: **Overview of EGCC.** Masked-state evidence guides weak-commitment selection, constrained local recommitment, and candidate acceptance through structural checks and evidence assessment of newly introduced objects.

Our first observation is that a dVLM can verify before it answers. For the same image, object-existence probe, and model, we read the answer distribution while its first position remains masked. This graded score yields higher unsupported-object detection ROC-AUC and PR-AUC than the hard decoded answer across LLaDA-V, MMaDA, and LaViDA (Figure 2).

Our second observation is that revisability does not imply correctability. Naive remasking reconstructs the same object in 80–86% of triggered responses, revealing a semantic reconstruction basin. Broad regeneration can escape this basin but may discard supported content or introduce new hallucinations, motivating a constrained local transition.

We introduce *Evidence-Guided Commitment Correction* (EGCC; Figure 1), which uses one masked-state evidence ordering to select weak source commitments and assess extracted replacement objects. Exact non-target clamping, temporary source-state exclusion, and conservative retention limit collateral rewriting, source reconstruction, and unsupported substitution, respectively.

EGCC is training-free and uses the same frozen dVLM for evidence and revision, without an auxiliary verifier or full-caption rollouts. One frozen numerical policy reduces all reported hallucination metrics relative to vanilla across the three backbones on COCO (Lin et al., 2014) and AMBER-Generative (Wang et al., 2023), and extends to generic noun-phrase proposals on OpenCHAIR Ben-Kish et al. (2024) without retuning.

Our contributions are threefold:

- We identify **masked-state visual evidence** in off-the-shelf dVLMs. Specifically, before a verification answer is committed, the native masked-token distribution is substantially more diagnostic of unsupported generated objects than the decoded answer to the same probe.
- We characterize a **semantic reconstruction basin** in completed dVLM responses and formulate **evidence-consistent recommitment**, which couples source selection and alternative retention through one model-native evidence ordering in the diffusion state space.
- We provide a **controlled cross-family evaluation** over three dVLMs and two generative hallucination benchmarks. Crucially, our analysis isolates semantic visual evidence from token confidence, generic remasking, and broad second-pass regeneration.

2 Related Work

Object hallucination in open-ended generation. CHAIR (Rohrbach et al., 2018) measures unsupported object mentions in free-form captions, POPE Li et al. (2023) probes object existence through

binary questions, and AMBER (Wang et al., 2023) provides both discriminative and generative protocols. THRONE (Kaul et al., 2024) shows that free-form hallucination cannot be reduced to short-form probing, while OpenCHAIR (Ben-Kish et al., 2024) exposes the limited coverage of closed object ontologies.

Internal evidence and selective correction in AR-VLMs. Probabilistic self-evaluation uses a model's answer distribution to assess proposed answers, as in Kadavath et al. (2022). AR-VLM mitigation modifies causal decoding through visual contrast (Leng et al., 2024), mutual-information amplification (Favero et al., 2024), attention penalties and retrospection (Huang et al., 2024), or focal visual contrast (Chen et al., 2024). Internal-state studies further show that hallucination or object grounding can be predicted from pre-generation representations, hidden states, or attention-derived evidence (Kogilathota et al., 2026; Hasanebrahimi et al., 2026; Yang et al., 2026b). VGD (Yang et al., 2026b) uses learned object-level readouts to selectively roll back a causal prefix, suppress the rejected object identity, and regenerate the continuation. REVERSE (Wu et al., 2025) trains a VLM to flag hallucinated phrases and uses the resulting token probabilities to trigger retrospective resampling.

Post-generation systems may instead decompose, verify, and rewrite claims. Woodpecker (Yin et al., 2024) extracts objects, verifies them with external visual tools, and corrects the response without retraining. LURE (Zhou et al., 2024) trains a revisor using uncertainty and position, while evidence-bounded editing and atomic self-verification provide conservative or calibrated refinement in AR-VLMs (Mishra et al., 2026; Liu et al., 2026).

Remasking and self-correction in diffusion models. Discrete diffusion supports infilling and flexible token revision (Lou et al., 2023; Nie et al., 2025). ReMDM (Wang et al., 2025) formalizes remasking samplers for pretrained masked diffusion models. RemeDi (Huang et al., 2026) learns token-quality scores and a remasking policy. CoRE (Zhai et al., 2026) identifies context-brittle tokens through masked-context perturbations and T2M (Yao, 2026) replaces direct token editing with training-free token-to-mask refinement. In multimodal masked diffusion, Ouyang et al. (2026) use current or accumulated token likelihoods to guide training-free remasking during generation. Remasking can also return a position to the same token, producing ineffective revision cycles (Xiang et al., 2026). EGCC instead uses a separate post-completion object-existence probe to guide source selection and replacement assessment.

Grounded inference in diffusion VLMs. LLaDA-V (You et al., 2026), MMaDA Yang et al. (2025), and LaViDA Li et al. (2025) instantiate masked multimodal diffusion with different architectures and schedules. ReDiff (Ji et al., 2025) learns to correct syntactic errors and semantic hallucinations through synthetic-error revision and expert-supervised self-correction, updating already generated tokens during decoding. Thinking Diffusion (Kim et al., 2026a) studies premature answer commitment and visual dependence during multimodal reasoning. VISAGE (Narnaware et al., 2026) re-ranks token commitments using spatial attention structure and ST-Veto (Kim et al., 2026b) uses temporal confidence dynamics to control an evolving generation trajectory. EGCC uses the same frozen dVLM for masked-state evidence and local revision, preserving all non-target tokens.

3 Problem Setting and Observations

3.1 Object commitments and correction requirements

Let p_θ be a dVLM that generates response $y = (y_1, \ldots, y_T)$ for image I and instruction x. We represent each expressed object as a commitment $c_i = (o_i, \Omega_i)$, where o_i is the object concept and $\Omega_i \subseteq \{1, \ldots, T\}$ contains the token positions that realize it:

$$\mathcal{C}(y) = \{c_i = (o_i, \Omega_i)\}_{i=1}^{N_y},$$

Repeated mentions are separate commitments, and one realization may span multiple subword tokens.

An effective correction operator must identify visually unsupported commitments, revise them without rewriting unrelated context, and avoid replacing one unsupported assertion with another. In particular, reopening positions Ω should preserve every unselected token,

$$y_j^\star = y_j \qquad \text{for every } j \notin \Omega. \tag{1}$$

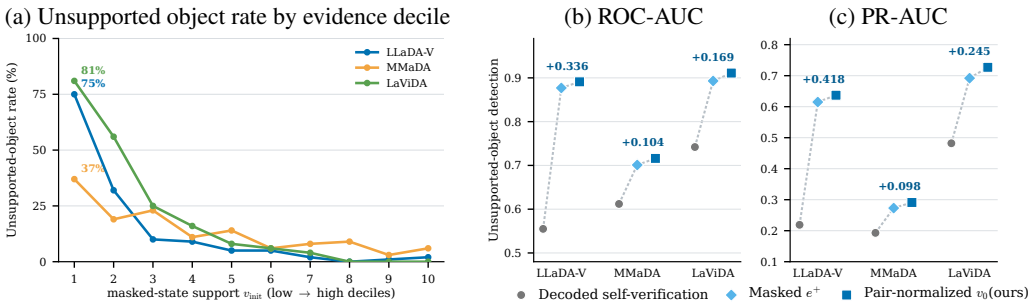

Figure 2: **Masked-state evidence stratifies unsupported-object incidence.** (a) Unsupported-object frequency shows a strong downward trend from low to high v_{init} deciles. (b–c) Across the extracted probe pool, masked-state support is more discriminative than decoded self-verification.

and an accepted transition should not increase the number of object commitments. Remasking alone neither identifies visually unsupported commitments nor ensures that newly introduced objects are better supported. We next establish the two empirical observations that motivate EGCC.

3.2 Masked states expose graded evidence beyond decoded answers

For each object in a completed response, we issue a separate object-existence probe to the same frozen dVLM. With the probe answer region still fully masked, we read affirmative and negative probabilities at its first position and normalize them into a continuous support score, v_{init}. Section 4.1 gives the formal definition. We compare this score with the fully decoded response to the same probe.

