ROAM: MITIGATING VISION-LANGUAGE HALLUCINATIONS WITH CALIBRATED GROUNDING MONITORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models can produce fluent responses that drift from visual evidence during generation. We introduce **Read-Only Attention Monitoring (ROAM)**, a training-free decoding method that decouples visual-grounding monitoring from intervention. Offline, ROAM selects reusable monitoring heads using text-to-vision attention statistics and their stability across controlled image and text perturbations. The resulting monitor is calibrated once per model and reused across inputs. Online, the selected heads remain read-only: their image-attention entropy provides a grounding signal that adaptively controls the sampling temperature, sharpening the next-token distribution when the signal indicates weaker grounding. The intervention acts solely on the output logits, requiring a single forward pass per token without auxiliary decoding branches or direct modifications to internal representations. Analyses on LLaVA show that lower grounding scores are associated with higher hallucination rates before intervention, while ablations support the value of cross-environment calibration. Experiments across three VLM families and four benchmarks show the largest gains in long-form generation. On CHAIR, ROAM reduces sentence-level hallucination by 2.4–6.4 percentage points relative to vanilla decoding while increasing object recall on all three models, with near-vanilla end-to-end latency.

1 INTRODUCTION

Conditioning a large vision–language model (LVLM) on an image does not guarantee that visual evidence continues to constrain its response. LVLMs can produce fluent descriptions while introducing objects, attributes, or relations that are unsupported by the image (Rohrbach et al., 2019; Li et al., 2023). One contributing failure mode is a weakening influence of visual evidence relative to language priors during generation (Leng et al., 2023; Favero et al., 2024). Once unsupported content enters the autoregressive context, subsequent tokens may compound the error (Zhang et al., 2023). These observations motivate monitoring visual grounding as generation unfolds and using the resulting signal to guide decoding when grounding weakens.

Decode-time mitigation addresses this problem without updating model parameters. Contrastive methods such as VCD (Leng et al., 2023) and M3ID (Favero et al., 2024) compare predictions under the original input with predictions obtained under weakened visual conditioning, requiring an auxiliary decoding branch. Other approaches intervene directly on internal attention or head behaviour, for example by penalising problematic continuations, suppressing selected heads, or constructing auxiliary predictions (Huang et al., 2024; Sarkar et al., 2025; Wan et al., 2025). These methods show that internal attention is useful for hallucination mitigation, but typically use it as part of the intervention itself. We investigate a complementary question: *can a reusable grounding monitor guide decoding without directly modifying the components from which its signal is measured?*

Building such a monitor requires distinguishing a head's current attention pattern from its suitability as a reusable monitoring component. A statistic computed on one image–text pair describes behaviour in that context, but does not establish whether the head provides a useful signal across different inputs. This motivates evaluating attention statistics and their stability across controlled variations of both modalities before selecting the monitoring heads. The selected head identities can then remain fixed while their attention responds to the current image, instruction, and generated

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.

context. This separates *which components to monitor*, *what their current activity indicates*, and *how to adjust generation*.

We introduce Read-Only Attention Monitoring (ROAM), a training-free decoding method built around this separation. Offline, controlled image and text perturbations identify attention heads with low mean text-to-vision entropy ratios and low cross-environment variability; the resulting head set and aggregation weights are fixed once per model. Online, the selected heads remain read-only: their normalised image-attention entropy provides a scalar grounding score that controls the sampling temperature, with lower scores inducing stronger sharpening of the next-token distribution. The intervention acts only on the output logits, leaving the monitored heads and internal representations unchanged. ROAM therefore requires a single model trajectory and one forward pass per generated token.

We evaluate ROAM across three LVLM families and four hallucination benchmarks. Its largest gains occur in long-form generation: relative to vanilla decoding, ROAM reduces CHAIR sentence-level hallucination by 2.4–6.4 percentage points while increasing object recall on all three models, at near-vanilla end-to-end latency. On LLaVA, lower grounding scores are associated with higher hallucination rates during unmodified generation, providing evidence that the calibrated readout carries hallucination-relevant information before intervention. Controlled ablations further isolate the two central design choices: calibrated monitors outperform random and input-local alternatives under the same controller, while read-only monitoring with output-logit control outperforms directly suppressing the selected heads.

2 Related Work

Hallucination mitigation during generation. Hallucination mitigation includes instruction tuning and preference alignment (Liu et al., 2024a; Yu et al., 2024), as well as post-generation verification and revision (Yin et al., 2024). This work focuses on decode-time interventions for frozen LVLMs. VCD and M3ID contrast predictions under original and weakened visual conditioning, deriving correction signals from additional predictive distributions (Leng et al., 2023; Favero et al., 2024). OPERA uses attention aggregation to penalise over-trusted continuations and retrospectively revise token choices during beam search (Huang et al., 2024). Recent methods couple internal grounding diagnostics with output control: CEC uses contextual-attention entropy and candidate-specific look-ahead evaluations to calibrate beam-search logits (Yin et al., 2026), while VGD fits a grounding verifier offline and selectively triggers object-level rollback, suppression, and regeneration (Yang et al., 2026). ROAM uses a grounding proxy from calibrated heads to adjust sampling temperature along a single decoding trajectory, without candidate look-ahead, auxiliary predictive branches, or rollback.

Attention-head selection and intervention. Head-level methods exploit differences in attention behaviour to guide targeted interventions. SPIN dynamically suppresses heads with weak image-token attention (Sarkar et al., 2025), while EAH broadcasts a selected shallow-layer head's attention map to other heads in the same layer (Zhang et al., 2024). ONLY selects heads using the text-to-visual entropy ratio at each decoding step and combines a text-enhanced single-layer prediction with the original logits (Wan et al., 2025). ASCD instead selects reusable text-centric heads offline and contrasts a positively head-steered prediction with one obtained by suppressing attention to critical visual tokens (Wang et al., 2026). ACT profiles temporal head dynamics on a calibration set, then amplifies selected visual-attention pathways and aggregates queries across parallel generation contexts (Yan et al., 2026). CausalLens identifies visually sensitive heads and directly adjusts their hidden-state contributions through a single-pass intervention (Ji et al., 2026). ROAM selects reusable monitors using the mean and cross-environment variability of head-level entropy ratios under controlled image and text perturbations; the selected heads remain read-only while their online readout controls sampling temperature.

Visual grounding and internal-state analysis. Component-level studies show that internal model states contain useful information about visual grounding. Kang et al. (2025) identify a small subset of attention heads that support object localisation, while related analyses examine the evolution of visual representations and their functional importance across layers (Chen et al., 2024; Neo et al., 2025). Offline component analysis also guides representation editing: ITI shifts activations along truth-related directions in language models (Li et al., 2024), while multimodal approaches use causal

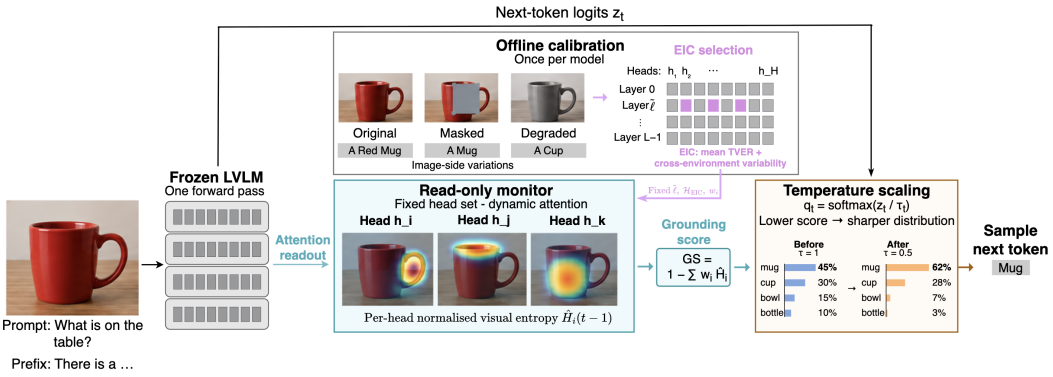


Figure 1: Overview of Read-Only Attention Monitoring (ROAM). Offline calibration fixes the monitoring layer, heads, and weights using controlled image and text variations. Online, the selected heads' visual-attention entropy provides a dynamic grounding score for next-token temperature control. Lower scores induce stronger sharpening without changing token rankings. The attention readout and logits share one forward pass; the monitored attention is not directly modified.

tracing to guide representation injection or remove hallucination-associated subspaces through null-space projection (Li et al., 2025; Yang et al., 2025). HALP instead trains probes on pre-generation hidden states to predict hallucination risk before any response token is generated (Kogilathota et al., 2026). ROAM calibrates heads without fitting a hallucination classifier and reads their attention during generation. This readout serves as a dynamic visual-focus proxy rather than a correctness test for individual tokens.

