ROLLOUT-VISUAL-COMMIT: MITIGATING OBJECT HALLUCINATION VIA COMMIT-TIME MULTIMODAL SCORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision–language models (LVLMs) answer complex visual questions fluently, yet still invent absent objects, deny present ones, or extend captions after an unsupported token. Fine-tuning remedies require extra supervision, current-step interventions act before a candidate has served as a query, and trajectory or reward methods add latency. All leave the multimodal state induced by a provisional next token unused. We propose Rollout-Visual-Commit (RVC), a decoding rule without LVLM post-training. RVC rolls top-K candidates forward in one branch-isolated packed pass, scores each candidate-induced state by probe-selected visual-attention mass times next-state logit margin, adds the standardized response to its base log-probability, and keeps only the winning KV entry. Thus, an unsupported but fluent token can be rejected before entering the prefix. On nine POPE settings with LLaVA-1.5 and Qwen2.5-VL, RVC attains the best Accuracy and F1 among the included methods (average per-setting margins $2.57/2.66$ and $1.44/2.01$). It also leads every reported CHAIR and AMBER metric, including AMBER coverage.

1 INTRODUCTION

Recent advances in large vision–language models (LVLMs) have shown strong potential on complex real-world multimodal tasks (Liu et al., 2024b; Bai et al., 2025). However, LVLMs frequently generate text that is inconsistent with the visual input: they invent absent objects, deny present ones, or misstate attributes such as color and count (Gunjal et al., 2024; Wang et al., 2023b; Huang et al., 2025; Li et al., 2023). This issue is commonly referred to as object hallucination (Rohrbach et al., 2018; Wang et al., 2023a), and it limits the use of LVLMs in settings that demand precise visual faithfulness.

Previous work points to two related causes. First, language priors are often stronger than visual evidence, so the model over-relies on textual cues (Lee et al., 2025; Tong et al., 2024). Second, for autoregressive decoding, *when* an error enters the generation matters: once an unsupported but locally fluent token is written into the prefix, later steps are conditioned on it and the mistake can snowball (Zhong et al., 2024). Mitigating object hallucination therefore depends on what the encoder represents and on how the decoder commits each next token.

Prior efforts address these issues along three main lines. (1) Fine-tuning with curated or preference data improves visual alignment (Sun et al., 2024; Yu et al., 2024b; Liu et al., 2024a; Rafailov et al., 2023; Chen et al., 2026), but needs extra annotation and updates the LVLM. (2) Current-step interventions reshape the forward pass or the distribution p_t by editing inputs, layers, or image-sensitive heads (Leng et al., 2024; Wang et al., 2024; Chuang et al., 2024; Kim et al., 2024; Chen et al., 2025; An et al., 2025; Huo et al., 2025); they act before any next-token candidate has served as a query, so the attention pattern and next-state margin induced by that candidate remain unobserved. (3) Trajectory-aware methods search, roll back, or score multi-token continuations with an external reward model (Huang et al., 2024; Chen et al., 2024b; Mañas et al., 2025); they can inspect longer futures, but multi-token expansion is slow, rollback discards already computed paths, and both waste decoding time. A remaining opportunity is to use one provisional write inside the frozen generator as the correction signal itself (Figure 1).

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.

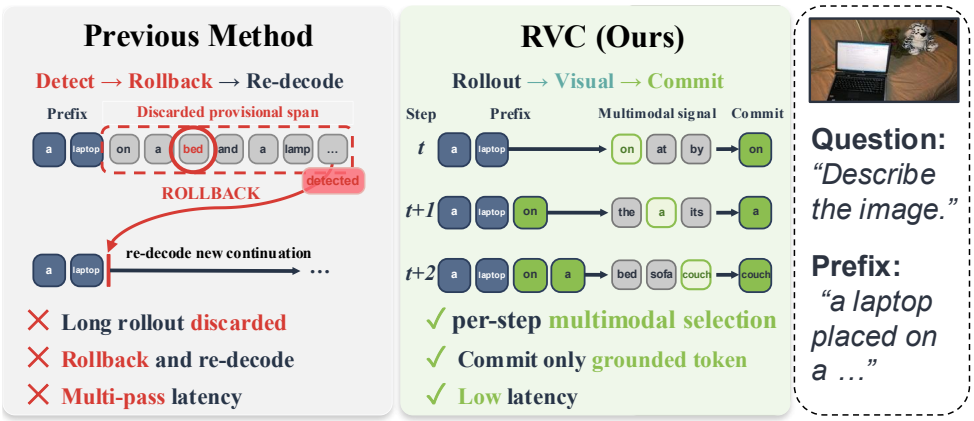


Figure 1: Why commit-time visual selection avoids rollback. Previous trajectory-aware methods detect errors only after a long rollout, discard the provisional span, and re-decode from the retained prefix. At each step, RVC applies the candidate-induced multimodal signal before commitment, keeps only the grounded token, and proceeds without rollback.

We propose Rollout-Visual-Commit (RVC) for that opportunity (Figures 1 and 2). After a candidate is provisionally written, RVC reads the candidate-induced multimodal state and decides whether to commit it. At each step, RVC *rolls out* the top-K candidates under a shared multimodal prefix with a branch-isolated mask, reads a *visual* response from each candidate-conditioned state (probe-selected attention mass times next-state logit margin), and *commits* the winning KV entry. The standardized response is added to $\log p_t$, so the base LVLM prior remains the anchor. The procedure uses no LVLM post-training and no auxiliary reward model, and it performs one provisional transition per candidate.

