TRACE: TRACING REPRESENTATION ANCESTRY FOR HALLUCINATION MITIGATION IN LARGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) often generate fluent but visually unsupported content. Prior studies commonly explain such hallucinations through reduced visual attention or increased reliance on textual context, implicitly associating information with the modality of the token positions that carry it. However, Transformer layers repeatedly mix information across positions, allowing visual evidence to propagate into question and previously generated tokens. As a result, token modality does not always reveal where information originally comes from. We introduce TRACE, a provenance-based framework for analyzing and mitigating hallucination in LVLMs. TRACE traces information back to its source by propagating contributions from visual inputs, user queries, and generation history through intermediate representations. This analysis reveals a recurring pattern before hallucination, which we call *provenance takeover*: generation-history provenance becomes increasingly dominant over visual provenance as the model approaches a hallucinated token. Building on this observation, TRACE detects abnormal takeover during decoding and reinforces the corresponding visual-origin component of the current representation. This enables TRACE to intervene selectively when the model begins to drift away from visual evidence, without requiring additional training. Across many benchmarks, TRACE consistently reduces hallucination-related errors across LVLMs while largely preserving overall multimodal performance with modest inference-time overhead. These results establish information provenance as a useful perspective for understanding hallucination and a practical signal for improving visual grounding in LVLMs.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have become a powerful interface for visual understanding, supporting tasks from visual question answering to open-ended image description Yin et al. (2024); Caffagni et al. (2024). Despite their strong multimodal capabilities, LVLMs remain prone to *hallucination*, where the model generates fluent and plausible content that is not supported by the visual input Liu et al. (2024a). A common explanation is that visual evidence becomes less influential during generation while the model increasingly relies on linguistic priors or its previously generated context Favero et al. (2024); Wu et al. (2025); Zhou et al. (2025). This raises a fundamental question: *what information actually drives the model's internal computation when hallucination occurs?*

Existing training-free methods approach this problem from several directions. Some modify the decoding distribution to suppress language-prior-driven predictions Leng et al. (2024); Huang et al. (2024). Others intervene inside the Transformer by strengthening visual interactions, identifying visually important components, or modifying intermediate representations Chefer et al. (2021); Yüksekgönül et al. (2024); Jiang et al. (2025); Zheng & Zhang (2025); Jung et al. (2025); Cheng et al. (2026). These methods are effective, but they often characterize visual influence through quantities associated with the *current token positions* that carry information, such as attention to image tokens or activations at visually relevant positions. This creates an important ambiguity: *the position currently carrying information does not necessarily reveal where that information originally came from.*

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.

This distinction matters because Transformer layers repeatedly mix information across token positions. Visual information can propagate into textual and previously generated representations and later influence predictions through those positions Kobayashi et al. (2021); Ferrando et al. (2022). Recent analyses further show that textual query representations can retain substantial visual information after cross-modal interaction, indicating that visual content need not remain localized to image-token positions Kaduri et al. (2025). For example, a generated token may interact only weakly with image-token positions while still receiving information that originated from the image in earlier layers. Conversely, a strong contribution from generation-history positions does not necessarily imply purely linguistic dependence, since those positions may themselves carry visual information. Token modality therefore identifies the current carrier of information, but not its original source.

Motivated by this limitation, we introduce an **information-provenance perspective** for studying LVLM hallucination. We separate the information available during autoregressive generation into three sources: the visual input $\mathcal{V}$, textual context $\mathcal{Q}$, and previously generated history $\mathcal{G}$. We then introduce **TRACE**, which propagates source-wise contributions through Transformer layers and tracks how much of the current computation can be traced back to each source. This provenance view follows visual information even after it has migrated into textual or generated representations.

Our analysis reveals a characteristic pattern before hallucination. Visual-origin information does not necessarily disappear from the model. Instead, generation-history provenance becomes increasingly dominant relative to visual provenance as the model approaches a hallucinated token. We call this phenomenon **provenance takeover**. While history-origin contribution can naturally increase as the autoregressive prefix grows, hallucinated predictions exhibit a much sharper transition toward generation-history dominance. This behavior suggests that hallucination can arise even when visual information remains internally available, but no longer exerts sufficient influence on the current computation.