3 Read-Only Attention Monitoring

ROAM separates *which attention heads to monitor* from *how their current readouts guide sampling*. Offline calibration fixes a monitoring layer, a head set, and aggregation weights once per model. During generation, these components remain fixed, but their attention changes with the image and generated prefix. We aggregate their visual-attention entropy into a grounding score and use this score to set the next-token sampling temperature: more diffuse visual attention induces more conservative sampling. At each controlled step, the attention readout and next-token logits come from the same frozen-model forward pass. The controller rescales only the output logits; it does not directly modify the monitored attention. Figure 1 summarises this calibration–readout–control pipeline.

3.1 Offline Calibration of Reusable Monitors

The goal of calibration is to obtain a reusable monitoring configuration $(\tilde{\ell}, \mathcal{H}_{\text{EIC}}, \{w_i\})$: where to read attention and how to aggregate the selected heads. Rather than select heads from a single input, we evaluate their statistics across calibration examples and controlled input conditions.

Calibration inputs. We use $N = 8{,}000$ COCO val2014 image–text pairs $\{(V^{(n)}, T^{(n)})\}_{n=1}^{N}$ (Lin et al., 2015). Each pair is evaluated under $K = 7$ deterministic conditions: the original input, three image-side perturbations (image mismatch, region masking, and appearance degradation), and three text-side perturbations (paraphrase, negation/conflict, and context rephrasing). Each perturbation changes one modality while holding the other fixed. These conditions vary the available evidence and textual framing; they need not preserve the appropriate response. For each example–condition pair, we run the frozen model and read the attention row at the final prompt position, before generating any response.

Head statistics across conditions. Let $a_{\ell,i}^{(n,k)}$ denote this attention row for query head i at layer ℓ. We separate accessible key positions into image tokens $\mathcal{P}_{\text{vis}}$ and non-image tokens $\mathcal{P}_{\text{text}}$, including

accessible textual and special-token positions. To measure how attention is distributed within each modality, we renormalise its corresponding entries. Suppressing (n, k), let $a^+_{\ell,i}$ denote the non-negative attention row after replacing non-finite values with zero, and define

$$\bar{a}^s_{\ell,i}[j] = \frac{a^+_{\ell,i}[j]}{\max\left\{\sum_{r\in\mathcal{P}_s} a^+_{\ell,i}[r], \epsilon\right\}}, \qquad j \in \mathcal{P}_s, \quad s \in \{\text{text}, \text{vis}\}. \tag{1}$$

We use $\epsilon = 10^{-12}$ and $H(p) = -\sum_j p_j \log(p_j + \epsilon)$ with natural logarithms. Numerical clamping, when needed, is applied within the entropy computation. A non-empty modality subset with zero attention mass retains an all-zero vector and has zero entropy.

Following Wan et al. (2025), we compute the text-to-vision entropy ratio (TVER):

$$r^{(n,k)}_{\ell,i} := \text{TVER}^{(n,k)}_{\ell,i} = \frac{H\left(\bar{a}^{\text{text},(n,k)}_{\ell,i}\right)}{\max\left\{H\left(\bar{a}^{\text{vis},(n,k)}_{\ell,i}\right), \epsilon\right\}}. \tag{2}$$

TVER compares text-attention entropy with visual-attention entropy; it does not measure the total attention mass assigned to either modality. We use it as an empirical statistic for selecting monitoring components, not as the grounding score of a decoding step. The online score, defined in Section 3.2, reads the selected heads' current visual-attention entropy instead.

Calibration summarises both the average TVER of each head and its sensitivity to the input conditions. Let $A^{(n)}_{\ell,i}$ be the condition-averaged TVER for example n:

$$\begin{aligned} A^{(n)}_{\ell,i} &= \frac{1}{K}\sum_{k=1}^{K} r^{(n,k)}_{\ell,i}, \\ \mu_{\ell,i} &= \frac{1}{N}\sum_{n=1}^{N} A^{(n)}_{\ell,i}, \\ \sigma^2_{\ell,i} &= \frac{1}{N}\sum_{n=1}^{N}\left[\frac{1}{K}\sum_{k=1}^{K}\left(r^{(n,k)}_{\ell,i} - A^{(n)}_{\ell,i}\right)^2\right]. \end{aligned} \tag{3}$$

Thus, $\mu_{\ell,i}$ measures the overall mean ratio, whereas $\sigma^2_{\ell,i}$ measures variation across conditions *within the same example*, averaged over examples. It is not the variance across different calibration examples.

Selecting and weighting monitoring heads. We first select one monitoring layer, then filter and weight its heads. Let $\mathcal{I}_\ell$ be the query heads at layer ℓ. We min–max normalise their mean ratios within each layer:

$$\hat{\mu}_{\ell,i} = \frac{\mu_{\ell,i} - \min_{j\in\mathcal{I}_\ell}\mu_{\ell,j}}{\max_{j\in\mathcal{I}_\ell}\mu_{\ell,j} - \min_{j\in\mathcal{I}_\ell}\mu_{\ell,j} + \epsilon}. \tag{4}$$

We choose the layer with the greatest dispersion of these range-normalised head means:

$$\tilde{\ell} = \arg\max_{\ell} \text{Std}_{i\in\mathcal{I}_\ell}\left(\hat{\mu}_{\ell,i}\right). \tag{5}$$

Within this layer, candidates are the heads with mean TVER below the layer-wide mean:

$$\mathcal{M} = \left\{ i \in \mathcal{I}_{\tilde{\ell}} : \ \mu_{\tilde{\ell},i} < \frac{1}{|\mathcal{I}_{\tilde{\ell}}|}\sum_{j\in\mathcal{I}_{\tilde{\ell}}}\mu_{\tilde{\ell},j} \right\}. \tag{6}$$

This yields a data-dependent cutoff without an additional tuned threshold.

For candidates in $\mathcal{M}$, we separately min–max normalise $\mu_{\tilde{\ell},i}$ and $\sigma^2_{\tilde{\ell},i}$ to obtain $\tilde{\mu}_i$ and $\tilde{\sigma}^2_i$. Both use the scaling form in Eq. 4, with extrema taken over $\mathcal{M}$ and the same stabiliser ϵ. We then define the *environment-informed calibration* (EIC) score:

$$\text{EIC}_{\tilde{\ell},i} = \begin{cases} (1-\tilde{\mu}_i)(1-\tilde{\sigma}^2_i), & i \in \mathcal{M}, \\ 0, & \text{otherwise}. \end{cases} \tag{7}$$

The mean-ratio cutoff determines candidate eligibility. Within this set, EIC gives greater weight to heads with lower relative mean ratios and lower within-example variability across conditions. The monitoring set and aggregation weights are

$$\mathcal{H}_{\mathrm{EIC}} = \{i : \mathrm{EIC}_{\tilde{\ell},i} > 0\}, \qquad w_i = \frac{\mathrm{EIC}_{\tilde{\ell},i}}{\sum_{j\in\mathcal{H}_{\mathrm{EIC}}} \mathrm{EIC}_{\tilde{\ell},j} + 10^{-8}}, \quad i \in \mathcal{H}_{\mathrm{EIC}}. \tag{8}$$

Calibration fixes $\tilde{\ell}$, $\mathcal{H}_{\mathrm{EIC}}$, and $\{w_i\}$ for inference, but does not freeze their attention patterns. If no candidate passes the cutoff, calibration returns an empty monitor and decoding uses the unmodified sampling distribution. The resulting monitor is evaluated in Section 4.3.