Observation 1 (Decoded answers omit a graded evidence ordering). Across all three backbones, v_{init} better discriminates unsupported objects than the hard decoded answer to the same probe on both ROC-AUC and PR-AUC; pair normalization also improves both metrics over raw affirmative evidence (Figure 2b–c). The unsupported-object rate shows a strong downward trend from the lowest to the highest v_{init} deciles (Figure 2a). This ordering remains informative even among commitments that all decode to *yes*: the lowest v_{init} decile is 2–5$\times$ more hallucination-dense than the within-group base rate, distinguishing risk levels that the shared hard answer cannot express. As a separate diagnostic, restricting evaluation to commitments with $v_{\text{init}} > 0.5$ (87–97% of the extracted probe pool) yields ROC-AUC scores of .878, .695, and .871 on LLaDA-V, MMaDA, and LaViDA, respectively. These results motivate identifying correction targets through graded support rather than decoded rejection alone. A response-level example and its denoising trajectory are provided in Appendix Table 12 and Figure 6.

3.3 Completed context forms a reconstruction basin

Observation 2 (Local remasking is attracted to the source state). On the responses selected for correction by the policy in Section 4.2, we compare local denoising with and without source-state exclusion, holding selected commitments and reopening scope fixed. Ordinary local denoising reconstructs the same object in 80–86% of these responses across the three backbones. Temporarily excluding the rejected surface forms reduces source retention to 0–17% (Figure 3; detailed protocol in Section 6.2). The completed bidirectional context therefore enables coherent local infilling but also attracts a reopened span back to its original semantic state. A corrector must leave this source basin without the collateral rewriting caused by broad regeneration.

4 Evidence-Guided Commitment Correction

The observations above motivate the three stages of EGCC (Figure 1): masked-state evidence screening, constrained local recommitment, and evidence-consistent candidate acceptance. Screening combines evidence readout with source selection; local recommitment discourages source reconstruction while preserving unselected context; and the acceptor checks structural constraints and assesses

newly introduced objects under the same evidence score. The contribution is this single-evidence correction interface as a whole: span remasking, clamping, and logit suppression are minimal implementation components rather than separate claims.

4.1 MASKED-STATE EVIDENCE READOUT

Generation confidence measures contextual predictability rather than visual support. EGCC instead queries the same dVLM with a fixed object-existence probe $q(o)$ and reads its native token distribution before it commits to an answer. Let $\mathbf{m}_{\text{init}}$ be a fully masked answer region, z_1 its first position, and $\mathcal{V}^+$ and $\mathcal{V}^-$ tokenizer-compatible affirmative and negative verbalizer sets. We compute

$$e^s_{\text{init}}(o \mid I) = \max_{w \in \mathcal{V}^s} p_\theta(z_1 = w \mid I, q(o), \mathbf{m}_{\text{init}}), \qquad s \in \{+, -\}. \tag{2}$$

and normalize the two sides,

$$v_{\text{init}}(o \mid I) = \frac{e^+_{\text{init}}(o \mid I)}{\max\{e^+_{\text{init}}(o \mid I) + e^-_{\text{init}}(o \mid I), \epsilon_e\}}. \tag{3}$$

The subscript "init" denotes the fully masked answer state before any reverse-denoising update. Here ϵ_e is a numerical stabilizer. For a commitment $c = (o_c, \Omega_c)$, $v_{\text{init}}(o_c \mid I)$ is its masked-state visual support. The queried object comes from the model's own response, and each score requires one masked-state forward; no reference caption, external verifier, or decoded probe trajectory is used.

4.2 SELECTION FROM MASKED-STATE EVIDENCE

For responses with multiple commitments, source selection requires both low absolute support and a sufficiently large support deficit relative to other commitments in the same response. The absolute ceiling limits intervention to low-support objects, while the relative criterion further restricts selection to objects whose support falls sufficiently below the response median. We normalize this deficit by the within-response median absolute deviation, a standard robust measure of scale (Rousseeuw & Croux, 1993). For $N_y > 1$, let

$$\begin{aligned} m_y &= \operatorname*{median}_{c' \in \mathcal{C}(y)} v_{\text{init}}(o_{c'} \mid I), \\ \delta_y(c) &= \frac{m_y - v_{\text{init}}(o_c \mid I)}{\operatorname*{median}_{c' \in \mathcal{C}(y)} |v_{\text{init}}(o_{c'} \mid I) - m_y| + \epsilon_\delta}. \end{aligned} \tag{4}$$

Here $\delta_y(c)$ measures how far commitment c falls below the within-response support median, and ϵ_δ is a numerical stabilizer. We select every commitment satisfying $v_{\text{init}}(o_c \mid I) \le \eta$ and $\delta_y(c) > \tau$. For $N_y = 1$, only the absolute condition is used. The same η and τ are selected once on the development split and then frozen across models and evaluation benchmarks. Degenerate cases and phrase boundaries are given in Appendix A.

Selection operates on semantic commitments rather than isolated tokens. Each selected object contributes a compact phrase span containing its complete token realization and the nearby words required to reform it. Multiple selected spans are reopened jointly so that their local syntax can resolve coherently.

4.3 FROM EVIDENCE TO A CONSTRAINED LOCAL TRANSITION

Let Ω be the union of selected phrase positions. These positions are returned to `[MASK]`, while every $j \notin \Omega$ is clamped to its original token throughout local denoising, enforcing Equation 1 by construction. Because ordinary remasking usually reconstructs the source, we temporarily reduce the logits of its tokenizer-compatible surface forms. If $\ell_{j,t}(v)$ is the logit for token v at local step t and $\mathcal{V}_{\text{src}}$ contains the selected source forms, then

$$\tilde{\ell}_{j,t}(v) = \ell_{j,t}(v) - \gamma\, \mathbf{1}[v \in \mathcal{V}_{\text{src}}], \qquad j \in \Omega. \tag{5}$$

Denoising the reopened positions with these logits yields candidate $\tilde{y}$. The fixed penalty γ acts only during this local transition and does not modify the model or globally suppress an object vocabulary.

Table 1: Generated object hallucination on COCO under the official CHAIR protocol. CHAIR scores are percentages (lower is better); Δ H is the net reduction in unsupported object mentions relative to vanilla (higher is better). Bold marks the best result for each backbone. "–" denotes an unavailable validated reproduction.

	LLaDA-V			MMaDA			LaViDA		
Method	CHAIR-S	CHAIR-I	Δ H	CHAIR-S	CHAIR-I	Δ H	CHAIR-S	CHAIR-I	Δ H
Vanilla	24.80	6.38	0	9.80	7.93	0	22.60	8.18	0
VISAGE	25.40	6.69	-7	**9.00**	7.29	4	21.20	7.95	7
MPD-PAC	24.00	6.40	4	–	–	–	22.40	7.86	14
Info-Gain	24.40	6.76	-8	11.80	8.91	-9	20.80	7.80	20
EGCC	**21.60**	**5.71**	**19**	**9.00**	**7.03**	**7**	**18.80**	**7.02**	**32**

Source exclusion is not used as evidence; it serves only as a transition constraint that discourages the reopened span from collapsing back to the rejected source state.

Escaping the source basin does not guarantee a safer candidate. We therefore reject any candidate that is empty, changes a token outside Ω, expands the total number of object commitments, duplicates an existing object, or adds an object without removing a selected source commitment. All object-level checks operate on commitments identified by the extractor. Let $\mathcal{O}_{\mathrm{src}}$ denote the selected source concepts and $\mathcal{O}_{\mathrm{add}}$ the newly introduced objects identified in $\tilde{y}$. For each $a \in \mathcal{O}_{\mathrm{add}}$, we compute the same masked-state score and retain the candidate only if

$$v_{\mathrm{init}}(a \mid I) > \max\left\{\frac{1}{2}, \max_{o \in \mathcal{O}_{\mathrm{src}}} v_{\mathrm{init}}(o \mid I)\right\}, \qquad a \in \mathcal{O}_{\mathrm{add}}. \tag{6}$$

Otherwise, EGCC returns the original response. A deletion introduces no new assertion and is authorized by weak-source selection together with locality and non-expansion. Thus, one masked-state evidence coordinate governs both ends of the transition: it identifies the source to reopen and orders any replacement against that source. Details are provided in Appendix A.1 and Algorithm 1.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate LLaDA-V (You et al., 2026), MMaDA (Yang et al., 2025), and LaViDA (Li et al., 2025). COCO (Lin et al., 2014) uses 500 random images with official CHAIR, AMBER-Generative uses its complete 1,004-image split with official CHAIR, Hal, and Cog measures (Wang et al., 2023), and OpenCHAIR random-300 tests generic noun-phrase proposals (Ben-Kish et al., 2024).

Each backbone's exact vanilla response is the primary reference. COCO comparisons include VISAGE (Narnaware et al., 2026), MPD-PAC (Hong et al., 2026), and Information-Gain sampling (Yang et al., 2026a) using public settings. AMBER includes VISAGE, whose attention-based ranking can be transferred under a matched open-ended setting across all three backbones. The pair-normalized selection policy is chosen once on a disjoint COCO development split and frozen across all models and benchmarks; no dataset-level edit budget or model-specific threshold is used. All main comparisons are paired against exact vanilla responses. Appendix A provides generation and evaluation details, Appendix E documents baseline integrations, and Appendix B.1 reports policy sensitivity.