This observation leads directly to our mitigation strategy. During decoding, TRACE monitors changes in provenance and identifies layers and generation steps that exhibit abnormal takeover. It then recovers the visual-origin contribution already present in the current forward pass and selectively reinforces it according to the detected takeover strength. In this way, TRACE restores the balance between visual and generation-history provenance without additional training or external supervision.

Across 3 complementary benchmarks (CHAIR, MMHal-Bench and AMBER), TRACE consistently reduces hallucination-related errors across LVLMs while largely preserving general multimodal performance with modest inference-time overhead. These results show that visual grounding depends not only on where information is currently represented, but also on where that information originated and how its influence evolves during generation.

We summarize our contributions as follows: **(1)** We introduce an **information-provenance framework** that separates the original source of information from the token positions that currently carry it, and traces contributions from visual input, textual context, and generation history across Transformer layers. **(2)** Using this framework, we identify **provenance takeover**, a recurring pattern before hallucination in which generation-history provenance becomes abnormally dominant relative to visual provenance while visual-origin information remains present. **(3)** We propose a **training-free provenance-guided intervention** that detects takeover during decoding and reinforces the corresponding visual-origin contribution. Across CHAIR, MMHal-Bench and AMBER, the resulting method consistently reduces hallucination-related errors while largely preserving general multimodal performance.

2 Related Work

Hallucination in LVLMs. Hallucination remains a major limitation of LVLMs, causing models to generate objects, attributes, or relations that are not supported by the visual input Liu et al. (2024a). Existing mitigation methods include training-based alignment and inference-time approaches, with the latter avoiding additional model optimization. Training-free methods commonly modify either the output distribution or the model's internal computation. Decoding-based methods such as VCD Leng et al. (2024) and OPERA Huang et al. (2024) suppress visually unsupported predictions through contrastive or attention-aware decoding. Other methods intervene directly on internal rep-

resentations. PAI Liu et al. (2024b) increases the influence of image tokens, while VTI Liu et al. (2025) modifies latent representations to improve visual stability. VISTA Li et al. (2025) further exploits internal visual-information dynamics to reinforce visually grounded predictions. These methods generally identify visual influence through image-token positions or visually relevant internal components. In contrast, TRACE tracks the *origin* of information after it has propagated across token positions. This allows the intervention to follow visual information even when it is carried by textual or previously generated representations.

Information Flow in Transformers. A related line of work studies how information propagates and mixes inside Transformers. Kobayashi et al. (2020) show that attention weights alone do not fully characterize token contributions and incorporate the transformed value vectors into their analysis. Their later work extends this formulation to residual and normalization operations Kobayashi et al. (2021). Ferrando et al. Ferrando et al. (2022) trace contextual information mixing across layers using token-level contribution matrices, while Mohebbi et al. Mohebbi et al. (2023) further study context mixing across Transformer components. These works provide tools for measuring how information moves between token positions, but do not distinguish the original source of information once representations have mixed. TRACE builds on this contribution-based view and introduces source-wise provenance for autoregressive LVLMs. We trace information originating from the visual input, textual context, and generation history across layers, and use changes in their relative influence to characterize and mitigate hallucination.

3 Understanding Provenance Takeover

3.1 Preliminaries: Layer-wise Token Contributions

Following Kobayashi et al. (2021), we express multi-head self-attention in terms of token-wise contributions to the output representation.[1] Consider token representations $X = [x_1, \ldots, x_T] \in \mathbb{R}^{d\times T}$, where $x_i \in \mathbb{R}^d$ denotes the representation at position i. For head h, with head dimension $d_H = d/H$, the query, key, and value vectors are $q_i^h = W_Q^h x_i$, $k_i^h = W_K^h x_i$, and $v_i^h = W_V^h x_i$. The causal attention weight from position j to position i is

$$A_{i,j}^h = \frac{\exp\left(\langle q_i^h, k_j^h\rangle/\sqrt{d_H} + M_{i,j}\right)}{\sum_{t=1}^T \exp\left(\langle q_i^h, k_t^h\rangle/\sqrt{d_H} + M_{i,t}\right)}, \tag{1}$$