3.2 Online Grounding Readout

The calibrated configuration specifies *where to read*; the current image and generated prefix determine *what is read*. Given an image V, an instruction T, and prefix $y_{<t}$, consider prediction of token y_t for $t \geq 2$. Processing y_{t-1} produces both the next-token logits z_t and the attention row whose query is y_{t-1}. We use this same row to compute the score that controls sampling of y_t; no additional model evaluation is required.

For each monitored head, let $\bar{a}^{\mathrm{vis}}_{\tilde{\ell},i}(t-1)$ denote its visual attention after the modality-specific normalisation in Eq. 1. For $L_{\mathrm{img}} > 1$ image tokens, define

$$\hat{H}_i(t-1) = \frac{H\left(\bar{a}^{\mathrm{vis}}_{\tilde{\ell},i}(t-1)\right)}{\log L_{\mathrm{img}}}, \tag{9}$$

and aggregate these normalised entropies into the grounding score:

$$\mathrm{GS}(t-1) = 1 - \sum_{i\in\mathcal{H}_{\mathrm{EIC}}} w_i \hat{H}_i(t-1). \tag{10}$$

The denominator $\log L_{\mathrm{img}}$ normalises entropy relative to uniform attention over image tokens. Concentrated visual-attention distributions produce higher scores, while diffuse distributions produce lower scores. Because attention is normalised within the image-token subset, GS describes visual concentration, not the total attention paid to the image. We use it as a visual-focus proxy rather than a token-level correctness signal; its relationship with hallucination is evaluated in Section 4.3. We do not clip GS, and an empty monitor yields $\mathrm{GS} = 1$.

3.3 Grounding-Guided Temperature Control

The controller uses GS to adjust *how conservatively to sample*, not to identify which candidate token is visually correct. When monitored visual attention is diffuse, we lower the temperature to favour the model's existing high-probability candidates. Higher scores induce less sharpening, while a temperature floor limits its strength.

For $t \geq 2$, a sharpening strength $\alpha \geq 0$ and $0 < \tau_{\mathrm{floor}} \leq 1$ define

$$\tau_{\mathrm{raw}}(t) = \max\left\{1 - \alpha\big(1 - \mathrm{GS}(t-1)\big), \tau_{\mathrm{floor}}\right\}, \qquad \tau_t = \begin{cases} \tau_{\mathrm{raw}}(t), & \tau_{\mathrm{raw}}(t) < 0.99, \\ 1, & \tau_{\mathrm{raw}}(t) \geq 0.99. \end{cases} \tag{11}$$

The 0.99 branch bypasses near-unit temperature changes. We sample the next token from

$$q_t = \mathrm{softmax}(z_t/\tau_t), \qquad y_t \sim \mathrm{Categorical}(q_t). \tag{12}$$

For each new input, we reset $\mathrm{GS}(0) = 1$ and apply no control during prefill, so the first generated token uses $\tau_1 = 1$. Control starts when predicting y_2: processing y_1 provides both z_2 and the attention used to compute $\mathrm{GS}(1)$. All reported models and benchmarks use $\alpha = 0.7$ and $\tau_{\mathrm{floor}} = 0.3$.

For a fixed prefix, positive temperature scaling preserves token rankings and changes only sampling probabilities. Thus, ROAM affects stochastic decoding but leaves greedy token selection unchanged. Sharpening is not a correctness guarantee: a high-probability unsupported token can also become more likely.

Read-only monitoring and computational cost. For a fixed input and prefix, ROAM leaves the model's attention weights, head outputs, and hidden representations unchanged; the controller acts only on the output logits after the forward pass. Changes in subsequent internal states therefore arise only through the sampled token and resulting autoregressive prefix. Online decoding follows one model trajectory and requires one forward pass per generated token, without an auxiliary predictive branch.

Calibration requires NK frozen-model evaluations once per model. Given the selected attention rows, the grounding readout requires $\mathcal{O}(|\mathcal{H}_{\mathrm{EIC}}|L_{\mathrm{img}})$ operations per monitored step, while temperature scaling adds an elementwise operation over the vocabulary logits. Total overhead also depends on how attention weights are exposed. On LLaVA, measured end-to-end latency and peak memory remain near vanilla levels; detailed measurements are reported in Section 4.2, with pseudocode in Appendix A.3.

4 Experiments

Our evaluation examines the effectiveness of ROAM in reducing hallucination, the information captured by its calibrated monitor, and the contribution of monitor selection and decoding control.

4.1 Experimental Setup

We evaluate ROAM on three LVLM families: LLaVA-v1.5-7B (Liu et al., 2024b), Qwen3-VL-8B-Instruct (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025), covering both MHA and GQA architectures. We evaluate free-form object hallucination on CHAIR (Rohrbach et al., 2019) (500 COCO images) and AMBER (Wang et al., 2024) (1,004 images), and assess short-answer performance on POPE (Li et al., 2023) (COCO random split; 3,000 questions) and MME (Fu et al., 2025) (2,374 questions; reported score sums existence, count, position, and color). We compare against vanilla decoding, VCD (Leng et al., 2023), M3ID (Favero et al., 2024), ONLY (Wan et al., 2025), and ASCD (Wang et al., 2026). CHAIR and AMBER use free-form generation of up to 128 new tokens, whereas POPE and MME use short answers of at most eight tokens. The base sampler uses temperature 1.0 and `top_p` $= 1.0$; ROAM applies the adaptive temperature defined in Section 3.3. All ROAM results use a single global setting, $\alpha = 0.7$ and $\tau_{\mathrm{floor}} = 0.3$, without benchmark-specific tuning. All experiments run on a single NVIDIA A100. Full model configurations, generation settings, and baseline hyperparameters are provided in Appendix A.

4.2 Hallucination Mitigation Across Models

Table 1 reports hallucination and coverage metrics alongside short-answer performance across the three model families.

Reducing hallucination while retaining object coverage. Relative to vanilla decoding, ROAM reduces CHAIRs by 2.4–6.4 percentage points, lowers CHAIRi, and increases object recall on all three models. LLaVA illustrates this joint improvement: CHAIRs falls from 55.2 to 48.8, while recall rises from 71.4 to 76.7. On AMBER, ROAM likewise reduces CHAIR and Hal while maintaining or increasing Cover relative to vanilla in every model family. The joint improvement supports more faithful object descriptions without sacrificing measured coverage.

Against the compared methods, ROAM achieves the lowest CHAIRs and CHAIRi on LLaVA and InternVL, and the best or tied-best AMBER CHAIR on all three models. On Qwen3, ONLY obtains slightly lower CHAIR scores, whereas ROAM retains higher recall and performs better on AMBER CHAIR and Hal. These comparisons highlight a favourable balance between reducing unsupported object mentions and retaining correctly described objects.

Short-answer performance. ROAM matches vanilla MME on all three models: 523.33 on LLaVA, 711.66 on Qwen3, and 721.67 on InternVL. In our MME evaluation, the yes/no decision is encoded in the first generated token, which ROAM samples without intervention. MME therefore measures preservation of this short-answer decision rather than a direct effect of the controller on it. POPE uses a different parsing rule: its evaluator scans the complete first output line for negative

Table 1: **Hallucination evaluation across three LVLM families.** Bold denotes the best reported value, including ties, within each model block. Across the three models, ROAM most consistently improves long-form hallucination metrics while increasing CHAIR object recall relative to vanilla decoding.