5.2 MAIN RESULTS

Consistent reduction across dVLM families. The frozen EGCC policy lowers both CHAIR endpoints for all three backbones (Table 1), with relative CHAIR-S reductions of 12.9%, 8.2%, and 16.8% on LLaDA-V, MMaDA, and LaViDA. It also achieves the largest net reduction in unsupported mentions (Δ H=19,7,32, respectively), showing that the gains extend beyond hallucination rates to absolute counts. MMaDA generates substantially shorter responses (12.3 versus 106–108 mean words), resulting in fewer object proposals and correction opportunities; Appendix B provides paired per-backbone accounting.

Table 2: Frozen-policy transfer to AMBER-Generative. CHAIR, Cover, Hal, and Cog are percentages; Δ H is the net reduction in unsupported object mentions relative to vanilla (higher is better).

Model	Method	CHAIR↓	Cover↑	Hal↓	Cog↓	Δ H↑
LLaDA-V	Vanilla	4.79	37.36	14.34	0.57	0
	VISAGE	5.00	**37.63**	15.04	0.63	-11
	EGCC	**4.09**	37.38	**12.15**	**0.51**	**34**
MMaDA	Vanilla	9.06	**33.54**	18.33	0.83	0
	VISAGE	9.09	33.31	18.23	0.81	2
	EGCC	**8.84**	33.02	**17.83**	**0.71**	**11**
LaViDA	Vanilla	8.68	**54.10**	44.92	3.72	0
	VISAGE	8.21	52.44	41.83	3.80	61
	EGCC	**6.44**	53.87	**33.27**	**2.34**	**201**

Table 3: Volume-matched selector comparison on COCO random-500 at final-policy intervention volume of each backbone. All rows share the same local transition and retention rule and differ only in commitment selection. Scores are percentages; ΔH is the net reduction in unsupported object mentions relative to vanilla.

	LLaDA-V			MMaDA			LaViDA		
Selector signal	CHAIR-S↓	CHAIR-I↓	ΔH↑	CHAIR-S↓	CHAIR-I↓	ΔH↑	CHAIR-S↓	CHAIR-I↓	ΔH↑
Vanilla	24.80	6.38	0	9.80	7.93	0	22.60	8.18	0
Decoded self-verif.	22.80	5.92	13	9.40	7.76	2	19.20	7.21	27
Low confidence	23.60	6.18	6	**8.80**	7.29	4	20.80	7.78	12
Denoising entropy	25.00	6.46	−1	9.40	7.44	4	21.80	7.93	9
Pair-norm., v_{init} (ours)	**21.60**	**5.71**	**19**	9.00	**7.03**	**7**	**18.80**	**7.02**	**32**

The same frozen policy reduces CHAIR, Hal, and Cog for every backbone without benchmark-specific retuning (Table 2). These gains also hold in absolute counts, with net reductions of 34, 11, and 201 unsupported object mentions on LLaDA-V, MMaDA, and LaViDA, respectively. Cover remains within 0.52 percentage points of vanilla, providing no evidence of broad suppression of grounded content; on LaViDA, Hal falls by 11.65 points while Cover changes by only -0.23 points. VISAGE worsens at least one hallucination measure on every backbone, whereas EGCC improves every reported hallucination endpoint. Together, these results support the transferability of a single correction policy across the evaluated backbones and benchmarks. Representative cross-family corrections and comparisons are provided in Appendix F.

Open-vocabulary transfer and efficiency. With generic noun-phrase proposals, the same frozen policy lowers the official OpenCHAIR score for all three backbones, with a maximum change of approximately 1.9% in extracted concrete-object occurrences. Mean COCO response length is unchanged, object recall stays within 0.31 points of vanilla, and AMBER Cover stays within 0.52 points. The conservative instrumented upper bound adds 0.50–4.74% amortized runtime (Table 6); detailed scores and accounting appear in Appendices G–J.

6 Mechanistic Analysis

End metrics alone cannot establish whether EGCC succeeds because it selects better states or because any second denoising pass helps. We therefore perform controlled interventions on selection, source-state transition, candidate retention, and reopening scope.

6.1 Visual evidence is distinct from token uncertainty

To compare selection quality at matched intervention volume, we hold the commitment pool, local transition, and retention rule constant. The pair-normalized row uses the exact frozen EGCC selector from Section 4.2 and therefore reproduces the final policy.Besides CHAIR-S/I, ΔH reports the net reduction in unsupported object mentions relative to the exact vanilla responses.

Pair-normalized masked-state evidence has the strongest cross-family profile (Table 3): it removes at least as many unsupported mentions as any alternative and gives the best CHAIR-S/I on LLaDA-V and LaViDA while improving both endpoints on MMaDA. At matched volume, it strictly outperforms decoded self-verification in unsupported-mention removal on every backbone. Low confidence achieves lower CHAIR-S on MMaDA, but EGCC achieves lower CHAIR-I and removes more unsupported mentions on every backbone. With intervention volume matched, these gains support evidence-based source selection over generation-time uncertainty.

Transfer to benchmark-native object queries. We isolate the evidence component from commitment parsing and recommitment on the unmodified full POPE benchmark (9,000 questions) (Li et al., 2023). Under the official protocol, replacing the decoded decision with the same frozen midpoint rule $v_{\text{init}} > 0.5$ improves both overall accuracy and F1 for every backbone (Table 4). The corresponding readout ROC-AUC scores are 0.938, 0.806 and 0.931. These results support transfer of the masked-state readout from generated commitments to benchmark-provided object queries, without local recommitment or candidate acceptance.

Table 4: Masked-state readout on full POPE.

	LLaDA-V		MMaDA		LaViDA	
Readout method	Acc	F1	Acc	F1	Acc	F1
Decoded answer	73.3	78.1	52.3	67.7	83.3	81.0
Pair-norm. v_{init}	80.8	83.0	63.7	72.8	83.8	81.9

6.2 Source-state exclusion enables alternative local transitions

On the same controlled subset, holding selected commitments and reopening scope fixed, ordinary local remasking returns to the source object—either by exact restoration or a source-preserving rewrite—in 85.7%, 85.7%, and 79.5% of LLaDA-V, MMaDA, and LaViDA cases, respectively. Source-state exclusion lowers these rates to 17.1%, 0.0%, and 14.5% (Figure 3). This controlled comparison shows that an additional denoising pass alone rarely escapes the reconstruction basin; exclusion is what makes alternative local states reachable.

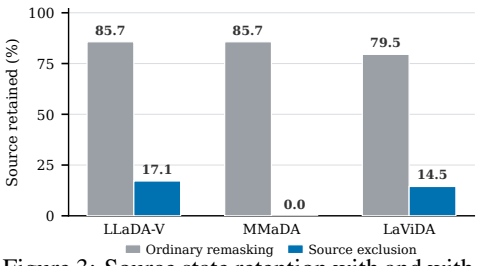

Figure 3: Source state retention with and without exclusion.

With selection and local candidates fixed, Table 5 reports hallucination-reducing (R) and harmful (H) edits retained by the combined structural-and-evidence candidate acceptor; neutral edits are excluded from these counts.

Given the alternative local states opened by exclusion, the combined candidate acceptor preserves 239 of 258 hallucination-reducing edits (92.6%) and blocks 15 of 34 harmful edits (44.1%; Table 5). It acts as a conservative filter on states made reachable by source exclusion, rather than as a second source selector. The remaining harmful edits primarily arise from lexical substitution outside the internal commitment vocabulary (Appendix H).

6.3 Locality controls the factuality-preservation trade-off

Holding the same controlled subset and all other components fixed, we vary only reopening scope: the local phrase used by EGCC, the suffix from the earliest selected position, or the full generated response. Suffix reopening serves as an AR-like scope control: revising an early commitment requires regenerating all subsequent positions, whereas local reopening exploits the arbitrary-position infilling capability native to dVLM denoising.

Across all scopes, non-target preservation is measured at the same original token positions outside the local target spans selected by EGCC. Local reopening preserves these tokens exactly by construction. This metric measures contextual rewriting, not whether each change is harmful; supported-object loss separately measures the net loss of grounded object mentions.

On LaViDA, suffix reopening removes 41 unsupported mentions but loses 60 supported ones, versus 32 and 9 under local correction (Figure 4). Local reopening also removes more unsupported mentions with fewer supported losses on LLaDA-V, whereas suffix reopening is slightly better on MMaDA.