On LLaVA-1.5 and Qwen2.5-VL, RVC attains the best Accuracy and F1 among the included methods on all nine POPE settings (Li et al., 2023) (average per-setting margins $2.57/2.66$ and $1.44/2.01$), and leads every reported CHAIR (Rohrbach et al., 2018) and AMBER (Wang et al., 2023a) metric while raising AMBER coverage. Our contributions are:

- A commit-time selection rule: after a provisional write, score the candidate-induced multimodal state, then commit only the winning token.
- A one-step packed instantiation without LVLM post-training: a branch-isolated rollout whose response (visual-attention mass times next-state logit margin) is standardized within the candidate set and added to the base log-probability.
- Stronger binary and open-ended faithfulness on LLaVA-1.5 and Qwen2.5-VL: best Accuracy and F1 on all nine POPE settings (average per-setting margins $2.57/2.66$ and $1.44/2.01$), and best reported CHAIR and AMBER scores with the highest AMBER coverage among compared methods.

2 Related Work

2.1 Object Hallucination in LVLMs

Modern LVLMs such as LLaVA and Qwen2.5-VL map visual features into a language-model stack (Liu et al., 2024b; Bai et al., 2025), yet answers can be dominated by language priors when visual evidence is weak (Lee et al., 2025; Tong et al., 2024). The failure is not confined to rare edge cases: object presence, attributes, and co-occurrence statistics can all pull decoding away from the image, especially once an early unsupported token enters the prefix. POPE, CHAIR, AMBER, and related protocols quantify these object-level failures under binary probing and open-ended captioning (Li et al., 2023; Rohrbach et al., 2018; Wang et al., 2023a; Kaul et al., 2024). Complementary analyses show that attention heads specialize unevenly in general (Voita et al., 2019; Clark et al., 2019), and that visual information in LVLMs concentrates in a subset of layers and heads (Yin et al., 2025; Kaduri et al., 2025; Chen et al., 2024a), which motivates reading attention geometry as evidence.

2.2 MITIGATING HALLUCINATIONS IN LVLMS

Existing mitigations fall mainly into training-time and inference-time families. Training-time methods introduce curated or preference data, stronger factual objectives, or related alignment recipes (Sun et al., 2024; Yu et al., 2024a;b; Liu et al., 2024a; Zhang et al., 2024; Rafailov et al., 2023; Chen et al., 2026). They can reduce hallucination, but require additional optimization and leave the decoder's commit rule unchanged: at step t, the model still chooses from the current distribution p_t before any candidate has served as a query.

Inference-time methods keep the LVLM frozen and differ in *when* the correction signal is read. Current-step approaches reshape p_t with contrastive decoding, layer contrasts, self-generated references, or head-level edits (Leng et al., 2024; Wang et al., 2024; Chuang et al., 2024; Kim et al., 2024; Liu et al., 2025; Huo et al., 2025; Zhu et al., 2025; Liu et al., 2024c; Woo et al., 2025; Favero et al., 2024; An et al., 2025; Chen et al., 2025); they intervene before a next-token candidate is written, so the multimodal states induced by alternative candidates remain unobserved. Trajectory-aware methods search, roll back, or reject and resample under retrospection or saliency checks (Huang et al., 2024; Chen et al., 2024b; Zhang et al., 2026); reward-guided decoding further scores multi-token continuations with an external multimodal reward model (Mañas et al., 2025). Multi-token expansion is slow, and rollback or resampling discards already computed paths, so both lines waste decoding time; a separate reward model adds further overhead. RVC stays in the inference-time family but moves the correction signal to commit time: one provisional transition exposes the generator's visual-attention response and next-state margin, with no auxiliary reward model and no emitted-token rollback.

3 METHOD

Overview. Let an image, a text prompt, and an already committed token prefix $y_{<t} = (y_1, \ldots, y_{t-1})$ be given at decoding step $t \geq 1$. An autoregressive LVLM produces vocabulary logits $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ and the next-token distribution $p_t = \mathrm{softmax}(\mathbf{z}_t)$, then selects $y_t \in \mathcal{V}$. Greedy or sampling decisions use only p_t, which ranks tokens before any candidate has been incorporated as a query. The resulting commitment is therefore made without observing how each option would reconfigure the multimodal state. RVC changes only this selection rule. Concretely, it forms a top-K candidate set from p_t ($K \geq 2$), shares the cached multimodal key–value (KV) context across those candidates, advances each candidate by one temporary step under a branch-isolated attention mask, reads a two-factor response from every candidate-conditioned state, and commits the highest-scoring candidate while discarding the others. We first construct comparable one-step states (Section 3.1), then define the response (Section 3.2), and fuse it with $\log p_t$ and update the KV cache (Section 3.3). Figure 2 summarizes the pipeline.

The design follows three constraints. First, scoring must stay inside the frozen generator: no second image encoding, no trained reward head, and no LVLM post-training. Second, the horizon is exactly one provisional transition, so the cost remains a single packed forward. Third, $\log p_t$ must remain the anchor, so the response can overturn only locally ambiguous decisions. Under these constraints, commit-time scoring contrasts with multi-token search, rollback, and reward-model scoring of full continuations.