	CHAIR			AMBER				Short-form	
Method	CHAIRs↓	CHAIRi↓	Rec.↑	CHAIR↓	Cover↑	Hal↓	Cog↓	POPE↑	MME↑
LLaVA-v1.5-7B									
Vanilla	55.20	16.18	71.4	11.1	49.7	48.0	4.2	84.63	523.33
VCD	55.20	15.43	75.2	10.5	51.3	45.0	4.5	84.60	556.67
M3ID	58.80	15.90	**79.4**	8.1	**56.8**	44.9	**3.3**	86.33	528.33
ONLY	52.40	15.13	73.5	8.5	48.8	**35.6**	3.6	87.97	**631.66**
ASCD	52.40	14.78	75.7	8.0	52.7	39.3	**3.3**	85.50	566.67
ROAM	**48.80**	**13.55**	76.7	**7.8**	51.6	36.4	4.0	**89.03**	523.33
Qwen3-VL-8B-Instruct									
Vanilla	24.80	8.51	48.3	7.7	48.9	37.4	1.3	90.63	**711.66**
VCD	23.20	8.97	48.7	8.3	47.7	37.9	1.3	**90.77**	686.66
M3ID	26.80	9.30	48.8	7.6	49.1	36.8	**1.0**	**90.77**	698.33
ONLY	**22.00**	**7.81**	46.8	7.2	**50.0**	34.0	1.4	90.00	675.00
ROAM	22.40	8.33	**49.2**	**6.8**	49.0	**33.6**	**1.0**	90.63	**711.66**
InternVL3.5-8B									
Vanilla	34.40	10.10	65.5	6.8	44.8	31.2	1.4	92.33	**721.67**
VCD	33.40	10.95	64.2	7.6	49.9	35.3	1.5	91.63	638.33
M3ID	34.40	10.13	65.9	6.4	45.2	29.1	**0.7**	92.67	698.33
ONLY	29.60	9.06	64.5	**5.7**	**51.2**	30.5	1.5	92.47	711.66
ROAM	**28.80**	**8.43**	**66.3**	**5.7**	47.0	**29.0**	1.2	**92.93**	**721.67**

terms. Consequently, later generated tokens can affect the final parsed answer even though ROAM does not alter the first-token distribution. Under this evaluation, ROAM improves POPE on LLaVA and InternVL and matches vanilla on Qwen3.

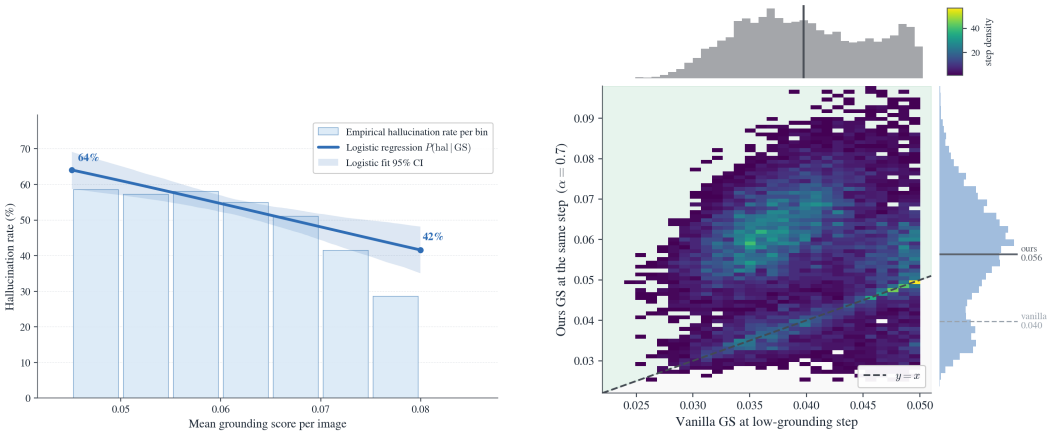
Inference cost. ROAM combines these long-form improvements with a single decoding trajectory and one forward pass per token. On LLaVA, ROAM has near-vanilla end-to-end latency (3.30 versus 3.32 seconds) and unchanged peak memory, while VCD and M3ID require approximately twice the vanilla latency. This highlights the efficiency of obtaining the grounding signal from the same forward pass used for generation.

4.3 Grounding Monitor Validation

We next analyse the calibrated monitor on LLaVA-v1.5-7B. We examine its reproducibility, its association with hallucination under vanilla decoding, and its behaviour along controlled trajectories.

Reproducibility of offline calibration. Splitting the calibration data into two disjoint halves and rerunning calibration independently selects the same layer and the same 13-head monitoring set (Jaccard overlap 1.0), with EIC score correlations above 0.99. The full-data run selects 12 of these 13 heads. This agreement supports stability with respect to the sampled calibration examples within the evaluated distribution. The appendix reports the complete split-half statistics and modality-specific entropy responses to the calibration perturbations.

Association with hallucination before control. We evaluate the online readout independently of the controller by recording the mean grounding score of each caption during *unmodified vanilla generation* on 2,000 COCO images. In Figure 2(a), the highest-score quartile has a hallucination rate 10.2 percentage points below the lowest-score quartile. The overall trend is negative (Cochran–Armitage $p = 3.55 \times 10^{-4}$). Thus, the calibrated readout carries information associated with caption-level hallucination even before it is used to control decoding. As a sanity check, a counterfactual monitor built from 12 zero-EIC heads yields a smaller hallucination-rate gap (3.4 vs. 10.2 points), supporting a stronger association for the calibrated head set (Appendix B).



(a) Caption-level association.

(b) Low-score trajectory comparison.

Figure 2: **Grounding-score analyses on LLaVA-v1.5-7B. (a)** During vanilla generation, lower caption-level grounding scores are associated with higher hallucination rates overall. **(b)** Position-aligned grounding scores at locations with vanilla GS < 0.05. The horizontal markers show the means across the 11,333 retained position pairs: 0.03974 for vanilla and 0.05641 for ROAM, a mean paired difference of 0.01667. Together, these analyses show that the calibrated readout is associated with hallucination before intervention and that ROAM exhibits higher grounding scores in the position-aligned low-score regime.

EIC ranking and head-level sensitivity. As an auxiliary diagnostic, we test whether EIC ranking tracks the sensitivity of individual head representations to head suppression. At the selected LLaVA layer and final query position, we compare each head's pre-output-projection vector between the original branch and an auxiliary branch in which selected heads are suppressed, while keeping the image and prompt unchanged. We average the resulting Euclidean change over generative and yes/no prompts. EIC rank correlates with this head-level sensitivity on both prompt types (Spearman $\rho = 0.59$–0.60). Thus, higher-ranked EIC heads tend to exhibit larger representation changes under this auxiliary perturbation, providing complementary evidence that the calibration ranking identifies responsive heads. The complete protocol and results are provided in Appendix B.2.

Grounding scores along generated trajectories. We compare vanilla and ROAM generations on 500 CHAIR images. Within each image, we align positions by generated-token ordinal and truncate the two responses to their common length. At position $t \geq 2$, the recorded $\mathrm{GS}(t-1)$ is computed from the attention row produced while processing y_{t-1}, before sampling y_t. We exclude non-finite scores and the initial $\mathrm{GS}(0) = 1$ sentinel, then retain positions with vanilla GS < 0.05. This yields 11,333 aligned position pairs.

Across these retained position pairs, mean vanilla GS is 0.03974 and mean ROAM GS is 0.05641, giving a mean paired difference of 0.01667 (Figure 2(b)). Thus, in positions that are low-grounding under vanilla decoding, ROAM trajectories exhibit higher grounding scores, consistent with the intended effect of the controller. Alignment is by generated-token ordinal rather than token identity, so once the generated prefixes diverge the scores describe different trajectories. We therefore interpret this as descriptive trajectory-level evidence rather than a controlled comparison at identical prefixes.