where $M_{i,j} = -\infty$ for $j > i$ and 0 otherwise. Including the residual connection, the attention-block output at position i is

$$y_i = x_i + \sum_{j=1}^T \sum_{h=1}^H W_O^h A_{i,j}^h W_V^h x_j. \tag{2}$$

We isolate the contribution transmitted from position j to position i as

$$a_{j\to i} = \mathbf{1}\{j = i\} x_i + \sum_{h=1}^H W_O^h A_{i,j}^h W_V^h x_j, \tag{3}$$

such that $y_i = \sum_{j=1}^T a_{j\to i}$. Following the contribution measure used in prior attribution analysis, we define the normalized contribution from j to i as

$$C_{i,j} = \frac{\max\left(0,\ \|y_i\|_1 - \|y_i - a_{j\to i}\|_1\right)}{\sum_{j'=1}^T \max\left(0,\ \|y_i\|_1 - \|y_i - a_{j'\to i}\|_1\right)}. \tag{4}$$

Applying equation 4 at every Transformer layer gives $C^{(\ell)} \in \mathbb{R}^{T\times T}$, where $C_{i,j}^{(\ell)}$ measures the local contribution from position j at layer $\ell - 1$ to position i at layer ℓ.[2]

[1] We focus on cross-position information transport. Token-wise operations such as normalization and the MLP transform representations locally but do not introduce new cross-token interactions.

[2] Under causal attention, $C_{i,j}^{(\ell)} = 0$ for $j > i$, so $C^{(\ell)}$ is lower triangular.

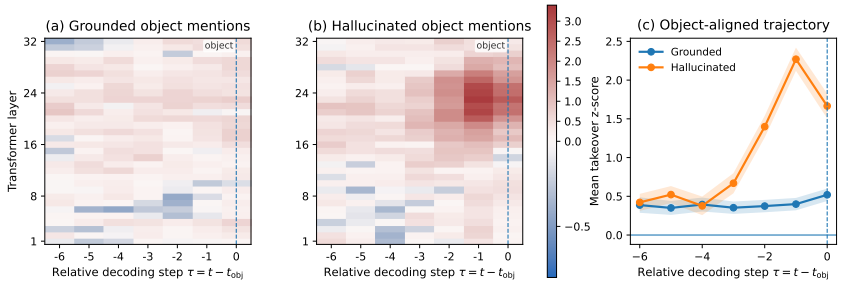


Figure 1: **Provenance takeover precedes hallucinated object generation.** Grounded mentions remain close to the reference provenance pattern, whereas hallucinated mentions exhibit a strong history-over-visual deviation in middle-to-late layers immediately before the object is generated.

However, $C_{i,j}^{(\ell)}$ only identifies the position through which information is transferred at the current layer. A textual position may have already absorbed visual information in earlier layers, yet its later contribution would still be assigned to that textual position by $C^{(\ell)}$. Therefore, the current token position does not necessarily reveal the original source of the information. Motivated by this limitation, we explicitly track the source of information as it propagates across Transformer layers.

3.2 Source-wise Contribution Provenance

To track where information originates as it moves through the model, we assign each input token to an initial source and propagate these source assignments through the layer-wise contribution matrices.

At decoding step t, let the current sequence contain T_t positions, partitioned into visual tokens $\mathcal{V}$, textual-context tokens $\mathcal{Q}$, and previously generated tokens $\mathcal{G}_t$. We define the source set $\mathcal{S} = \{V, Q, G\}$ and initialize $S^{(0)} \in \mathbb{R}^{T_t \times 3}$ as

$$S_{i,s}^{(0)} = \begin{cases} 1, & i \in \mathcal{V},\ s = V, \\ 1, & i \in \mathcal{Q},\ s = Q, \\ 1, & i \in \mathcal{G}_t,\ s = G, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Each position then passes on the source composition it has accumulated from earlier layers. We propagate this provenance through the Transformer as

$$S^{(\ell)} = C^{(\ell)} S^{(\ell-1)}, \qquad \ell = 1, \ldots, L, \tag{6}$$

or, for a single position i,

$$S_{i,:}^{(\ell)} = \sum_{j=1}^{T_t} C_{i,j}^{(\ell)} S_{j,:}^{(\ell-1)}. \tag{7}$$