3.1 ONE-STEP CANDIDATE ROLLOUT

At step t, $\mathbf{z}_t$ ranks vocabulary tokens before any candidate has served as a query in the prefix, so candidate-conditioned attention and next-state logits are unavailable. Heuristics that reweight p_t using only the present-step attention or logit geometry leave unobserved how the multimodal state would react once a specific token is written into the cache. We form a candidate set $\mathcal{C}_t = \{c_1, \ldots, c_K\} \subset \mathcal{V}$ as the top-K tokens under p_t, without an external proposal sequence. Restricting proposals to this support keeps the search local to tokens the LVLM already considers plausible and avoids retrieving a separate response hypothesis. The goal is to discriminate among tokens already near the decision boundary.

To compare candidates under an identical history, we process each $c_j \in \mathcal{C}_t$ as if it were appended, while preventing cross-candidate interference. Isolation is essential: if temporary candidates could

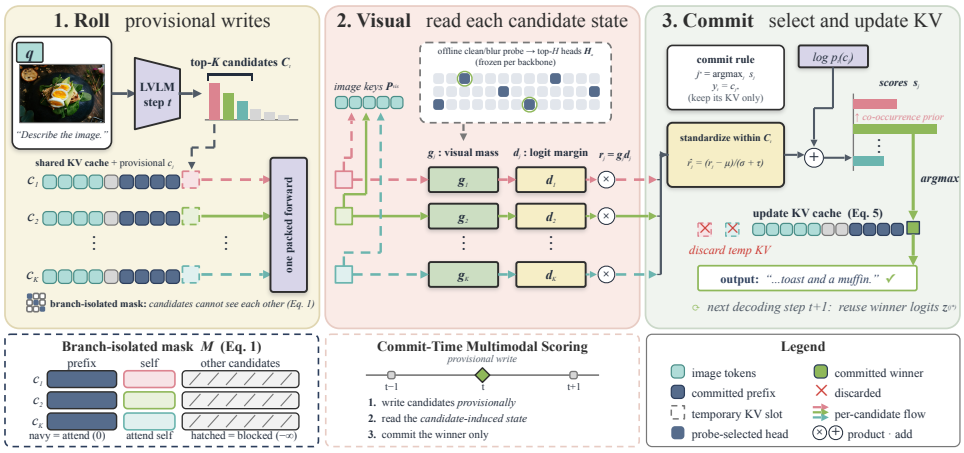


Figure 2: RVC pipeline. (1) Rollout: top-K provisional writes under a shared KV cache in one packed forward. (2) Visual: read probe-selected visual mass g_j and next-state logit margin d_j as $r_j{=}g_j d_j$. (3) Commit: score $s_j{=}\log p_t(c_j) + \widehat{r}_j$ and keep only the winning KV entry.

attend to one another, their induced attention and logits would no longer be counterfactuals under the same prefix. Let S_t denote the number of cached multimodal tokens in the committed prefix (image tokens, prompt tokens, and $y_{<t}$). Candidate c_j occupies temporary query position $q_j = S_t + j$. The packed queries are indexed by candidate $j \in \{1, \dots, K\}$, and keys by sequence positions $q \in \{1, \dots, S_t + K\}$. Define the attention mask entry $M^{(t)}_{j,q} \in \{0, -\infty\}$ by

$$M^{(t)}_{j,q} = \begin{cases} 0, & q \le S_t \text{ or } q = q_j, \\ -\infty, & \text{otherwise}, \end{cases} \tag{1}$$

so query j attends to the committed prefix keys and to its own key at q_j, while the other $K-1$ temporary keys remain invisible. Here q_j is only the physical cache/mask slot; all K candidates share one logical next-position ID for rotary embeddings after the committed prefix. We evaluate them together in one packed forward under Eq. 1. Packing reuses the prefix KV once, so it is faster than K sequential forwards over the same cache. The forward returns, for each c_j, next-state logits $\mathbf{z}^{(j)}_{t+1} \in \mathbb{R}^{|\mathcal{V}|}$, per-head attention weights $\mathbf{A}^{(j)}$, and a temporary KV entry for position q_j. Here $A^{(j)}_{\ell h}[q_j, v]$ denotes the attention weight at layer ℓ and head h from query position q_j to key position v. After scoring, only the winning temporary KV is retained for the committed history.

3.2 Response Scoring

Let ℓ index transformer layers and h index attention heads within a layer, and let $\mathcal{P}^{\text{vis}}$ be the set of key positions that correspond to image tokens. Because head-level visual signals are useful but uneven across the full stack, we select a compact subset for reading visual mass. Offline, on $N{=}100$ COCO train image–prompt pairs with a fixed generic prompt (disjoint from POPE, CHAIR, and AMBER images), we compare each image x with a Gaussian-blurred copy $\tilde{x}$ (blur radius 5) and read visual attention mass at the last prompt query $\bar{q}$ before generation,

$$m_{\ell h}(x) = \sum_{v \in \mathcal{P}^{\text{vis}}} A_{\ell h}[\bar{q}, v], \qquad s_{\ell h} = \frac{1}{N} \sum_{i=1}^{N} \big(m_{\ell h}(x_i) - m_{\ell h}(\tilde{x}_i) \big), \tag{2}$$

and keep the H highest-scoring pairs $\mathcal{H}_a = \text{TopH}(s_{\ell h})$, $|\mathcal{H}_a| = H$. Heads that attend more under the clean image than under the blur receive higher $s_{\ell h}$. The ranking is a one-time cost of $2N$ forwards per backbone, then frozen and used only to choose readout heads for g_j. For candidate c_j, define the visual attention mass g_j, the next-state logit margin d_j, and the response r_j by

$$g_j = \frac{1}{|\mathcal{H}_a|} \sum_{(\ell,h) \in \mathcal{H}_a} \sum_{v \in \mathcal{P}^{\text{vis}}} A^{(j)}_{\ell h}[q_j, v], \quad d_j = z^{(j)}_{t+1,(1)} - z^{(j)}_{t+1,(2)}, \quad r_j = g_j d_j, \tag{3}$$

where $z^{(j)}_{t+1,(1)}$ and $z^{(j)}_{t+1,(2)}$ are the largest and second-largest entries of $\mathbf{z}^{(j)}_{t+1}$ over the full vocabulary $\mathcal{V}$.