4.4 Ablation Study

We examine the monitoring configuration and the intervention applied during generation in Table 2. Additional experiments assess the contribution of calibration environments and the reuse of the selected heads in another decoder.

Calibrated monitors and input-local alternatives. Table 2(a) compares EIC with all-head, random-head, and inverted-EIC controls, as well as two input-local alternatives. The *per-step TVER signal* uses $\exp(-\overline{\mathrm{TVER}}(t))$ directly as the online control signal. The *per-step TVER-*

Table 2: **Ablation of monitoring configuration and intervention on LLaVA-v1.5-7B. (a)** The head set or online signal is varied while retaining the same temperature mapping and controller parameters. **(b)** Static suppression of the calibrated heads is compared with ROAM's read-only monitoring and output control, alongside vanilla. EIC gives the lowest CHAIRs among the evaluated monitoring variants, while read-only output control outperforms static suppression of the same calibrated heads.

(a) Monitoring configuration

Monitor / signal	CHAIRs↓
All heads	52.60
Random heads	51.40
Inverted EIC	50.00
Per-step TVER signal	51.60
Per-step TVER-selected heads	51.20
EIC	**48.80**

(b) Intervention

Method	CHAIRs↓	POPE↑
Vanilla	55.20	84.63
Static suppression	55.80	84.67
ROAM	**48.80**	**89.03**

selected heads variant instead recomputes TVER at every decoding step and retains heads satisfying $\mathrm{TVER}_i(t) < \overline{\mathrm{TVER}}(t)$, replacing the fixed EIC-selected monitor with an input-local head set.

Under the same temperature mapping, EIC obtains the lowest CHAIRs among the evaluated monitoring configurations, improving over both per-step TVER variants (48.8 versus 51.2–51.6), with similar object recall. Thus, the monitor calibrated once per model outperforms both input-local variants on sentence-level hallucination while retaining similar object coverage.

Read-only monitoring versus static head suppression. With EIC calibration fixed, directly suppressing the selected heads does not reproduce the CHAIRs improvement obtained by ROAM (Table 2(b)). Suppression remains near vanilla, whereas using these heads to monitor generation and control the output distribution reduces hallucination. This comparison supports using the calibrated heads as read-only monitors rather than direct intervention targets.

Calibration environments and reuse across decoders. On InternVL, CHAIRs decreases progressively as more calibration environments are added, from 31.0 at $K = 1$ to 30.8 at $K = 3$, 30.2 at $K = 5$, and 28.8 at $K = 7$, while recall at $K = 7$ remains essentially unchanged from $K = 1$. This supports the value of combining complementary image- and text-side calibration conditions. Separately, replacing ONLY's per-input head selection with the offline EIC set on LLaVA reduces CHAIRs from 52.4 to 47.6 while leaving the rest of its decoding pipeline unchanged. This transfer shows that the calibrated head selection can also benefit a decoder that uses the heads for a different intervention. Additional environment, monitoring-layer, and controller-strength ablations are provided in the appendix.

5 Conclusion

We introduced Read-Only Attention Monitoring (ROAM), a training-free decode-time framework that separates visual-grounding monitoring from the intervention used to control generation. Rather than modifying the attention heads from which grounding is measured, ROAM calibrates a fixed set of read-only monitoring heads across controlled image and text perturbations and uses their evolving image-attention concentration to adapt the next-token sampling temperature. Across three LVLM families and four hallucination benchmarks, ROAM reduces object hallucination most consistently in long-form generation, while increasing CHAIR object recall on all three models and preserving broader multimodal capability at near-vanilla inference cost. Our analyses further show that the calibrated monitor is reproducible, that lower grounding scores are associated with higher hallucination rates before intervention, and that grounding-guided control shifts low-grounding trajectories toward stronger subsequent grounding. Together, these results show that internal visual grounding can serve not only as a diagnostic of hallucination, but as a lightweight online signal for controlling generation as it unfolds.

AI USE STATEMENT

Generative AI tools were used during the preparation of this paper for language editing, including grammar, clarity, and readability improvements, and as a programming assistant for debugging. The research questions, methodology, experimental design, evaluation, analysis, and scientific conclusions were developed by the authors. All AI-assisted manuscript and code changes were reviewed and validated by the authors, and all reported experimental results were produced using the authors' own experimental pipeline. The authors take full responsibility for the content of the paper.

REPRODUCIBILITY STATEMENT

We report the calibration-set size and controlled calibration environments, model-specific monitoring layers and head-set sizes, global decoding hyperparameters, generation settings, model configurations, evaluation benchmarks, baseline settings, and inference measurements. Appendix A provides the complete calibration and generation settings together with pseudocode for offline EIC calibration and online ROAM decoding. The remaining appendix sections provide additional monitor validation, ablations, capability results, statistical uncertainty, efficiency measurements, and qualitative success and failure cases. Anonymous code and evaluation scripts will be provided as supplementary material.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

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, 2024. URL https://arxiv.org/abs/2403.06764.

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, 2024. URL https://arxiv.org/abs/2403.14003.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. URL https://arxiv.org/abs/2306.13394.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

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, 2024. URL https://arxiv.org/abs/2311.17911.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4199–4209, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Ji_CausalLens_Sensitivity-Guided_Multi-Head_Causal_Intervention_for_Hallucination_Mitigation_in_Large_CVPR_2026_paper.html.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding, 2025. URL https://arxiv.org/abs/2503.06287.

Sai Akhil Kogilathota, Sripadha Vallabha E G, 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, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.287. URL https://aclanthology.org/2026.eacl-long.287/.

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, 2023. URL https://arxiv.org/abs/2311.16922.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation, 2025. URL https://arxiv.org/abs/2511.05923.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023. URL https://arxiv.org/abs/2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL https://arxiv.org/abs/1405.0312.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning, 2024a. URL https://arxiv.org/abs/2306.14565.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024b. URL https://arxiv.org/abs/2310.03744.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024c. URL https://arxiv.org/abs/2307.06281.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2025. URL https://arxiv.org/abs/2410.07149.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning, 2019. URL https://arxiv.org/abs/1809.02156.

Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Beerel, and Souvik Kundu. Mitigating hallucinations in vision-language models through image-guided head suppression, 2025. URL https://arxiv.org/abs/2505.16411.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models, 2025. URL https://arxiv.org/abs/2507.00898.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation, 2024. URL https://arxiv.org/abs/2311.07397.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Yujun Wang, Aniri, Jinhe Bi, Soren Pirk, and Yunpu Ma. Ascd: Attention-steerable contrastive decoding for reducing hallucination in mllm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):10306–10314, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i12.38000. URL http://dx.doi.org/10.1609/aaai.v40i12.38000.

Bei Yan, Yuecong Min, Jie Zhang, Shiguang Shan, and Xilin Chen. ACT now: Preempting LVLM hallucinations via adaptive context integration. *arXiv preprint arXiv:2604.00983*, 2026. URL https://arxiv.org/abs/2604.00983.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection, 2025. URL https://arxiv.org/abs/2412.13817.

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*, 2026. URL https://arxiv.org/abs/2607.27823.

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), December 2024. ISSN 1869-1919. doi: 10.1007/s11432-024-4251-x. URL http://dx.doi.org/10.1007/s11432-024-4251-x.

Xuanyu Yin, Daowan Peng, and Wei Wei. Mitigating hallucinations in vision-language models via contextual entropy calibration. In *Proceedings of the 2026 International Conference on Multimedia Retrieval*, pp. 2022–2031. Association for Computing Machinery, June 2026. doi: 10.1145/3805622.3810847. URL https://doi.org/10.1145/3805622.3810847.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024. URL https://arxiv.org/abs/2312.00849.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. How language model hallucinations can snowball, 2023. URL https://arxiv.org/abs/2305.13534.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms, 2024. URL https://arxiv.org/abs/2411.09968.