Here, $C_{i,j}^{(\ell)}$ determines how strongly position j contributes to position i, while $S_{j,:}^{(\ell-1)}$ records the sources already carried by position j. Because the rows of $C^{(\ell)}$ are normalized and $S^{(0)}$ is one-hot, the source weights remain normalized at every layer:

$$S_{i,V}^{(\ell)} + S_{i,Q}^{(\ell)} + S_{i,G}^{(\ell)} = 1. \tag{8}$$

We therefore use $S_{i,V}^{(\ell)}$, $S_{i,Q}^{(\ell)}$, and $S_{i,G}^{(\ell)}$ to measure the contribution attributed to the visual input, textual context, and generation history, respectively.

The key property of this propagation is that provenance follows information across token positions. If a textual position absorbs visual information at an early layer, its later contribution still carries visual provenance rather than being reassigned to the textual source. As a result, textual and generated positions can carry visual-origin contribution even when the model interacts only weakly with the original image-token positions. This lets us track how the balance between visual and generated sources evolves during decoding, which we use next to characterize *provenance takeover*.

3.3 Provenance Takeover in Hallucinated Generation

We next examine how source provenance changes before grounded and hallucinated object mentions. At decoding step t, let i_t denote the current decoding position and define $p_{t,s}^{(\ell)} = S_{i_t,s}^{(\ell)}$ for $s \in \{V, Q, G\}$. To measure the relative dominance of generation-history over visual provenance, we define

$$r_t^{(\ell)} = \log \frac{p_{t,G}^{(\ell)} + \epsilon}{p_{t,V}^{(\ell)} + \epsilon}. \tag{9}$$

A larger $r_t^{(\ell)}$ indicates stronger generation-history provenance relative to visual provenance.

Because the scale of $r_t^{(\ell)}$ varies across layers and decoding contexts, we normalize it against its recent trajectory. For the preceding window $\mathcal{W}_t = \{t - w, \ldots, t - 1\}$, we compute

$$z_t^{(\ell)} = \frac{r_t^{(\ell)} - \mathrm{Median}_{k \in \mathcal{W}_t} r_k^{(\ell)}}{\mathrm{MAD}_{k \in \mathcal{W}_t} \left(r_k^{(\ell)} \right) + \epsilon}. \tag{10}$$

Here, MAD denotes the median absolute deviation over $\mathcal{W}_t$ Rousseeuw & Croux (1993), and $\epsilon > 0$ ensures numerical stability. Larger positive $z_t^{(\ell)}$ indicates that generation-history provenance has become unusually dominant relative to visual provenance compared with the recent decoding trajectory. We activate takeover detection once w preceding decoding steps are available; before that, we set $\gamma_t^{(\ell)} = 0$.

Figure 1 compares this behavior for grounded and hallucinated object mentions, aligned by $\tau = t - t_{\mathrm{obj}}$. Grounded mentions maintain a nearly stable trajectory before the object is generated. In contrast, hallucinated mentions show a sharp rise as the model approaches the hallucinated object, with the strongest increase appearing immediately before generation. We refer to this pre-generation shift toward generation-history provenance as *provenance takeover*. Since the shift appears before the hallucinated token, it provides a useful signal for intervention during decoding.

4 Method

4.1 Online Takeover Detection

The analysis in Sec. 3.3 shows that provenance takeover emerges before hallucinated object generation. We use this behavior as an online signal during decoding. At each step t and layer ℓ, we compute the takeover deviation $z_t^{(\ell)}$ from equation 10 using only the preceding decoding window $\mathcal{W}_t$. We convert the takeover deviation into a continuous gate

$$\gamma_t^{(\ell)} = \mathrm{clip}\left(\frac{z_t^{(\ell)} - \kappa}{\beta},\, 0,\, 1 \right), \tag{11}$$

where κ controls when intervention begins and $\beta > 0$ controls how quickly its strength increases. When $z_t^{(\ell)} \leq \kappa$, we set $\gamma_t^{(\ell)} = 0$ and leave the representation unchanged. As the deviation grows beyond κ, $\gamma_t^{(\ell)}$ increases toward 1, allowing stronger correction for more pronounced takeover. The resulting gate adapts the intervention to the degree of takeover at each decoding step and layer.