3.3 Scoring and Commitment

Raw response magnitudes vary across prompts and decoding steps, so absolute thresholds on r_j are brittle. We instead standardize $\{r_j\}_{j=1}^{K}$ within the current candidate set $\mathcal{C}_t$. Let $\mu_r = \frac{1}{K}\sum_{j=1}^{K} r_j$ and let σ_r be the population standard deviation $\left(\frac{1}{K}\sum_{j=1}^{K}(r_j-\mu_r)^2\right)^{1/2}$. With a fixed numerical floor $\tau{=}10^{-3}$, the standardized response is

$$\widehat{r}_j = \frac{r_j - \mu_r}{\sigma_r + \tau}. \tag{4}$$

Within-set standardization hard-bounds the correction: $|\widehat{r}_j| \le \sqrt{K-1}$ (about 1.41 at $K{=}3$), so at unit scale the response cannot overwhelm $\log p_t$ by more than roughly 1.4 nats. Consequently, since the pairwise gap $\widehat{r}_j - \widehat{r}_{j'}$ is at most $\sqrt{2K}$ (about 2.4 nats at $K{=}3$, a probability ratio of $\approx 12\times$), a candidate trailing the greedy token by a larger $\log p_t$ margin can never be selected; this makes precise the sense in which RVC overturns only locally ambiguous decisions. The floor τ keeps the map continuous at $\sigma_r{=}0$: when responses agree, $\widehat{r}_j \to 0$ and selection falls back to $\log p_t$. Ablations that replace the product by g_j, d_j, or additive $r_j{=}g_j{+}d_j$ apply this same one-shot standardization to that scalar r_j before adding to $\log p_t$; g_j and d_j are standardized only after they form the scalar r_j, so raw scale differences are absorbed before fusion. Whether either factor dominates the within-set ranking is settled empirically in Table 4(a). The base log-probability is left unnormalized. The selection score s_j and the winning index j^* are

$$s_j = \log p_t(c_j) + \widehat{r}_j, \qquad j^* = \arg\max_{j\in\{1,\dots,K\}} s_j, \tag{5}$$

we add the within-set response at unit scale (implicit $\lambda{=}1$), which is commensurate with the $\sqrt{K-1}$ bound above. Turning the response off recovers $\arg\max_j \log p_t(c_j)$ inside $\mathcal{C}_t$ (greedy whenever the greedy token is in $\mathcal{C}_t$). A non-greedy token wins only when the relative response overcomes the current $\log p_t$ gap. Scoring uses the generator's one-step state and keeps $\log p_t$ as the anchor, in contrast to reward-model reranking of full continuations (Mañas et al., 2025).

After choosing c_{j^*}, we set $y_t = c_{j^*}$, retain only its temporary KV entry, and reuse $\mathbf{z}^{(j^*)}_{t+1}$ as the logits for step $t{+}1$; the other $K-1$ entries are discarded so hypothetical candidates never enter the persistent cache. This commit rule keeps lookahead safe for autoregressive generation: branches are evaluated, and only one state is written forward. Because the discarded branches leave the prefix clean, later steps remain ordinary LVLM decoding conditioned on a single committed history. Each step therefore uses one packed evaluation that amortizes shared-prefix compute across the K temporary positions and attention tensors. White-box access to the attention mask, per-head attention, and KV cache is required; black-box APIs that expose only next-token probabilities are outside the present scope.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. POPE (Li et al., 2023) is designed to assess how effectively LVLMs recognize the presence of specific objects, and thus exposes object-level hallucination under binary probing. It asks Yes/No questions derived from object annotations and reports Accuracy, Precision, Recall, and F1; following common practice we emphasize Accuracy and F1. The dataset is balanced (50% existing / 50% non-existing objects) and uses three sampling strategies (Random, Popular, and Adversarial) that progressively strengthen frequency and co-occurrence priors. Drawing from MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019), POPE comprises $27{,}000$ query–answer pairs across nine settings; we evaluate all nine.

CHAIR (Rohrbach et al., 2018) measures unsupported object mentions in free-form image captions. We report sentence-level C_S and instance-level C_I at maximum generation lengths 64 and 512 on a

random sample of 500 images from the COCO 2014 *validation* set. The short horizon stresses early object commitments, while the long horizon allows unsupported nouns to accumulate later in the caption.

AMBER (Wang et al., 2023a) evaluates generative faithfulness at multiple granularities. We report object-mention CHAIR, descriptive Cover, response-level Hal, and cognitive Cog. Together with CHAIR, AMBER tests whether reduced hallucination comes at the cost of omitting annotated objects.

Baselines. We adopt the widely used LLaVA-1.5-7B (Liu et al., 2024b) and Qwen2.5-VL-7B-Instruct (Bai et al., 2025) as backbone LVLMs; tables refer to them as LLaVA-1.5 and Qwen2.5-VL. We compare the unmodified base decoder (*Vanilla* in the tables) against representative inference-time hallucination mitigators spanning three families: (i) contrastive or layer-contrast decoding, including VCD (Leng et al., 2024), DoLa (Chuang et al., 2024), and SID (Huo et al., 2025); (ii) head- or attention-oriented edits, including ICT (Chen et al., 2025); and (iii) trajectory-aware search or rollback, including OPERA (Huang et al., 2024) and HALC (Chen et al., 2024b). Baseline numbers use the authors' public implementations under matched prompts, decoding lengths, and answer parsers whenever available.