A Reproducibility Details

A.1 Configurations and Generation Settings

All main experiments use a single global sharpening strength $\alpha = 0.7$ and temperature floor $\tau_{\text{floor}} = 0.3$. Table 3 summarises the model-specific calibration configuration and benchmark generation settings.

Table 3: **ROAM configuration and generation settings.** All main ROAM results use the same α and τ_{floor} without benchmark-specific tuning.

(a) Model-specific configuration

	LLaVA	Qwen3	InternVL
Calibration N	8000	8000	8000
Environments K	7	7	7
Layer $\tilde{\ell}$	1	0	1
Selected query heads	12/32	14/32	24/32
α	.7	.7	.7
τ_{floor}	.3	.3	.3

(b) Generation settings

	CHAIR	AMBER	POPE	MME
Seed	3407	42	42	42
Max tokens	128	128	8	8
Samples	500	1004	3000	2374

All benchmarks use sampling with temperature 1.0 and `top_p` $= 1.0$.

For VCD and M3ID, we use $\alpha_{\text{cd}} = 1.0$ and $\beta_{\text{cd}} = 0.1$. For ONLY, $(\alpha_{\text{pos}}, \alpha_{\text{neg}}) = (3.0, 1.0)$ for all models. On LLaVA, CHAIR uses $\beta = 0.01$ and $\gamma = 0.25$, whereas AMBER, POPE, MME, MMVP, and MMBench use $\beta = 0.1$ and $\gamma = 0.2$. On Qwen3 and InternVL, we use $\beta = \gamma = 0.1$. For ASCD, we use the official LLaVA-v1.5-7B configuration with $\alpha = 1.0$, $\beta = 0.1$, and the released 32-head set.

A.2 Calibration Inputs and Perturbations

For each model, we randomly select $N = 8{,}000$ COCO val2014 calibration images using the fixed seed $S = 0$. We cycle deterministically through four base instructions: "Describe this image.", "What is happening in this picture?", "What do you see in this image?", and "Can you describe what's in this photo?". Each instruction is inserted into the model's native chat template. The same four-instruction scheme is used for all three models.

Each example is evaluated under seven conditions: (1) the original input; (2) an image mismatch obtained by circularly shifting the calibration image ordering by 997 positions; (3) a grey square mask with side length $0.2 \min(W, H)$; (4) appearance degradation combining Gaussian blur of radius 1, independently sampled brightness, contrast, and colour factors from $[0.8, 1.2]$, and half-resolution bilinear downsampling and upsampling; (5) rule-based paraphrasing; (6) semantic-conflict rewriting using object-collocation substitutions when applicable and context-rephrasing otherwise; and (7) locative/context rephrasing. Stochastic choices within these perturbations are seeded deterministically by the calibration seed, example ID, and condition.

Calibration–evaluation overlap and sensitivity. For each model, 106 of the 8,000 original calibration image IDs overlap with the 500 CHAIR evaluation images, while there is no image-ID overlap with POPE, AMBER, or MME. Calibration uses no evaluation labels, CHAIR annotations, generated benchmark answers, or benchmark scores. To test whether the CHAIR overlap affects the results, we recompute each monitor using 8,000 COCO images sampled after excluding all evaluation image IDs, while retaining the originally selected monitoring layer and otherwise matching the calibration and evaluation setup. The paired-bootstrap 95% confidence intervals for the CHAIRs change (disjoint minus original) are $[-1.6, 0.6]$ for LLaVA, $[-2.0, 0.4]$ for Qwen3, and $[0.0, 0.0]$ percentage points for InternVL. Thus, removing evaluation images from calibration does not degrade CHAIRs in this fixed-layer sensitivity analysis.

A.3 Offline Calibration and Online Decoding

The offline procedure is run once per model and reused across all evaluation inputs. Importantly, the EIC variance measures cross-environment variability *within each calibration example*, averaged across examples.

Table 4: **Pseudocode for ROAM.** Left: one-time offline EIC calibration. Right: single-pass online decoding.

Algorithm 1: Offline EIC calibration

Require: Frozen LVLM; calibration set $\{(V^{(n)}, T^{(n)})\}_{n=1}^{N}$; environments $\{g_k\}_{k=1}^{K}$

for $n = 1, \ldots, N$ **do**
  for $k = 1, \ldots, K$ **do**
    Apply g_k and extract the final-prompt attention row
    Treat empty modality subsets as invalid inputs
    Replace non-finite attention values by zero
    Compute $r_{\ell,i}^{(n,k)}$ using Eq. 2
  end for
  Compute the condition mean $A_{\ell,i}^{(n)}$ and within-example variance across k
end for
Compute $\mu_{\ell,i}$ and $\sigma_{\ell,i}^2$ using Eq. 3
Normalise $\mu_{\ell,i}$ within each layer using Eq. 4
Select $\tilde{\ell}$ using Eq. 5
Form candidate set $\mathcal{M}$ using Eq. 6
if $\mathcal{M} = \varnothing$ **then**
  return $\tilde{\ell}, \varnothing, \varnothing$
end if
Min–max normalise μ and σ^2 over $\mathcal{M}$
Compute $\mathrm{EIC}_{\tilde{\ell},i}$ using Eq. 7
Form $\mathcal{H}_{\mathrm{EIC}}$ and weights $\{w_i\}$ using Eq. 8
if $\mathcal{H}_{\mathrm{EIC}} = \varnothing$ **then**
  return $\tilde{\ell}, \varnothing, \varnothing$
end if
return $\tilde{\ell}, \mathcal{H}_{\mathrm{EIC}}, \{w_i\}$

Algorithm 2: Online ROAM

Require: Frozen LVLM; $\tilde{\ell}$; $\mathcal{H}_{\mathrm{EIC}}$; $\{w_i\}$; α; τ_{floor}

Set $\mathrm{GS}(0) \leftarrow 1$
Prefill image and prompt to obtain z_1
Sample $y_1 \sim \mathrm{Categorical}(\mathrm{softmax}(z_1))$
if $y_1 = \texttt{EOS}$ **then**
  return generated sequence
end if
for $t = 2, 3, \ldots$ **do**
  Process y_{t-1} to obtain z_t and its query attention row
  if $\mathcal{H}_{\mathrm{EIC}} = \varnothing$ **then**
    $\mathrm{GS}(t-1) \leftarrow 1$
  else
    for $i \in \mathcal{H}_{\mathrm{EIC}}$ **do**
      Compute $\hat{H}_i(t-1)$ using Eq. 9
    end for
    Compute $\mathrm{GS}(t-1)$ using Eq. 10
  end if
  Compute τ_t using Eq. 11
  Sample y_t using Eq. 12
  if $y_t = \texttt{EOS}$ **then**
    break
  end if
end for
return generated sequence

B Additional Grounding-Monitor Validation

This section provides additional evidence for the validity and reproducibility of the calibrated monitor, together with a head-level sensitivity analysis of the selected heads.

B.1 Environment Validation and Calibration Stability

We evaluate two properties of the calibration procedure: whether the controlled environments produce the intended modality-specific entropy responses, and whether the selected monitor is stable to the particular calibration examples used. For image-side environments, the reported ratio compares the dataset-level mean absolute change in visual entropy with the largest corresponding change produced by a text-side environment. For text-side environments, it compares the dataset-level mean absolute change in text entropy with the visual-entropy change under the same environment. For calibration stability, we rerun the full procedure on two disjoint halves of the calibration set and compare the selected layer, head set, and EIC scores.

Table 5: **Calibration-environment validation and split-half reproducibility on LLaVA-v1.5-7B.** The controlled environments predominantly affect their intended modality, while independent split-half calibrations recover the same monitoring layer and head set with highly correlated EIC scores.