4.2 Provenance Restoration

When provenance takeover is detected, we reinforce the visual-origin information reaching the current decoding position i_t. Visual information may no longer reside only in visual-token positions, since it can propagate through textual and previously generated tokens across layers. We therefore weight each incoming contribution by the visual provenance carried by its position:

$$u_{t,V}^{(\ell)} = \sum_{j=1}^{T_t} S_{j,V}^{(\ell-1)}\, a_{j \to i_t}^{(\ell)}. \tag{12}$$

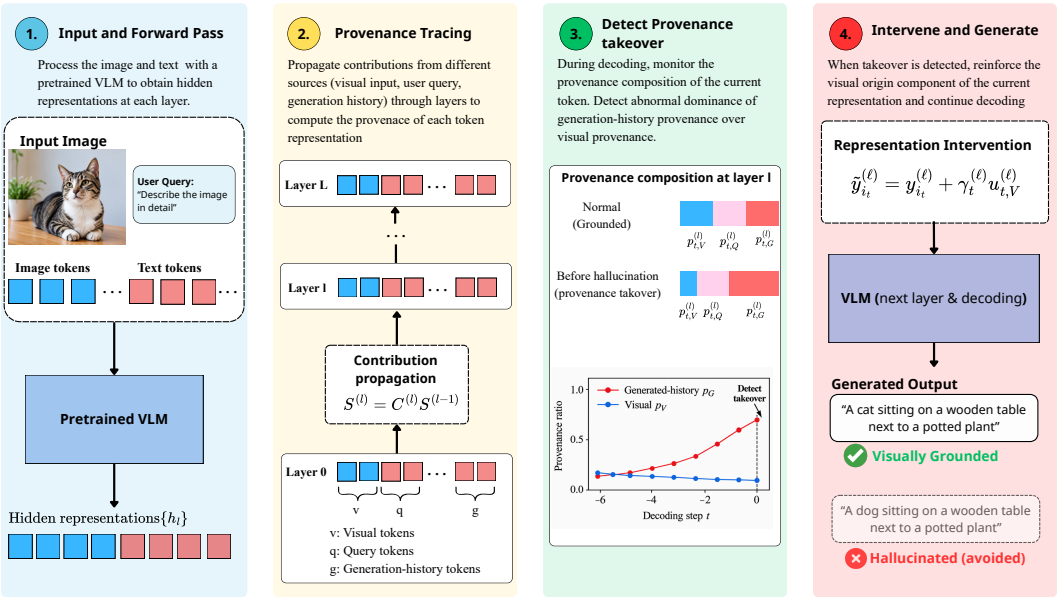


Figure 2: **Overview of our provenance tracing and intervention framework.** We trace visual, query, and generation-history contributions through the VLM, detect provenance takeover before hallucinated object generation, and reinforce the visual component of the hidden representation to maintain grounding.

Here, $S_{j,V}^{(\ell-1)}$ measures how much of the information at position j originates from the visual input. Thus, $u_{t,V}^{(\ell)}$ aggregates the visual-origin contribution reaching i_t at layer ℓ.

We then use the takeover gate $\gamma_t^{(\ell)}$ to control how much of this contribution is restored:

$$\tilde{y}_{i_t}^{(\ell)} = y_{i_t}^{(\ell)} + \gamma_t^{(\ell)} u_{t,V}^{(\ell)}. \tag{13}$$

When $\gamma_t^{(\ell)} = 0$, the representation remains unchanged. As the detected takeover becomes stronger, $\gamma_t^{(\ell)}$ increases and TRACE reinforces more visual-origin information before continuing the forward pass. By reinforcing the visual-origin contribution, TRACE restores the balance between visual and generation-history provenance during decoding.

5 Experiments

5.1 Models

We evaluate TRACE on three representative LVLMs: LLaVA-1.5-7B Liu et al. (2023), Qwen-VL-Chat-7B Bai et al. (2023), and Qwen2.5-VL-7B-Instruct Bai et al. (2025).