Table 5: Effect of the structural-and-evidence candidate acceptor on AMBER-Generative with selection and recommitment candidates fixed. R/H denotes hallucination-reducing/harmful edits.

Model	Off (R/H)	On (R/H)	R kept↑	H blk↑
LLaDA-V	33/7	32/4	97.0%	42.9%
MMaDA	24/17	19/9	79.2%	47.1%
LaViDA	201/10	188/6	93.5%	40.0%
All	258/34	239/19	**92.6%**	**44.1%**

Table 6: Conservative instrumented runtime upper bound on paired COCO random-500. Measurements use one H100 80GB at batch size one.

Model	Added (s)	Overhead
LLaDA-V	0.52	2.53%
MMaDA	0.11	0.50%
LaViDA	0.26	4.74%
Macro avg.	0.30	2.59%

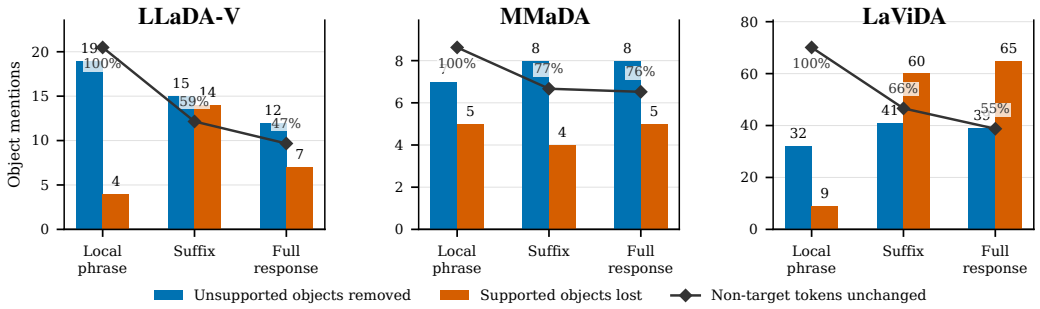

Figure 4: **Local recommitment preserves non-target tokens by construction.** Bars compare net unsupported-object removal and supported-object loss as reopening scope widens; black curves show the fraction of original tokens outside the local target spans left unchanged, using the same reference positions across all scopes.

Thus, locality provides exact non-target preservation while retaining substantial correction gains, rather than uniformly optimizing object-level outcomes across backbones.

7 Limitations

Our evaluation focuses on generated object commitments; extending the evidence interface to attributes, relations, counts, and external-world facts remains future work. Evidence quality and correction opportunities depend on the frozen backbone and its output regime. Object-level safeguards apply to extracted commitments and can miss substitutions outside the extractor's coverage (Appendix H). Although the frozen policy transfers to generic noun-phrase proposals on OpenCHAIR without retuning, the gains are modest; broader evaluation is needed to establish its effectiveness across open-vocabulary settings.

8 Conclusion

We studied object hallucination as a semantic-state correction problem specific to diffusion vision–language models. Unlike causal AR correction, a dVLM must revise an arbitrary region of a completed bidirectional canvas, where ordinary remasking can be attracted back to the same semantic source. Our central finding is that the native masked answer distribution, read before a verification trajectory commits, exposes graded visual evidence that is substantially more diagnostic than the decoded answer to the same query. EGCC turns this evidence into an actionable correction interface: one evidence ordering governs which semantic source is reopened and which locally reachable alternative may be retained, while the transition preserves unrelated context by construction. One frozen policy improves all COCO and AMBER hallucination metrics across three heterogeneous dVLMs, while OpenCHAIR evaluation shows that the same frozen policy extends to generic object proposals without retuning. The contribution is a dVLM-native connection between masked-state visual evidence and post-completion semantic recommitment.

AI USE STATEMENT

In this work, we used generative AI tools for surveying prior work, generating images for figures, feedback on the interpretation of experimental results, and checking grammar. We have not used generative AI tools for generating synthetic datasets, cleaning datasets, or reformatting datasets. Additionally, we used generative AI tools for drafting parts of the paper and improving prose readability and sentence structure. We have reviewed all AI-assisted work. All AI-refined text was manually reviewed and verified by the authors to ensure clarity, accuracy, and appropriate context. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Md Atabuzzaman and Chris Thomas. Reliability challenges in diffusion vision-language models. *arXiv preprint arXiv:2609.01318*, 2026. URL https://arxiv.org/abs/2609.01318.

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*, pp. 22680–22698, 2024.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846. PMLR, 21–27 Jul 2024.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312. IEEE, 2024.

Afsaneh Hasanebrahimi, Hanxun Huang, Christopher Leckie, and Sarah Erfani. VisER: Visual evidence and reliance for object hallucination detection in LVLMs. *arXiv preprint arXiv:2608.30480*, 2026. URL https://arxiv.org/abs/2608.30480.

Sujung Hong, Chanyong Yoon, and Seong Jae Hwang. Mitigating mask prior drift and positional attention collapse in large diffusion vision-language models. *arXiv preprint arXiv:2605.14530*, 2026. URL https://arxiv.org/abs/2605.14530.

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

Zemin Huang, Yuhang Wang, Zhiyang Chen, and Guo-Jun Qi. Don't settle too early: Self-reflective remasking for diffusion language models. In *International Conference on Learning Representations*, volume 2026, pp. 56998–57026, 2026.

Yatai Ji, Teng Wang, Yuying Ge, Zhiheng Liu, Sidi Yang, Ying Shan, and Ping Luo. From denoising to refining: A corrective framework for vision-language diffusion model. *arXiv preprint arXiv:2510.19871*, 2025. URL https://arxiv.org/abs/2510.19871.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, Christopher J Taylor, and Stefano Soatto. THRONE: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27218–27228. IEEE, 2024.

Keuntae Kim, Mingyu Kang, and Yong Suk Choi. Thinking Diffusion: Penalize and guide visual-grounded reasoning in diffusion multimodal language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5154–5164, June 2026a.

Keuntae Kim, Beomseok Lee, Hyunwoo Kim, and Yong Suk Choi. ST-Veto: Spatio-temporal token veto for diffusion MLLMs via Taylor prediction and visual grounding. *arXiv preprint arXiv:2607.17884*, 2026b. URL https://arxiv.org/abs/2607.17884.

Sai Akhil Kogilathota, Sripadha Vallabha EG, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 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 *Advances in Neural Information Processing Systems*, volume 38, pp. 105101–105134, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 2023.

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 *European conference on computer vision*, pp. 740–755. Springer, 2014.

Jiarui Liu, Weihao Xuan, Zhijing Jin, and Mona Diab. Taming object hallucinations with verified atomic confidence estimation. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5430–5444, 2026.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. *arXiv preprint arXiv:2310.16834*, 2023. URL https://arxiv.org/abs/2310.16834.

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.

Vishal Narnaware, Animesh Gupta, Kevin Zhai, Zhenyi Wang, and Mubarak Shah. Seeing to ground: Visual attention for hallucination-resilient MDLLMs. *arXiv preprint arXiv:2603.25711*, 2026. URL https://arxiv.org/abs/2603.25711.

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 *Advances in Neural Information Processing Systems*, volume 38, pp. 50608–50646, 2025.

Yidong Ouyang, Panwen Hu, Zhengyan Wan, Zhe Wang, Liyan Xie, Dmitriy Bespalov, Ying Nian Wu, Guang Cheng, Hongyuan Zha, and Qiang Sun. Training-free self-correction for multimodal masked diffusion models. *arXiv preprint arXiv:2602.02927*, 2026. URL https://arxiv.org/abs/2602.02927.

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.

Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical association*, 88(424):1273–1283, 1993.

Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/d877ea0dd78bdbe54830670618c1de09-Abstract-Conference.html.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. URL https://arxiv.org/abs/2311.07397.

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. URL https://arxiv.org/abs/2504.13169.

Yanzheng Xiang, Lan Wei, Yizhen Yao, Qinglin Zhu, Hanqi Yan, Chen Jin, Philip Alexander Teare, Dandan Zhang, Lin Gui, Amrutha Saseendran, et al. Stop the flip-flop: Context-preserving verification for fast revocable diffusion decoding. *arXiv preprint arXiv:2602.06161*, 2026. URL https://arxiv.org/abs/2602.06161.

Kaisen Yang, Jayden Teoh, Kaicheng Yang, Yitong Zhang, and Alex Lamb. Improving sampling for masked diffusion models via information gain. *arXiv preprint arXiv:2602.18176*, 2026a. URL https://arxiv.org/abs/2602.18176.