Implementation Details. Unless otherwise stated, main tables use one-step candidate rollout with $K{=}3$, within-set response standardization (Eq. 4) with $\tau{=}10^{-3}$, and $H{=}32$ offline-probe heads for reading g_j (Eq. 2; COCO train pairs only, frozen once per backbone). These defaults are held fixed across POPE, CHAIR, and AMBER. Nearby widths remain effective in Table 4(b). POPE answers are decoded greedily up to 16 new tokens and parsed as Yes/No; CHAIR uses length limits 64/512 on the validation sample above; AMBER follows its default generative setting. All runs use `bfloat16` on a single NVIDIA A100 (80 GB). Ablations of K, latency, and the response factors appear in Section 5.1 and the subsequent analysis subsections.

4.2 Main Results

Results on POPE. Table 1 shows that RVC improves object-presence decisions on every POPE setting. Across both LLaVA-1.5 and Qwen2.5-VL, it records the best Accuracy and the best F1 in all nine settings. Relative to the base decoder, the nine-setting macro Accuracy/F1 rises from $79.83/79.29$ to $86.59/87.32$ on LLaVA-1.5 ($+6.76/+8.03$) and from $83.38/82.14$ to $88.38/88.18$ on Qwen2.5-VL ($+5.00/+6.04$). The same ordering holds against strong inference-time baselines. Averaged over settings, RVC still leads the strongest non-RVC competitor by 2.81 Accuracy / 2.72 F1 on LLaVA-1.5 (ICT) and by 1.72 / 2.70 on Qwen2.5-VL (ICT). Under Adversarial sampling, where co-occurrence priors are strongest, LLaVA-1.5 reaches Adversarial F1 of $87.21/82.32/80.76$ on COCO/A-OKVQA/GQA, and Qwen2.5-VL reaches $86.16/84.48/84.69$. Current-step contrast, self-contrast, and head edits (VCD, DoLa, SID, ICT) and search/rollback methods (OPERA, HALC) all trail RVC across these settings. This pattern matches the intended role of commit-time scoring: the decoder keeps $\log p_t$ as the language prior and uses the one-step visual response to resolve locally ambiguous Yes/No writes. Section 5.1 isolates that response directly.

Model	Method	COCO						A-OKVQA						GQA					
		Random		Popular		Adversarial		Random		Popular		Adversarial		Random		Popular		Adversarial	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Qwen2.5-VL	Vanilla	85.23	83.09	84.53	82.58	83.37	81.57	86.40	85.07	85.77	84.49	80.37	79.68	85.10	83.87	80.87	80.33	78.77	78.56
	DoLa	86.53	84.44	85.93	84.41	83.47	81.51	87.40	86.46	87.53	86.06	81.83	81.33	87.13	86.05	82.57	82.35	82.00	81.51
	OPERA	87.30	86.92	87.43	86.68	84.77	83.45	88.20	88.43	87.90	87.13	80.83	81.54	86.03	85.29	81.97	82.12	80.23	80.64
	HALC	85.73	83.65	84.13	82.30	84.50	82.69	86.13	84.82	86.57	85.23	79.83	79.34	86.07	84.76	80.67	80.10	79.43	79.19
	VCD	88.63	87.81	87.13	86.40	84.27	83.90	89.23	89.01	87.83	87.81	81.27	82.38	85.60	85.33	81.83	82.23	80.00	80.74
	ICT	87.53	85.84	86.77	84.93	86.17	84.32	88.97	87.90	87.43	86.39	83.60	83.02	88.97	87.89	86.43	85.47	84.10	83.53
	SID	87.90	86.73	86.90	86.00	83.83	83.02	88.50	87.93	87.63	87.41	80.83	81.51	84.87	84.25	81.60	81.83	79.57	79.88
	RVC (ours)	**90.50**	**89.62**	**89.17**	**88.54**	**86.67**	**86.16**	**92.33**	**92.18**	**89.73**	**89.79**	**83.83**	**84.48**	**91.63**	**91.32**	**87.27**	**86.81**	**84.33**	**84.69**
LLaVA-1.5	Vanilla	83.30	81.33	81.87	80.06	78.97	77.57	83.43	82.56	79.90	79.59	74.03	75.15	83.73	82.95	78.17	78.38	75.07	76.06
	DoLa	85.97	86.14	82.93	83.80	77.17	79.41	83.23	84.83	76.47	79.86	68.00	74.48	83.70	85.29	74.03	78.74	68.73	74.78
	OPERA	89.20	88.81	86.63	84.89	81.23	81.38	88.03	87.00	83.23	84.15	73.83	77.91	88.13	88.91	79.27	82.11	75.00	78.71
	HALC	84.00	81.97	81.40	79.62	80.17	78.48	83.10	82.34	80.80	80.25	73.43	74.82	84.27	83.45	77.93	78.24	76.07	76.88
	VCD	87.73	87.16	85.37	85.06	80.87	81.33	86.17	86.34	81.87	82.82	74.97	77.73	86.67	86.99	80.73	82.24	76.10	78.78
	ICT	89.17	88.48	86.07	86.40	83.30	82.54	89.00	88.71	83.40	84.45	75.57	78.69	89.23	89.34	80.87	82.64	77.40	80.11
	SID	88.03	87.34	85.83	85.44	81.40	81.54	86.47	86.52	82.47	83.21	75.90	77.93	86.97	87.17	82.07	82.95	76.63	78.99
	RVC (ours)	**90.83**	**90.37**	**90.00**	**89.62**	**87.33**	**87.21**	**91.83**	**91.87**	**87.67**	**88.22**	**80.17**	**82.32**	**91.50**	**91.63**	**82.17**	**83.91**	**77.80**	**80.76**

Table 1: POPE Accuracy and F1 (%) under Random / Popular / Adversarial negative sampling on COCO, A-OKVQA, and GQA. Bold / underline: best / second-best per backbone–setting; the RVC rows are highlighted.