5.2 Benchmarks and Metrics

We evaluate TRACE on 3 benchmarks. CHAIR Rohrbach et al. (2018) measures object hallucination in image captioning using CHAIR_s and CHAIR_i. MMHal-Bench Sun et al. (2024) evaluates hallucination in open-ended multimodal responses using its official evaluation protocol. AMBER Wang et al. (2023) provides a unified evaluation of both generative and discriminative hallucination (we only use the generation result). We exclude POPE Li et al. (2023) and MME Fu et al. (2025) because they primarily consist of short yes/no questions, which provide insufficient generation history for our takeover detector.

Table 1: Results on the CHAIR benchmark. Max new tokens are set to 512. Lower C_S and C_I indicate fewer hallucinations, while higher Recall is better. Best results are **bolded** and second-best results are underlined.

	LLaVA-1.5-7B				Qwen-VL-Chat				Qwen2.5-VL-7B-Instruct			
Method	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Len.	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Len.	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Len.
Regular	52.5	16.1	77.1	93.8	49.9	14.4	76.6	92.5	31.2	8.4	69.4	94.7
VCD	51.2	15.1	77.4	101.5	47.6	13.5	75.6	98.4	30.4	8.1	68.8	101.2
OPERA	44.2	13.0	78.2	95.7	42.6	12.1	79.1	97.2	28.7	7.6	70.3	96.5
PAI	39.4	12.2	77.1	94.1	36.9	12.0	**82.0**	96.6	27.1	7.3	71.6	95.9
VTI	37.9	12.7	**79.0**	94.0	35.3	11.7	81.5	95.0	26.5	7.1	72.0	95.2
CRG	34.5	11.0	78.1	97.8	32.1	10.6	81.3	96.9	25.0	6.8	72.4	97.1
Devils	30.0	**9.0**	78.0	95.0	28.5	8.6	80.1	96.0	23.4	6.2	71.9	95.6
TRACE	**27.3**	**9.0**	78.6	96.5	**25.1**	**7.4**	81.5	96.6	**21.8**	**5.9**	**72.8**	96.1

Table 2: Results on the AMBER-Generative benchmark. Lower CHAIR, Hal, and Cog indicate fewer hallucinations, while higher Cover is better.

	LLaVA-1.5-7B				Qwen-VL-Chat				Qwen2.5-VL-7B-Instruct			
Method	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓	CHAIR↓	Cover↑	Hal↓	Cog↓
Regular	8.2	49.3	36.5	4.3	7.3	52.7	34.1	3.5	5.3	47.2	23.6	1.9
VCD	7.1	49.0	33.0	3.5	6.4	52.4	30.8	3.0	4.7	45.9	20.5	1.6
OPERA	6.7	49.5	31.0	3.0	6.0	52.8	28.9	2.7	5.1	46.2	21.6	1.7
PAI	6.6	48.9	42.0	2.0	5.6	52.3	30.2	2.2	4.8	46.6	21.0	1.5
VTI	5.8	48.5	27.1	2.4	5.2	51.8	25.6	2.1	4.6	46.4	20.0	1.4
CRG	4.7	53.1	23.4	2.1	4.5	53.4	22.7	1.8	4.3	47.1	19.2	1.3
Devils	4.2	51.8	**21.8**	1.7	4.1	53.0	21.0	**1.5**	4.1	46.9	**18.7**	1.2
TRACE	**4.0**	**53.6**	22.0	**1.6**	**3.9**	**53.6**	**20.7**	1.6	**3.9**	**47.8**	18.9	**1.1**

5.3 Baselines and Implementation Details

We compare TRACE with representative training-free hallucination mitigation methods, including VCD Leng et al. (2024), OPERA Huang et al. (2024), PAI Liu et al. (2024b), VTI Liu et al. (2025), Devils Jiang et al. (2025) and CRG Cheng et al. (2026). We use greedy decoding unless otherwise specified and keep the decoding configuration consistent across methods whenever applicable. For TRACE, takeover detection is activated once w preceding decoding steps are available, and the gate is evaluated independently at each Transformer layer. Detailed model configurations, method-specific hyperparameters, and decoding settings are provided in Appendix A.