(a) Environment selectivity

Type	Environment	Ratio
Image	Mismatch	15.70×
	Masking	5.59×
	Appearance	7.86×
Text	Paraphrase	18.27×
	Negation/conflict	3.83×
	Context rephrase	3.73×

(b) Split-half calibration

Statistic	Result
Selected layer	1/1
Selected-head overlap	13/13
Jaccard	1.000
Pearson EIC	0.9963
Spearman EIC	0.9976
Mean $\lvert\Delta\mathrm{TVER}\rvert$	0.000259
Max. $\lvert\Delta\mathrm{TVER}\rvert$	0.003418
Full-data overlap	12/13 ($J = .923$)

Table 5(a) reports the environment analysis over 200 images and the 12 EIC-selected heads at layer $\tilde{\ell} = 1$. For all six environments, paired-bootstrap 95% confidence intervals on the corresponding per-image entropy differences exclude zero; Wilcoxon signed-rank tests give p-values between 5.42×10^{-33} and 6.10×10^{-20}.

Table 5(b) reports the split-half stability analysis, in which the 8,000 calibration examples are divided into two disjoint 4,000-example subsets and calibration is rerun independently. The two split-half runs select identical head sets, while the full-data monitor overlaps with 12 of the 13 selected heads ($J = .923$), indicating close agreement across calibration subsets.

Random-head counterfactual. To complement the caption-level association in Figure 2(a), we construct a counterfactual monitor from 12 randomly selected zero-EIC heads at the same LLaVA layer, using the calibrated monitor's weights. On the same 2,000 vanilla generations, the hallucination-rate gap between the lowest- and highest-score quartiles is 10.2 percentage points for the EIC monitor and 3.4 points for the random-head monitor. The corresponding point-biserial correlations are $r_{\mathrm{pb}} = -0.089$ ($p < 0.001$) and $r_{\mathrm{pb}} = -0.021$ ($p = 0.36$), respectively. This supports a stronger hallucination-relevant association for the calibrated head set than for this random-head control.

B.2 Head-Level Sensitivity of EIC-Ranked Heads

EIC is computed from observational attention statistics, so we separately test whether its ranking tracks the sensitivity of individual head representations to an auxiliary head-suppression perturbation. At the selected LLaVA layer and final query position, let v_h^{real} and v_h^{sup} denote head h's pre-output-projection vectors under the original branch and an auxiliary ONLY-style branch in which selected heads are suppressed while the image and prompt are unchanged. For each head, we compute

$$\Delta_h = \left\| v_h^{\mathrm{real}} - v_h^{\mathrm{sup}} \right\|_2 , \tag{13}$$

average this quantity over 200 generative prompts and 200 yes/no prompts, and correlate the resulting head-level values with EIC rank. As shown in Figure 3, EIC rank correlates with this sensitivity measure on both prompt types ($\rho_{\mathrm{gen}} = 0.59$, $\rho_{\mathrm{y/n}} = 0.60$; both $p < 0.001$). Thus, higher-ranked EIC heads tend to exhibit larger representation changes under this auxiliary perturbation, providing complementary evidence that EIC prioritises responsive heads.

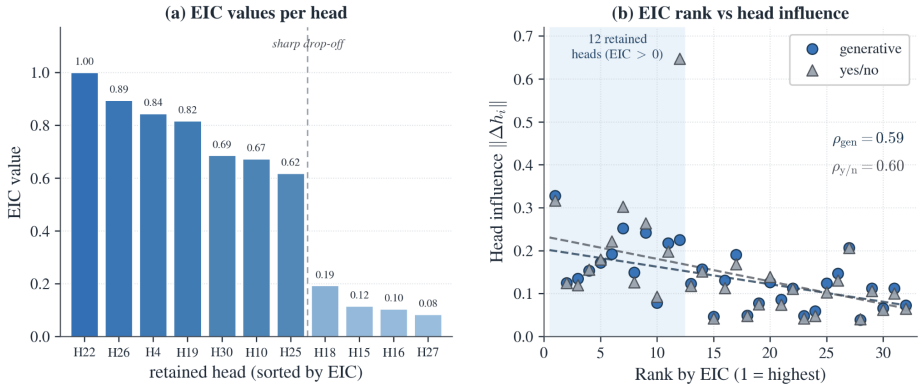


Figure 3: **Head-level sensitivity of EIC-ranked heads.** Higher-ranked EIC heads tend to exhibit larger pre-output-projection representation changes between the original and auxiliary head-suppressed branches at the selected LLaVA layer. The correlation provides complementary evidence that EIC prioritises heads whose representations are more responsive to this perturbation.

C Additional Ablation Studies

The main paper reports the head-set controls and the comparison between static suppression and adaptive output control. Here we report complementary calibration, signal-selection, transfer, and sensitivity ablations.

C.1 Calibration Environments and Per-Step TVER

Table 6: **Additional calibration and monitor ablations.** **(a)** Number of calibration environments on InternVL3.5-8B. **(b)** Offline EIC versus input-local TVER variants on LLaVA-v1.5-7B. Together, the results support both the use of complementary calibration environments and fixed offline EIC over the evaluated input-local TVER alternatives.

(a) Calibration environments

K	CHAIRs	CHAIRi	Rec.
1	31.00	9.16	66.11
3	30.80	9.16	65.18
5	30.20	9.20	64.98
7	**28.80**	**8.43**	**66.31**

(b) Grounding signal

Signal	CHAIRs	CHAIRi	Rec.
Vanilla	55.20	16.18	71.4
Per-step TVER signal	51.60	14.60	76.9
Per-step TVER-selected heads	51.20	**13.20**	**77.4**
EIC	**48.80**	13.55	76.7

Table 6(b) complements the main ablation by reporting CHAIRi and recall for the same input-local TVER variants. Both improve over vanilla, but offline EIC yields the lowest CHAIRs at comparable recall.

As shown in Table 6(a), InternVL CHAIRs decreases from 31.0 at $K = 1$ to 30.8 at $K = 3$, 30.2 at $K = 5$, and 28.8 at $K = 7$, while recall at $K = 7$ remains essentially unchanged from $K = 1$. This supports the use of complementary image- and text-side calibration conditions.

C.2 Reusing EIC within ONLY

To separate offline head selection from ROAM's temperature controller, we replace ONLY's per-input TVER selection with the fixed offline EIC head set while leaving the remainder of ONLY's decoding pipeline and hyperparameters unchanged. Table 7 reports the result.

Table 7: **EIC head selection within ONLY on LLaVA-v1.5-7B.** Only the head-selection rule is changed in ONLY+EIC. The improvement with ONLY+EIC shows that the calibrated head set remains useful under a different downstream intervention.

Method	CHAIRs↓	CHAIRi↓	Rec.↑	Hal↓	POPE↑	MME↑
Vanilla	55.20	16.18	71.4	48.0	84.63	523.33
ONLY	52.40	15.13	73.5	35.6	87.97	**631.66**
ROAM	48.80	13.55	76.7	36.4	89.03	523.33
ONLY+EIC	**47.60**	13.90	75.0	36.9	88.87	**631.66**

Replacing ONLY's online TVER selection with EIC reduces CHAIRs from 52.4 to 47.6 while leaving the remainder of ONLY's decoding pipeline unchanged. This shows that the calibrated EIC head set remains useful under a different downstream intervention, supporting offline head selection as a transferable component beyond ROAM's temperature controller.

C.3 Layer and Controller Sensitivity

We additionally test sensitivity to the two main choices fixed in the reported experiments: the monitoring layer and controller strength. Figure 4 varies the monitored layer on InternVL, while Figure 5 varies the sharpening strength α on LLaVA. The EIC-selected layer $\tilde{\ell} = 1$ achieves the lowest CHAIRs among the tested layers, while stronger sharpening over the evaluated α range reduces hallucination with recall remaining above vanilla.