Model	Method	CHAIR				AMBER			
		$C_S^{64}\downarrow$	$C_I^{64}\downarrow$	$C_S^{512}\downarrow$	$C_I^{512}\downarrow$	CHAIR↓	Cover↑	Hal↓	Cog↓
Qwen2.5-VL	Vanilla	20.8	17.8	35.5	24.5	5.3	<u>47.2</u>	23.6	1.9
	SID	18.1	15.6	33.0	22.8	4.9	46.3	21.5	1.4
	DoLa	20.1	17.3	35.9	24.6	5.5	46.8	25.1	2.2
	OPERA	18.6	15.9	32.9	22.8	5.1	46.2	21.6	1.7
	HALC	19.5	16.7	34.2	23.6	5.0	45.8	22.0	1.7
	VCD	19.3	16.7	33.9	23.3	<u>4.7</u>	45.9	<u>20.5</u>	1.6
	ICT	<u>17.0</u>	<u>14.5</u>	<u>31.9</u>	<u>22.1</u>	4.9	45.1	21.8	<u>1.2</u>
	RVC (ours)	**16.5**	**14.3**	**31.4**	**21.8**	**4.6**	**47.7**	**20.1**	**1.1**
LLaVA-1.5	Vanilla	31.5	20.3	50.6	30.4	7.3	50.7	33.7	3.8
	SID	26.0	17.2	46.2	27.6	6.1	50.2	27.5	2.8
	DoLa	29.1	21.5	52.4	30.2	7.6	<u>51.6</u>	36.0	4.0
	OPERA	26.7	17.9	46.9	27.8	7.3	49.6	32.0	3.5
	HALC	28.9	18.7	48.5	28.8	6.6	49.1	30.9	3.3
	VCD	28.3	19.0	48.7	28.9	6.8	49.6	30.4	3.5
	ICT	<u>24.2</u>	<u>15.9</u>	<u>44.5</u>	<u>26.6</u>	<u>5.2</u>	51.3	<u>23.1</u>	<u>2.1</u>
	RVC (ours)	**23.5**	**15.3**	**43.6**	**26.0**	**4.9**	**52.8**	**22.5**	**1.9**

Table 2: Open-ended results. Left: CHAIR C_S/C_I (%) at max lengths 64/512 on COCO 2014 val. Right: CHAIR, Cover, Hal, and Cog on AMBER (Cover ↑; others ↓). Bold / underline: best / second-best per backbone; RVC rows highlighted.

Results on CHAIR. Table 2 (left) shows that the same rule also lowers unsupported object mentions in free-form captions. On both backbones, RVC obtains the lowest sentence-level C_S and instance-level C_I at max length 64 and at 512. The short horizon (64) rewards cleaner early object commitments. The long horizon (512) still improves, so the benefit extends beyond the first few tokens of a caption. The strongest baseline here, ICT, reduces CHAIR relative to the base decoder, as do OPERA and HALC with matched multi-token search, yet all of them remain above RVC at every reported length. Commit-time response scoring therefore remains effective against both head-level contrast and matched multi-token search on open-ended generation.

Results on AMBER. Table 2 (right) shows that lower hallucination coexists with higher coverage. RVC is best on every reported AMBER metric for both backbones. Versus the base decoder, Cover rises from 47.2 to 47.7 on Qwen2.5-VL and from 50.7 to 52.8 on LLaVA-1.5, while object-mention CHAIR falls from 5.3 to 4.6 and from 7.3 to 4.9. Response-level Hal and cognitive Cog fall as well, from 23.6/1.9 to 20.1/1.1 on Qwen2.5-VL and from 33.7/3.8 to 22.5/1.9 on LLaVA-1.5. OPERA and HALC often cut hallucination proxies while losing Cover (e.g., Qwen2.5-VL Cover 46.2/45.8 vs. base 47.2). RVC improves both coverage and faithfulness. Together with the length-wise CHAIR gains, these results show that candidate-conditioned scoring transfers from binary POPE probes to open-ended captioning.

5 Analysis and Discussions

5.1 Impact of Response Factors and Head Selection

Table 3 is a within-step mechanism check on AMBER generative captions with Qwen2.5-VL-7B, not an end-to-end Hal metric. We log each *object-commit* step: the first decode step at which an object word begins, and record g, d, and $r{=}g{\cdot}d$ for every candidate in $\mathcal{C}_t$ (one image can yield several commits). Each string is labeled *safe* if it is a prefix of a ground-truth or association object, *halluc* if it is a prefix of an AMBER-vocabulary object absent from the ground truth, and *other* otherwise (dropped). The analysis unit is a same-step safe–hallucination pair. On 100 images this yields 197 pairs. Fusion ranks the safe token above the hallucinated one in 83.2% of pairs ($p{=}2.2{\times}10^{-22}$), above g (71.1%) and d (76.1%). Relative to g alone, fusion gains +0.121 in pairwise accuracy; when the two disagree on the winner, fusion helps on 31 pairs and hurts on 7. Thus $r{=}g{\cdot}d$ is the strongest separator of the object token that should be written from an unsupported object token competing in the same step.