Lei Yang, Xinze Liu, Dayan Wu, Ding Wang, Hengjie Zhu, Zihao Zhang, Tianzhu Hu, Hanqi Wu, Peng Fu, and Zheng Lin. Hallucinations leave a grounding signature: Verifier-guided decoding for selective object correction. *arXiv preprint arXiv:2607.27823*, 2026b. URL https://arxiv.org/abs/2607.27823.

Ling Yang, Ye Tian, Bowen Li, Xinchen Zhang, Ke Shen, Yunhai Tong, and Mengdi Wang. MMaDA: Multimodal large diffusion language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 138867–138907, 2025.

Lin Yao. Remask, don't replace: Token-to-mask refinement in diffusion large language models. *arXiv preprint arXiv:2604.18738*, 2026. URL https://arxiv.org/abs/2604.18738.

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. doi: 10.1007/s11432-024-4251-x.

Zebin You, Shen Nie, Xiaolu Zhang, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. LLaDA-V: Large language diffusion models with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10093–10105, 2026.

Kevin Zhai, Sabbir Mollah, Zhenyi Wang, and Mubarak Shah. CoRE: Context-robust remasking for diffusion language models. *arXiv preprint arXiv:2602.04096*, 2026. URL https://arxiv.org/abs/2602.04096.

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

A Reproducibility Details

A.1 Complete inference procedure

Algorithm 1 states the executable procedure corresponding to the evidence, selection, local-transition, and acceptance equations in Section 4.

Frozen policy. The main method uses visual-support ceiling $\eta = 0.60$, response-relative threshold $\delta_y(c) > 0.5$, all evidence-eligible commitments, source-state penalty $\gamma = 8$, local joint reopening, and the acceptor in Section 4. The selection parameters are chosen once on a COCO val2014 random-500 development split by three-model macro $F_{0.25}$ and then frozen. Test data comprise a disjoint COCO random-500 split and the full 1,004-image AMBER-Generative split. No model- or test-benchmark-specific threshold is fitted.

Benchmarks and metrics. COCO (Lin et al., 2014) evaluation uses 500 val2014 images sampled at random, following standard CHAIR practice (Rohrbach et al., 2018; Huang et al., 2024; Leng et al., 2024): we report sentence-level CHAIR-S and mention-level CHAIR-I, where lower is better. AMBER-Generative uses its complete 1,004-image split with the official CHAIR, Hal, and Cog measures, and the OpenCHAIR random-300 evaluation replaces the benchmark-native proposal interface with a generic noun-phrase parser; its evaluator is accessed only after correction (Appendix I).

Reopening-scope implementation. The main local-phrase operator expands each selected object realization by at most three lexical words on either side, stopping at punctuation or a coordination, relative-clause, or copular boundary. The union of these compact spans is masked and denoised jointly, while every position outside the union is clamped to the completed response. In the scope ablation, *suffix* instead masks every position from the earliest selected local token through the end of the response, and *full response* masks every generated position. Selection, source-state exclusion, and candidate retention are held fixed across these controls.

Degenerate relative-risk cases. In Equation 4, we set $\epsilon_\delta = 10^{-8}$. For multiple commitments whose denominator dispersion is at most 10^{-12}, we replace m_y by the arithmetic mean and the dispersion term by the population standard deviation. Commitments are ranked first by $\delta_y(c)$ and then by raw support $v_{\text{init}}(o_c \mid I)$. These deterministic cases implement the same absolute-plus-relative rule without model- or benchmark-specific parameters.

Masked-state evidence probe. For object concept o, the fixed probe asks whether an instance of o is present in the image and requests a yes-or-no answer. The answer region is initialized entirely with mask tokens, and Equation 2 is evaluated at this initial state using only its first position. The sets $\mathcal{V}^+$ and $\mathcal{V}^-$ contain tokenizer-compatible forms of *yes* and *no*, including model-specific leading-space and capitalization variants; the maximum probability within each semantic set is used. We set the denominator floor in Equation 3 to $\epsilon_e = 10^{-12}$. Initial selection uses low $v_{\text{init}}(o \mid I)$ together with the response-relative score in Equation 4. Source scores are cached by image and normalized phrase: repeated mentions remain separate commitments but reuse the same score. For a newly introduced candidate object, the acceptor computes the same v_{init} from one fully masked initial-state forward and applies Equation 6. Thus, source selection and candidate retention are compared in an identical diffusion state and neither decision uses a decoded probe response.

Commitment extraction. We use the object parsing protocol paired with each evaluation benchmark. COCO captions are mapped with the public CHAIR categories and aliases; AMBER uses its public global object vocabulary. These resources identify object-denoting spans in generated text. They provide no image-specific labels, references, or correctness decisions to the selector, recommitment operator, or acceptor. Character spans are aligned back to each model tokenizer, and all subword positions realizing a selected phrase are reopened jointly.

Local-transition safeguards. Let $n_y(o)$ and $n_{\tilde{y}}(o)$ be the numbers of commitments expressing object o before and after local recommitment. A candidate must be nonempty and identical to y at every position outside Ω. It must satisfy $|\mathcal{C}(\tilde{y})| \leq |\mathcal{C}(y)|$ and $n_{\tilde{y}}(o) \leq n_y(o)$ for every object already present in y. If any object count increases, at least one selected source object's count must decrease.

Algorithm 1 Evidence-Guided Commitment Correction EGCC

Require: dVLM p_θ, image I, instruction x, complete response y; support ceiling η, relative threshold τ, and exclusion strength γ
Ensure: corrected response $y^\star$

Extract commitments $\mathcal{C}(y)$ and their token positions
if $\mathcal{C}(y) = \varnothing$ **then**
    return y
end if
for all $c = (o, \Omega_c) \in \mathcal{C}(y)$ **do**
    Retrieve $v_{\text{init}}(o \mid I)$ cached by image and normalized phrase, or compute and cache it with Equations 2–3
end for
if $N_y > 1$ **then**
    Compute response-relative score with Equation 4
end if
Set $\mathcal{C}_{\text{sel}}$ to all commitments satisfying $v_{\text{init}}(o \mid I) \le \eta$ and, when $N_y > 1$, $\delta_y(c) > \tau$
if $\mathcal{C}_{\text{sel}} = \varnothing$ **then**
    return y
end if
Expand selected commitments into local phrase spans
Set Ω to the union of token positions in these spans
Initialize $\tilde{y}^{(L)}$ by masking every position in Ω
Denoise the positions in Ω with Equation 5, clamping every $j \notin \Omega$ to y_j; obtain $\tilde{y}$
Identify $\mathcal{O}_{\text{add}}$ and $\mathcal{O}_{\text{drop}}$ from their mention counts
if $\tilde{y}$ violates a local-transition safeguard **then**
    return y
end if
if $\mathcal{O}_{\text{add}} = \varnothing$ **then**
    return $\tilde{y}$
end if
for all $a \in \mathcal{O}_{\text{add}}$ **do**
    Compute $v_{\text{init}}(a \mid I)$ from one fully masked initial-state forward
end for
Evaluate Equation 6 for every $a \in \mathcal{O}_{\text{add}}$
if every added object satisfies Equation 6 **then**
    return $\tilde{y}$
end if
return y

These checks permit deletion, object-preserving local revision, and one-for-one replacement while rejecting collateral rewriting, duplication, and unconstrained object insertion.

Generation settings. All models use validated native chat templates, image formatting, and model-compatible precision without 4/8-bit weight quantization. Table 7 reports the initial free-form caption-generation settings. The shorter object probes and sparse local recommitment passes are not included in these output budgets.

LLaDA-V uses a single diffusion block on both benchmarks. MMaDA uses its blockwise semi-autoregressive sampler, four 128-token blocks on COCO and one 64-token block on AMBER-Generative. The 256-step COCO entry denotes MMaDA's initial caption-generation schedule, not the cost of each EGCC operation. For MMaDA, decoded self-verification and trajectory diagnostics use a 16-token probe block with 16 denoising steps; the final EGCC policy reads evidence from a single initial-state forward without probe decoding. Local recommitment uses 64 scheduled steps over the 512-token sequence (16 updates per 128-token block), reopening only the selected positions. LaViDA follows its shifted transfer schedule ($s = 1/3$) (Li et al., 2025). LLaDA-V and LaViDA use 384-base any-resolution image grids, whereas MMaDA uses a fixed 512-pixel input.

Table 7: Baseline caption-generation configurations. Output budget, denoising steps, and block length follow each model's validated benchmark protocol. Token proposals are deterministic in all cases.