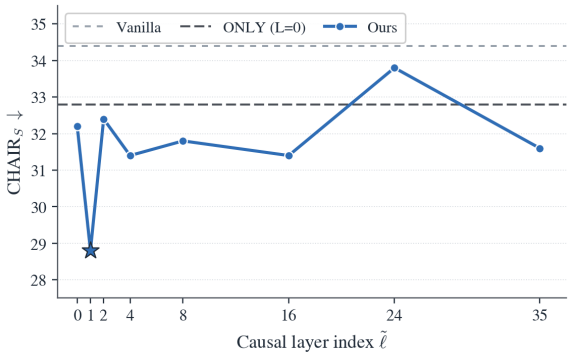


Figure 4: **Monitoring-layer sensitivity on InternVL3.5-8B.** The EIC-selected layer $\tilde{\ell} = 1$ obtains the lowest CHAIRs (28.8) among the eight tested layers.

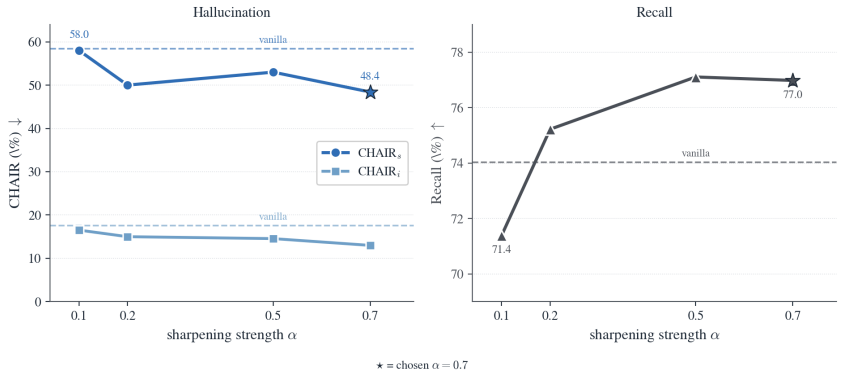


Figure 5: **Sharpening-strength sensitivity on LLaVA-v1.5-7B.** Hallucination decreases over the tested α range while recall remains above vanilla. All main experiments use the single global setting $\alpha = 0.7$.

D Additional Evaluation and Failure Analysis

D.1 Additional Capability Metrics

We assess broader multimodal capability using METEOR (Banerjee & Lavie, 2005), a CLIPScore-based alignment metric (Hessel et al., 2022), MMBench (Liu et al., 2024c), and MMVP (Tong et al., 2024). Table 8 reports the complementary capability results.

Table 8: **Additional general-capability metrics.** Higher is better. METEOR is computed on the 500 CHAIR captions against COCO val2014 references. MMB-single denotes standard MMBench accuracy and MMVP-p denotes paired MMVP accuracy. Across the evaluated LLaVA and Qwen3 settings, ROAM shows no broad degradation on these complementary capability metrics.

Model	Method	METEOR	CLIP-A	MMB-single	MMVP-p
LLaVA	Vanilla	.1652	.2789	72.0	10.67
	VCD	.1752	.2813	72.8	10.67
	ONLY	.1697	.2819	73.1	12.00
	ASCD	.1801	**.2832**	72.15	**14.67**
	ROAM	**.1816**	.2819	**73.5**	11.33
Qwen3	Vanilla	.138	**.290**	**89.7**	**62.00**
	VCD	.136	**.290**	89.1	57.33
	ONLY	.136	.286	65.1	42.00
	ROAM	.138	**.290**	89.4	60.00

On LLaVA, ASCD obtains the highest CLIP-A and paired-MMVP scores, while ROAM obtains the highest standard MMBench accuracy. On Qwen3, ROAM remains close to vanilla across the reported metrics, whereas ONLY shows substantially lower MMBench and MMVP performance. On the 500-image CHAIR caption set, ROAM increases LLaVA METEOR from 0.1652 to 0.1816 relative to vanilla decoding. On Qwen3, both methods score 0.138 at the reported three-decimal precision.

D.2 Statistical Uncertainty and Inference Cost

The main paper reports the primary long-form gains and the near-vanilla inference cost of ROAM on LLaVA. Here we complement those results with paired uncertainty estimates for the long-form hallucination metrics and detailed latency and peak-memory measurements.

Table 9: **Statistical uncertainty and detailed inference cost. (a)** Percentile 95% confidence intervals from $10{,}000$ paired image-level bootstrap resamples, with $\Delta = \text{ROAM} - \text{Vanilla}$; p-values are from two-sided exact McNemar tests on paired image-level binary outcomes. **(b)** Detailed LLaVA latency and peak-memory comparison.

(a) Long-form uncertainty

Metric	Model	95% CI for Δ	p
CHAIRs	LLaVA	$[-11.60, -1.20]$	.016
	Qwen3	$[-6.20, +1.40]$	.227
	InternVL	$[-9.70, -1.50]$	.008
AMBER Hal	LLaVA	$[-15.44, -7.87]$	$< 10^{-4}$
	Qwen3	$[-6.67, -0.80]$	.014
	InternVL	$[-5.68, +1.29]$	.236

(b) LLaVA inference cost

Method	Lat. (s)	Relative	Mem. (GB)
Vanilla	3.32	1.00×	13.96
VCD	6.51	1.96×	14.70
M3ID	6.34	1.91×	14.13
ONLY	3.37	1.02×	13.96
ASCD	9.40	2.77×	14.38
ROAM	3.30	0.99×	13.96

The uncertainty analysis supports the CHAIRs reductions on LLaVA and InternVL and the AMBER-Hal reductions on LLaVA and Qwen3. On LLaVA, ROAM achieves these gains with essentially vanilla measured latency and memory, whereas VCD and M3ID require approximately twice the measured latency. Together, these results support the statistical robustness of the main long-form improvements while preserving the practical efficiency of single-pass decoding.



Figure 6: **Representative qualitative successes on LLaVA-v1.5-7B.** Compared with vanilla decoding, ROAM avoids unsupported objects and scene details while retaining the central image-supported description.

D.3 Qualitative Successes and Failure Cases

Figure 6 shows representative cases in which ROAM avoids unsupported details present under vanilla decoding while preserving image-supported content. For example, in the portrait case, vanilla introduces unsupported details such as a *long, elegant gown* and a *clock on the wall*, whereas ROAM remains focused on the visible portrait and background. Figure 7 illustrates the complementary limitation: the grounding score measures concentration of image-token attention rather than verifying that the attended visual evidence supports the exact generated label.

In the kitchen example, the hallucinated word *microwave* has $\mathrm{GS} = 0.058$, close to the grounded word *oven* at $\mathrm{GS} = 0.064$. In the cat example, the hallucinated word *table* has $\mathrm{GS} = 0.057$, close to the grounded word *bed* at $\mathrm{GS} = 0.060$. Thus, ROAM can remain visually focused while the model assigns an incorrect semantic label to the attended content. This is consistent with treating GS as an online grounding proxy rather than a token-level verifier.

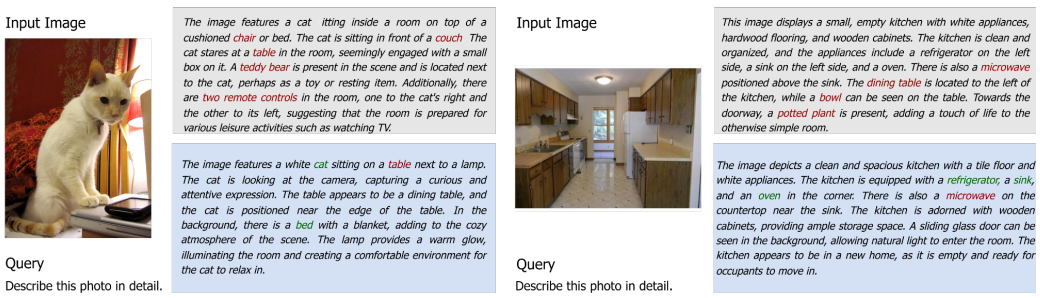


Figure 7: **Representative failure cases on LLaVA-v1.5-7B.** ROAM omits several unsupported mentions that appear in the vanilla responses but does not eliminate every hallucination.