Signal	P(safe>halluc)	Mean Δ	Sign-test p
Grounding (g)	0.711	+0.029	$1.5{\times}10^{-9}$
Margin (d)	0.761	+0.930	$4.9{\times}10^{-14}$
Fusion ($r{=}g{\cdot}d$)	0.832	+0.281	$2.2{\times}10^{-22}$

Table 3: Within-step ranking on AMBER-gen (Qwen2.5-VL-7B): safe vs. hallucination candidates in the same $\mathcal{C}_t$ (100 images, 197 pairs).

We first ask whether the product response and the offline-probe head set are necessary. Table 4 ablates both under a fixed scaffold (K=3) on Qwen2.5-VL. Panel (a) keeps $s_j = \log p_t + \widehat{r}_j$ and changes only r_j. Removing the response collapses to greedy. Using g_j or d_j alone raises F1 to 85.6 and 83.1. Additive fusion $g_j + d_j$ reaches 86.4. The product $g_j d_j$ is strongest at F1 88.2, with the best AMBER CHAIR and Cover. Empirically, requiring agreement between the two factors helps more than using either factor alone or adding them. We therefore keep the product as the main response.

(a) Response r_j	F1↑	CHAIR↓	Cover↑
Greedy	82.1	5.3	47.2
g_j only	85.6	5.0	47.3
d_j only	83.1	5.2	47.2
$g_j + d_j$ (additive)	86.4	4.8	47.5
$g_j d_j$ (RVC main)	**88.2**	**4.6**	**47.7**
(b) Head readout	F1↑	CHAIR↓	Cover↑
Zero probe heads	82.2	5.3	47.2
Probe heads H=16	87.5	4.8	47.6
Probe heads H=32 (default)	**88.2**	**4.6**	**47.7**
Probe heads H=64	88.0	4.7	47.6
Probe heads H=128	87.3	4.9	47.5
Random heads ($\lvert\mathcal{H}\rvert$=32)	84.1	5.1	47.3
All heads (784)	83.5	5.2	47.3

Table 4: Qwen2.5-VL ablations (K=3). Panel (a) varies r_j; panel (b) varies the head readout with $\mathcal{C}_t$ and $\log p_t$ fixed. F1 averages nine POPE settings; CHAIR/Cover are from AMBER.

Panel (b) freezes $\mathcal{C}_t$ and $\log p_t$ and varies only the head readout over 784 layer–head pairs. Zeroing probe-selected heads returns near-greedy performance. Random heads of size 32 reach only 84.1, and reading all 784 heads reaches 83.5, because non-visual heads dilute g_j. Probe top-H sets are consistently better across nearby widths. Section 5.2 locates the sensitive heads and returns to this ablation.

5.2 Analysis of Attention Heads for Visual Readout

Prior analyses show that attention heads specialize at different granularities (Voita et al., 2019; Clark et al., 2019); in LVLMs, head-level visual signals are correspondingly useful yet uneven and often noisy across the full stack (Yin et al., 2025; Kaduri et al., 2025). Averaging over all layer–head pairs therefore mixes informative heads with weak ones. Eq. 2 uses an offline clean/blur probe to rank heads and keep a Top-H subset for reading g_j.

Figure 3 visualizes probe ranking on Qwen2.5-VL-7B and LLaVA-1.5-7B. Darker cells indicate higher probe scores. Early layers stay weak on both backbones, while stronger responses concentrate in the middle and deep layers, with additional scattered peaks later in the stack. The heads used by RVC come from this more sensitive region of the stack.

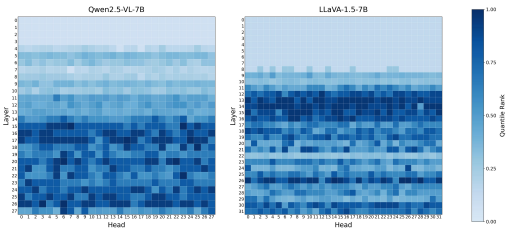


Figure 3: Heatmap of per-head offline clean/blur probe scores (Eq. 2). Left: Qwen2.5-VL-7B; right: LLaVA-1.5-7B. Darker cells indicate higher visual sensitivity.

Table 4(b) tests whether that filter matters on the decoding task. Under a frozen candidate set and $\log p_t$, zeroing the probe-selected heads, replacing them with random heads, or reading all heads loses most of the gain. Using the probe Top-H subset recovers the improvement. The ablation therefore shows that reading g_j from visually sensitive heads outperforms a random or full-stack average.

5.3 Width-Matched Quality and Decoding Cost

A natural concern is that RVC wins only because it inspects more candidates than greedy decoding. Table 5 therefore matches width W=4 for beam search, OPERA (Huang et al., 2024), and HALC (Chen et al., 2024b), and reports wall-clock captioning cost on Qwen2.5-VL under matched hardware (batch size 1, warm-up discarded, repeated timed runs; one fresh process per configuration). At this matched width, RVC (K=4) reaches the best AMBER CHAIR (4.5) and Cover (47.9) with POPE F1 (88.5) within 0.3 of the K=3 default, while running at only 1.30× greedy; beam, OPERA, and HALC cost 3.77×, 6.59×, and 7.70×, yet none close the quality gap. VCD pays 1.75× for its extra distorted-image forward. RVC instead packs K temporary candidate positions over the shared prefix in one forward: K=2 already runs at 1.13×, and peak memory rises monotonically with K yet stays within about 0.6% of the baseline. Sweeping K shows that most POPE

F1 appears by K=2–3 while relative latency rises only mildly through K=4 (curve in the supplementary material), so additional width alone does not explain the main-table gains; the difference lies in reading the generator's own one-step visual response before commit, rather than expanding longer futures or reallocating emitted paths.