Model (checkpoint)	Benchmark	Budget	Steps	Block	Precision	Vision
LLaDA-V (`LLaDA-V`)	COCO	128	128	128	FP16	384
LLaDA-V (`LLaDA-V`)	AMBER-G	64	64	64	FP16	384
MMaDA (`MMaDA-8B-Base`)	COCO	512	256	128	BF16	512
MMaDA (`MMaDA-8B-Base`)	AMBER-G	64	64	64	BF16	512
LaViDA (`lavida-llada-v1.0`)	COCO	128	128	128	BF16	384
LaViDA (`lavida-llada-v1.0`)	AMBER-G	128	128	128	BF16	384

Evaluation. COCO CHAIR constructs supported objects from COCO instance annotations and human reference captions after all generated responses are fixed; AMBER uses its official generative evaluator.

B MMaDA Output Regime

Table 8: COCO random-500 response and intervention statistics under the frozen policy. Words and Obj./resp. are per-response means; other columns are totals. Object mentions follow official CHAIR counting; evidence probes cover phrases captured by EGCC's local span extractor, deduplicated by normalized phrase within each image.

Model	Words	Obj./resp.	Dedup. probes	Eligible	Jobs	Accepted
LLaDA-V	108.1	5.3	872	35	35	34
MMaDA	12.3	1.4	356	21	21	17
LaViDA	106.2	5.0	704	92	83	64

Table 8 summarizes the per-backbone response and intervention statistics. MMaDA's short output regime (12.3 mean words vs. 108 for the other backbones) yields fewer correction opportunities per response. This relative tendency is not unique to our implementation. In an independent evaluation under a common captioning setup, MMaDA-MixCoT produces the shortest caption among six evaluated dVLMs, while LLaDA-V and LaViDA are among the longest (Atabuzzaman & Thomas, 2026). Although its checkpoint and prompts differ from ours, the relative pattern supports concise generation as a model-family output regime. Our token traces confirm model-native early EOS, and all reported gains use paired comparisons within each backbone.

B.1 Sensitivity of the frozen policy

Table 9 varies one selection parameter independently around the frozen setting $(\eta, \tau) = (0.60, 0.50)$. All other components of the final policy are held fixed, and every evidence-eligible commitment is processed. All outputs use the same official CHAIR protocol as Table 1, which constructs image support from both COCO instance annotations and human reference captions.

Every displayed setting improves both CHAIR endpoints over vanilla on all three backbones. The support ceiling controls the factuality–preservation trade-off: a looser ceiling admits more weakly supported commitments, whereas a tighter ceiling intervenes more conservatively. Performance is stable over the tested relative-threshold range. We retain the development-frozen default rather than selecting the strongest test-set row.

C Evidence Diagnostics

Evidence diagnostics cover the object phrases captured by EGCC's local span extractor, deduplicated by normalized phrase within each image.

Commitment timing is included only as a diagnostic and is not used by the final policy. It is competitive on MMaDA but substantially less discriminative than masked-state evidence on LLaDA-

Table 9: Sensitivity of the frozen selection policy on COCO random-500. CHAIR values are percentages and lower is better; ΔH↑ is the net reduction in unsupported object mentions and ΔS↓ the net loss of supported object mentions relative to vanilla. Each row changes one parameter from the default $(\eta, \tau) = (0.60, 0.50)$ while holding the remaining correction pipeline fixed; no setting is selected separately for a backbone. C_S and C_I denote CHAIR-S and CHAIR-I, respectively.

			LLaDA-V				MMaDA				LaViDA			
Setting	η	τ	C_S	C_I	ΔH↑	ΔS↓	C_S	C_I	ΔH↑	ΔS↓	C_S	C_I	ΔH↑	ΔS↓
Vanilla	–	–	24.80	6.38	0	0	9.80	7.93	0	0	22.60	8.18	0	0
Default	0.60	0.50	21.60	5.71	19	4	9.00	7.03	7	5	18.80	7.02	32	9
Looser ceiling	0.64	0.50	21.20	5.71	19	7	9.00	7.06	7	8	18.80	6.95	34	13
Tighter ceiling	0.56	0.50	22.60	5.85	15	4	9.40	7.42	4	3	19.00	7.14	29	11
Lower rel. threshold	0.60	0.25	21.60	5.71	19	6	9.20	7.03	7	5	18.80	6.95	34	12
Higher rel. threshold	0.60	0.75	21.80	5.78	17	4	9.00	7.03	7	5	19.20	7.06	31	9

Table 10: Where the evidence is measured. Unsupported-object detection on the extracted probe pool from held-out COCO responses (ROC-AUC / PR-AUC). The first three rows use the same object probe and compare the decoded decision with two scores obtained while the first answer position remains masked; timing is a separate diagnostic. Pair-normalization v_{init} is explained in Equation 3.

Probe signal	LLaDA-V	MMaDA	LaViDA
Decoded self-verification	0.555 / 0.219	0.612 / 0.193	0.742 / 0.482
Raw masked-state evidence, e^+_{init}	0.877 / 0.615	0.701 / 0.273	0.893 / 0.692
Pair-normalized evidence, v_{init} (ours)	**0.891 / 0.637**	**0.716 / 0.291**	**0.911 / 0.727**
Commitment timing only	0.729 / 0.370	0.722 / 0.311	0.735 / 0.405

V and LaViDA, whereas the latter transfers consistently across all three backbones. The final selector therefore uses evidence alone and contrasts the affirmative and negative verbalizer scores at the initial masked state through v_{init}. Both masked-state variants are more discriminative than decoded self-verification on every backbone. Pair normalization further improves over raw e^+ on both endpoints for all three backbones. Table 11 reports the decile boundaries, within-positive restriction, and sample sizes that underlie Figure 2(a) and Observation 1.

Evidence across verification states. Figure 5 traces the continuous evidence score v_t over the same 16-step verification trajectories; the hollow diamond at D instead reports the ROC-AUC of the terminal decoded answer and is not an additional timestep. The masked-state score is largely insensitive to which masked timestep is read. LLaDA-V declines from 0.891 to 0.749 as its masked fraction collapses, before the decoded decision falls to 0.555. The MMaDA aggregate curve remains near 0.70 because it mixes states: at the final step, already-committed trajectories yield 0.547, whereas still-masked trajectories retain 0.670. LaViDA remains masked in 92% of trajectories through step 15, so its flat $0.911 \rightarrow 0.907$ curve establishes temporal stability while masked; its decoded decision is lower at 0.742. Thus, this diagnostic does not claim that one early timestep is uniquely informative. It shows that graded evidence is stable while available in a masked state and is more discriminative than the terminal discrete decision.

Table 11: Full evidence diagnostics on the same extracted COCO probe pool as Figure 2. The lowest and highest decile unsupported rates bracket the downward trend in panel (a). The within-positive column restricts evaluation to commitments with $v_{\text{init}} > 0.5$.

			Decile rate		Within-positive	
Model	n	Unsup.	Low	High	%	ROC-AUC
LLaDA-V	872	123	75%	2%	96.6	0.878
MMaDA	356	48	37%	6%	96.1	0.695
LaViDA	704	138	81%	0%	87.2	0.871

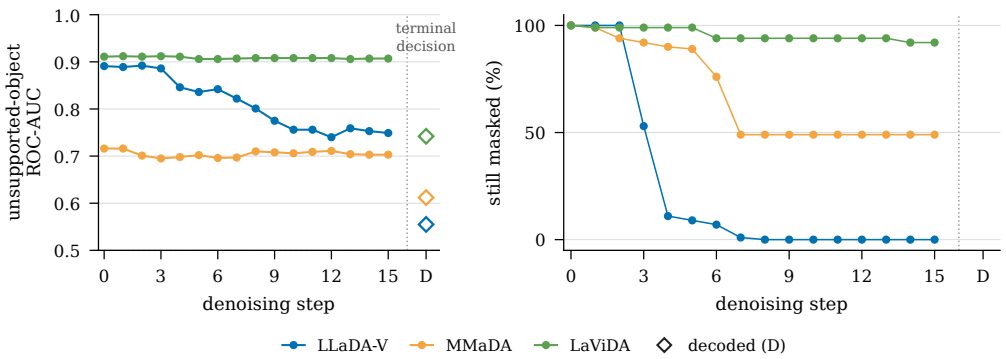

Figure 5: **ROC-AUC of v_t across the verification trajectory.** Left: detection AUC at each denoising step for all three backbones. Right: fraction of trajectories whose answer position is still masked. On LLaDA-V, readout AUC declines as the answer position becomes committed. LaViDA's readout remains stable while most answer positions are masked; on MMaDA, the still-masked subset has higher final-step AUC than the already-committed subset.