Method	W/K	F1↑	CHAIR↓	Cover↑	Lat. (64 tok.)↓	Mem.↓
Greedy	1	82.1	5.3	47.2	1181 (1.00×)	16288.7
VCD	–	85.1	4.7	45.9	2063 (1.75×)	–
Beam Search	4	83.5	5.1	46.9	4449 (3.77×)	–
HALC	4	82.5	5.0	45.8	9098 (7.70×)	–
OPERA	4	84.7	5.1	46.2	7786 (6.59×)	–
RVC (K=3)	3	88.2	4.6	47.7	1389 (1.18×)	16387.5
RVC (K=4)	4	**88.5**	**4.5**	**47.9**	1535 (1.30×)	16390.2

Table 5: Width-matched quality and cost on Qwen2.5-VL. F1: nine-setting POPE; CHAIR/Cover: AMBER. Latency: ms/sample at 64 tokens (× greedy); memory: peak MB. Beam/HALC/OPERA use W=4.

5.4 Decision Analysis

POPE presence questions expose the clearest case. Yes and No are both locally fluent under language priors and usually share the shortlist $\mathcal{C}_t$, yet only one write should gain visual support. Within-set standardization compares which provisional write raises grounding mass and next-state margin before commitment. When the prior-favored answer has little visual gain and the alternative has more, $\widehat{r}_j$ can overturn $\log p_t$ before an unsupported token enters the prefix. Table 4(b) supports this account directly: zeroing the probe-selected heads removes most of the POPE gain.



Figure 4: Open-ended cases. (a) Success: RVC removes unsupported objects. (b) Failure: it recovers background detail but still makes a near-synonym error on the main object.

Open-ended CHAIR/AMBER is harder for the same score. Unsupported objects often enter as multi-token phrases whose onset may fall outside a small $\mathcal{C}_t$, so RVC can rerank only shortlisted candidates and cannot recover an excluded token. This matches the smaller Cover gains in Table 2 and the K sweep, where most POPE F1 appears once $|\mathcal{C}_t|$=2–3. Figure 4 shows both a clean correction and a residual near-miss. As a side check under the same greedy protocol, Qwen2.5-VL scores 39.0 (base) versus 39.4 (RVC) on MMMU (Yue et al., 2024) validation.

6 Conclusion

A one-step provisional write exposes candidate-conditioned visual attention and next-state margin that current-step decoding never sees. RVC turns that internal response into a relative correction of $\log p_t$, scores top-K candidates in one branch-isolated packed forward, and commits a single KV entry, without LVLM post-training and without an auxiliary reward model. Across LLaVA-1.5 and Qwen2.5-VL it attains the best score among the included methods on every POPE, CHAIR, and AMBER comparison we report, reducing open-ended hallucination while raising AMBER coverage.

Limitations. We study object hallucination on two white-box 7B backbones. Because RVC only reranks the current top-K set with a one-step response, excluded tokens stay unreachable and long-horizon corrections remain unguaranteed; visual attention is used as a support proxy.

AI USE STATEMENT

We have not used generative AI tools for any task with required disclosure. Additionally, we used generative AI tools to polish wording. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, jun 2025.

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

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4209–4221, 2025.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Chao Sun, Rongzhou Zhang, Guanyu Zhou, and Lijie Wen. OmniDPO: A preference optimization framework to address omni-modal hallucination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20172–20180, 2026.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. *An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models*, pp. 19–35. Springer Nature Switzerland, nov 2024a. ISBN 9783031730047.

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

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 *The Twelfth International Conference on Learning Representations*, 2024.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pp. 276–286, 2019.

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*, pp. 14303–14312, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):18135–18143, mar 2024. ISSN 2159-5399.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, jan 2025. ISSN 1558-2868.

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

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14549–14558, 2025.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C.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*, pp. 27228–27238, 2024.

Hyun Kim, Junho Kim, Yeon Kim, and Yong Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. In *Advances in Neural Information Processing Systems 37*, NeurIPS 2024, pp. 133571–133599. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. VLind-Bench: Measuring language priors in large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4129–4144, 2025.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, 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, 2014.

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28556–28572. Association for Computational Linguistics, 2025.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*, pp. 125–140, 2024c.

Oscar Mañas, Pierluca D'Oro, Koustuv Sinha, Adriana Romero-Soriano, Michal Drozdzal, and Aishwarya Agrawal. Controlling multimodal LLMs via reward-guided decoding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1391–1401, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

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.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pp. 146–162, 2022.

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

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, jun 2024.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, 2019.

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 MLLM hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023a.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023b.

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

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in MLLMs: Unlocking pathways to faster inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9391, 2025.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12944–12953. IEEE, jun 2024a.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. *Reflective Instruction Tuning: Mitigating Hallucinations in Large Vision-Language Models*, pp. 196–213. Springer Nature Switzerland, nov 2024. ISBN 9783031731136.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sjnErRHXf3.

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

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1615–1624. IEEE, jun 2025.