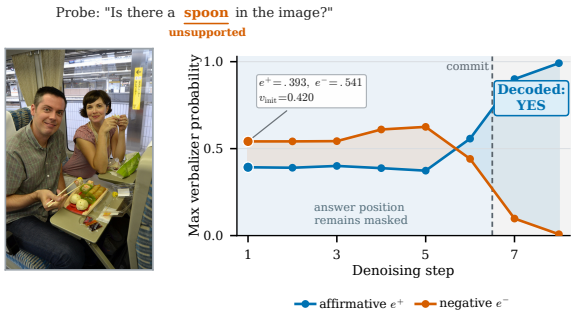

Figure 6: **An individual trajectory (LLaDA-V).** For the unsupported *spoon*, the initial state has $e^- > e^+$; during denoising, e^+ overtakes e^- before the answer commits to *yes*.

Table 12: Worked example from the COCO. All commitments decode to the same answer, while v_{init} ranks the unsupported commitment last.

Commitment	Decoded answer	v_{init}	In image
person	yes	0.98	yes
train	yes	0.91	yes
spoon	yes	0.42	no

A case study of an LLaDA-V probe trajectory. Table 12 and Figure 6 report the probe scores and denoising trajectory for one LLaDA-V response. Although all three object probes decode to *yes*, the initial masked-state score ranks the unsupported *spoon* below the supported *person* and *train*. In this case, e^+ overtakes e^- before the spoon probe commits to *yes*.

This individual trajectory illustrates that the initial score can favor a negative answer while the final decoded answer is affirmative. Separately, the decile ordering and within-positive AUC in Table 11 demonstrate the continuous score's usefulness for ranking unsupported objects across the evaluated commitments.

Initial-state scores and decoded answers. The decoded probe answer is affirmative for 98%, 84%, and 86% of all generated commitments on the three backbones, and LLaDA-V emits only 17 *no* answers in 872 probes. The displacement is bidirectional: 76% of LLaDA-V commitments that the initial state weakly rejects ($v_{\text{init}} \in [0.3, 0.5]$) are answered *yes*, while 63% of MMaDA's mid-grade commitments ($v_{\text{init}} \in [0.5, 0.8]$) are answered *no*.

Prompt robustness. Across *is there*, *contain*, and *visible* paraphrases, raw masked-state evidence rankings have Spearman correlations of approximately 0.895–0.991. Pair normalization places selection and retention on one support-versus-contradiction scale across tokenizers and backbones; it is not an additional learned calibrator.

Shared-state control. To test whether candidate retention benefits from a later masked state, we compared evaluating both sides of Equation 6 at the fully masked initial state versus at the last pre-commitment masked state on the development split. The two readouts produce the same accept/reject decisions on LLaDA-V and MMaDA and differ on at most one LaViDA transition. This agreement supports the shared initial masked state as a common evidence coordinate. We use it throughout the final method to make source and candidate evidence state-matched while eliminating probe denoising trajectories.

Selector definitions. All selector variants in Table 3 operate on the same commitment pool with the local transition and retention rule held fixed. The pair-normalized row uses the exact frozen EGCC selection rule from Section 4.2, yielding 35, 21, and 92 selected commitments for LLaDA-V, MMaDA, and LaViDA. These counts are outcomes of threshold-based selection, not imposed budgets for EGCC. For the alternative selectors, we retain the first commitments in each induced ordering up to the corresponding EGCC count to match intervention volume. Ties for these alternatives are broken by commitment order in the response. The selection rules are:

- **Decoded self-verification**: the probe trajectory is decoded to a yes/no answer; commitments not decoded as *yes* are ranked before those decoded as *yes*. Outputs containing neither keyword follow the fixed affirmative fallback, which conservatively avoids prioritizing an unparsed answer for intervention.
- **Raw masked evidence (e^+_{init})**: the raw affirmative masked-state score from Equation 2; lower e^+_{init} is ranked first.
- **Pair-normalized (v_{init}, ours)**: select every commitment satisfying $v_{\text{init}}(o_c \mid I) \leq \eta$ and, for responses with $N_y > 1$, $\delta_y(c) > \tau$, using the frozen thresholds $(\eta, \tau) = (0.60, 0.50)$ and the degenerate-case handling in Appendix A. For $N_y = 1$, only the absolute support condition applies. This is the final EGCC selector, with no top-K truncation.
- **Low confidence**: final-precommit confidence, measured immediately before each token is finalized and averaged across the subword positions realizing the commitment; lower mean confidence is ranked first.
- **Denoising entropy**: vocabulary-wide Shannon entropy averaged over the denoising trajectory and across the subword positions realizing the commitment; higher mean entropy is ranked first.

Both signals are derived from generation-time logits. For MMaDA, entropy is restricted to the language-model vocabulary (excluding image tokens), matching the model's native output space.

D Transition Analysis

All transition-component analyses use the commitments selected by the final policy (35/21/92 commitments for LLaDA-V/MMaDA/LaViDA). Each comparison changes only the indicated transition component while keeping selection, reopening scope, and acceptance fixed. Jobs count response-level joint transitions; thus LaViDA's 92 commitments correspond to 83 jobs.

Table 13: Source-state transition outcomes on the final-policy intervention set used in Figure 3, with generic remasking (γ=0) and source-state exclusion (γ=8). "Source retained" combines exact restoration and source-preserving rewrites.

Model	γ	Jobs	Restore	Rewrite	Deletion	Replaced
LLaDA-V	0	35	27	3	5	0
	8	35	6	0	28	1
MMaDA	0	21	13	5	1	2
	8	21	0	0	13	8
LaViDA	0	83	62	4	15	2
	8	83	11	1	49	22

Source retention under plain remasking (γ=0) is 85.7%, 85.7%, and 79.5% on the three backbones; with exclusion (γ=8) it drops to 17.1%, 0.0%, and 14.5% respectively. Table 13 provides the exact per-model outcome counts under the frozen main policy. This is the numeric basis for the 80–86% and 0–17% ranges reported in Observation 2.

Reopening-scope source table. Table 14 reports the exact numeric basis for Figure 4. Each scope varies only the reopened token range while selection, exclusion, and retention remain fixed on the frozen main-policy selection.

Table 14: Reopening-scope ablation on the frozen main-policy set selected with $\eta = 0.60$. Net hall. removed and net supported loss are relative to exact vanilla. Non-target preservation is the fraction of original tokens outside the local target spans selected by EGCC that remain unchanged, using the same reference positions across local, suffix, and full-response reopening. Local preservation is 100% by construction.

Model	Scope	CHAIR-S/CHAIR-I	Hall rem.	Supp. loss	Non-target
LLaDA-V	Local	21.60/5.71	19	4	100%
	Suffix	22.40/5.87	15	14	59%
	Full	23.00/5.97	12	7	47%
MMaDA	Local	9.00/7.03	7	5	100%
	Suffix	8.80/6.88	8	4	77%
	Full	8.80/6.89	8	5	76%
LaViDA	Local	18.80/7.02	32	9	100%
	Suffix	18.00/6.82	41	60	66%
	Full	17.80/6.91	39	65	55%

Table 15: Suppression strength on the frozen main-policy set selected with $\eta = 0.60$ (35/21/92 selected commitments for LLaDA-V/MMaDA/LaViDA) with pair-normalized retention fixed. CHAIR values use the same official instance-and-reference support protocol as the main table and are percentages. Both net columns are relative to exact vanilla; net supported added is supported-object additions minus removals.

Model/γ	CHAIR-S/ CHAIR-I	Net hall. removed	Net supp. added
LLaDA-V/0	23.80/6.20	5	0
LLaDA-V/4	22.80/5.96	12	-3
LLaDA-V/8	21.60/5.71	19	-4
LLaDA-V/12	21.20/5.64	21	-7
MMaDA/0	9.80/7.83	1	-3
MMaDA/4	9.00/7.01	7	-4
MMaDA/8	9.00/7.03	7	-5
MMaDA/12	9.00/7.03	7	-5
LaViDA/0	21.20/7.83	10	-5
LaViDA/4	19.00/7.23	26	-5
LaViDA/8	18.80/7.02	32	-9
LaViDA/12	18.80/6.91	35	-12

$\gamma = 12$ further lowers mention-level hallucination on LLaDA-V and LaViDA but removes more supported commitments; its sentence-level score saturates on LaViDA. The development-frozen $\gamma = 8$ is the middle point on the factuality–preservation frontier and already saturates on MMaDA.

E CONTROLLED BASELINE IMPLEMENTATION

VISAGE. VISAGE is the closest diffusion-native pre-commitment baseline, but no released implementation covers our open-ended three-backbone protocol. We therefore implement its commitment ranking directly from the published formulation. For each still-masked text position, we extract final-layer attention to visual tokens, normalize it over image positions, compute spatial Shannon

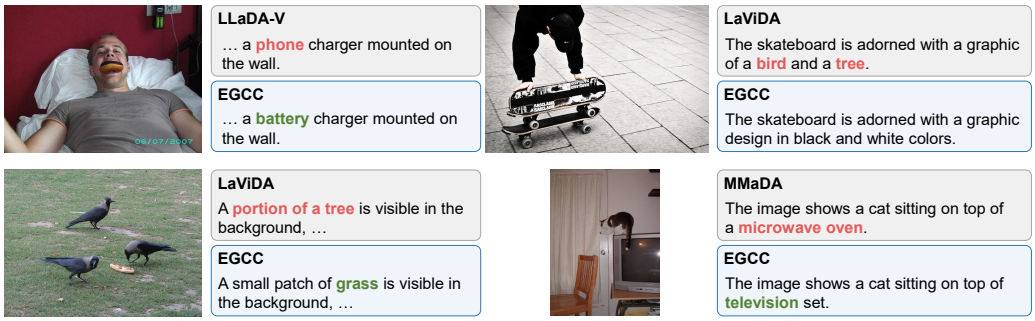

Figure 7: **Representative corrections across three dVLM baselines.** Qualitative comparison on COCO and AMBER-Generative benchmarks. Colored phrases mark unsupported commitments (red) and visually grounded alternatives (green).

entropy per head, and use the $\beta = 0.25$ lower-entropy head quantile as the robust visual entropy H_i. Native confidence c_i is ranked as $u_i = c_i(1 + H_i)^{-\alpha}$; token proposals, commitment counts, prompts, and native denoising schedules remain unchanged. We use $\alpha = 0.5$ for LLaDA-V and $\alpha = 0.1$ for MMaDA and LaViDA, as fixed in the disjoint development protocol. Setting $\alpha = 0$ exactly recovers native confidence ordering. MMaDA is the architecture studied by VISAGE; the LLaDA-V and LaViDA rows are architecture-faithful task transfers and are identified as such rather than as reproductions of published CHAIR results.

MPD-PAC and Information-Gain. For MPD-PAC, we use the public implementations and release settings for LLaDA-V and LaViDA. Although the original paper reports MMaDA results, the public release used in our experiments does not include an MMaDA-specific integration. Our adaptation produced lower CHAIR scores alongside severely degraded captions, preventing us from validating a faithful reproduction under our captioning protocol. We therefore mark this entry as unavailable rather than treating these scores as a validated baseline result.

For Information-Gain, we use the public implementation with position temperature 0.1 and confidence bypass 0.8; MMaDA scoring is restricted to its language vocabulary. All baseline scores in Table 1 use the same image manifest, model-native generation configuration, and official CHAIR evaluator as EGCC. On LaViDA, Information Gain's lower CHAIR coincides with a 2.93-point object-recall decrease and 2.32 fewer mean words, whereas EGCC preserves both quantities within 0.25 points and 0.01 words, respectively.

F Qualitative Corrections

Figure 7 presents representative paired examples across the three dVLM families. In these cases, EGCC replaces an unsupported commitment with an image-consistent alternative while preserving the surrounding description.

G Content Preservation

Table 16 reports the raw COCO object-mention accounting underlying the main-table CHAIR values, and Table 17 reports preservation statistics for the same outputs. COCO object recall uses the official CHAIR object mapping, and AMBER Cover uses the benchmark's reference-object coverage.

Across both benchmarks, the applied corrections leave mean response length unchanged and keep object recall or reference-object coverage within 0.52 percentage points of vanilla.

Acceptor transition flow. Table 18 follows one consistent accounting chain on AMBER-Generative under the frozen policy. A reopening job is a selected response-level intervention. A changed candidate differs from the vanilla response, and an applied edit is a changed candidate retained by the acceptor. Reducing, neutral, and harmful are then measured by the official evaluator according to whether the applied edit decreases, preserves, or increases the number of unsupported mentions.

Table 16: Raw COCO object-mention accounting under the frozen policy, paired against exact vanilla responses.

	LLaDA-V	MMaDA	LaViDA
vanilla			
Object mentions	2,634	681	2,506
Unsupported	168	54	205
Responses with hallucination	124	49	113
EGCC			
Object mentions	2,611	669	2,465
Unsupported	149	47	173
Responses with hallucination	108	45	94

Table 17: Content preservation under the frozen main policy. Values are percentage-point changes except mean words, which is an absolute change.

Model	COCO Δ words	COCO Δ recall	AMBER Δ Cover
LLaDA-V	0.00	-0.12	$+0.02$
MMaDA	0.00	-0.31	-0.52
LaViDA	0.00	-0.25	-0.23

Of the 781 reopening jobs, 150 reconstruct the vanilla response exactly and therefore do not count as edits. The acceptor rejects 76 of the 631 changed candidates: 50 for commitment-count expansion, 13 for failure to reallocate a selected source, 10 for non-monotonic candidate evidence, and three for duplicate expansion. Under the official evaluator, the 10 candidates rejected for non-monotonic evidence comprise three harmful, two hallucination-reducing, and five neutral edits. The remaining 555 applied edits give exactly the evaluator-level on-arm counts used in Table 5.

H Proposal-Vocabulary Boundary Analysis

Among the 555 applied edits in Table 18, 19 increase evaluator-visible unsupported mentions. Of these, 18 introduce an evaluator-visible object outside the internal proposal vocabulary; the remaining case is a tracked one-for-one replacement whose candidate evidence passes the gate. Thus, the residual harmful edits are concentrated at the boundary between the proposal vocabulary and the evaluator lexicon. Appendix I separately evaluates the frozen policy with a generic noun-phrase proposal-interface variant.

I Extension to Open-Vocabulary Object Proposals

We evaluate whether the final correction policy extends to generic noun-phrase proposals on the deterministic OpenCHAIR random-300 subset. We retain the pair-normalized masked-state evidence, development-frozen selection thresholds, source-state exclusion, and pair-evidence acceptor. Evaluation uses the authors' official concrete-noun extraction, prompt, answer parser, and Llama-2-70B judge. OpenCHAIR references and evaluator outputs are accessed only after correction.

The improvement direction holds for all three backbones, with a maximum change of approximately 1.9% in extracted object occurrences. This extension result supports compatibility with an open-vocabulary proposal interface with limited changes in object counts.

J Runtime Accounting

Table 6 reports a conservative measured upper bound on the paired COCO random-500 set. The timed implementation includes proposal extraction, one initial-state forward per distinct normalized source phrase within each image, sparse source and candidate probe trajectories, local recommitment, and acceptance. In contrast, the final policy in Algorithm 1 reuses source initial-state scores. After

Table 18: Acceptor transition flow on AMBER-Generative under the frozen policy with $\eta = 0.60$. Outcome columns use the official evaluator and sum to the number of applied edits.

Model	Jobs	Changed	Applied	Reducing	Neutral	Harmful
LLaDA-V	98	76	69	32	33	4
MMaDA	149	133	100	19	72	9
LaViDA	534	422	386	188	192	6
All	781	631	555	239	297	19

Table 19: Official OpenCHAIR random-300 proposal-interface expansion under the frozen pair-normalized policy. The final columns report the number of concrete-object occurrences extracted by the official evaluator before and after correction.

Model	Vanilla	EGCC	Differences	Objects (V)	Objects (EGCC)
LLaDA-V	0.3506	0.3404	-0.0103	4,039	3,962
MMaDA	0.2979	0.2937	-0.0042	624	622
LaViDA	0.3424	0.3365	-0.0059	3,725	3,658

the structural checks, each distinct object newly introduced by a candidate requires one initial-state forward before the evidence decision; no probe-denoising trajectory is decoded. Table 20 reports the initial-state forward counts per backbone. Model loading and disk I/O are excluded.

Forward-count accounting. Each initial-state forward is a single masked-state pass (no denoising trajectory). On the COCO random-500 set, source scoring requires one forward per distinct normalized phrase within each image, with the score reused across repeated commitments. Each distinct object introduced by a structurally admissible candidate that reaches the evidence gate requires one additional forward, whether the candidate subsequently passes or fails the evidence condition.

Table 20: Initial-state forward counts on the COCO random-500 test split. Source forwards equal the per-image deduplicated phrase count. In these runs, candidate forwards, including evaluations of subsequently rejected candidates, do not exceed the corresponding numbers of evidence-eligible source commitments.

Model	Source forward	Candidate forward	Total/image
LLaDA-V	872	$\leq$35	$\leq$1.9
MMaDA	356	$\leq$21	$\leq$0.8
LaViDA	704	$\leq$92	$\leq$1.6