5.4 Main Results

Results on CHAIR. Table 1 shows that TRACE reduces object hallucination across all three LVLMs. On LLaVA-1.5-7B, C_S decreases from 52.5 to 27.3 and C_I from 16.1 to 9.0. Similar reductions are observed on Qwen-VL-Chat and Qwen2.5-VL-7B-Instruct. Importantly, Recall is maintained or improved on all three models, reaching 78.6, 81.5, and 72.8, respectively, while the generated response length remains close to the regular decoding baseline.

Results on AMBER-Generative. Table 2 shows a similar pattern on AMBER-Generative. TRACE obtains the lowest CHAIR score and the highest Cover on all three backbones. For LLaVA-1.5-7B, CHAIR decreases from 8.2 to 4.0 while Cover increases from 49.3 to 53.6. On Qwen2.5-VL-7B-Instruct, CHAIR decreases from 5.3 to 3.9 with Cover increasing from 47.2 to 47.8. TRACE is also competitive on Hal and Cog, achieving either the best or second-best result across the three models.

Results on MMHal-Bench. Figure 3 reports performance across hallucination categories on MMHal-Bench. TRACE achieves the highest average score on both LLaVA-1.5 and Qwen-VL, and improves over Regular, VCD, and OPERA across most categories, including attributes, relations, counting, and holistic reasoning. The gains are therefore not restricted to object-existence hallucination.

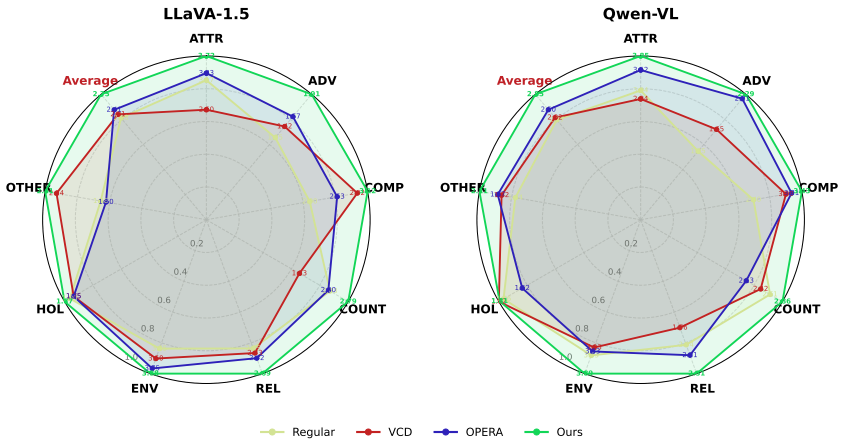


Figure 3: Per-category radar comparison on MMHal-Bench for two VLMs (LLaVA-1.5-7B and Qwen-VL-Chat).

Table 3: Ablation study on CHAIR with LLaVA-1.5-7B. Lower C_S and C_I are better; higher Recall is better. *Pos.* denotes position-based attribution and *Prov.* denotes propagated source provenance. In (a), the arrow denotes detection → restoration.

(a) Position vs. provenance

Variant	$C_S \downarrow$	$C_I \downarrow$	Rec.↑
Pos. → Pos.	36.8	11.8	77.7
Prov. → Pos.	33.7	10.9	78.0
Pos. → Prov.	31.6	10.3	78.1
Prov. → Prov.	**27.3**	**9.0**	**78.6**

(b) Takeover gating

Gate	$C_S \downarrow$	$C_I \downarrow$	Rec.↑
Always-on	30.5	10.0	76.8
Hard	28.9	9.6	78.2
Adaptive	**27.3**	**9.0**	**78.6**

5.5 Ablation Study and Hyperparameter Sensitivity

We ablate the main design choices of TRACE on CHAIR using LLaVA-1.5-7B. Detailed definitions of the ablated variants and implementation settings are provided in the Appendix A.3.

Ablation study. We first compare position-based and provenance-based detection and restoration. Results are shown in Table 3. Replacing current token positions with propagated provenance improves both stages, with the full provenance-based variant achieving the lowest C_S and C_I. We also compare always-on restoration, hard thresholding, and the continuous takeover gate. Always-on intervention reduces hallucination but slightly lowers Recall, whereas takeover-guided restoration provides a better balance between hallucination reduction and retained object coverage.

Hyperparameter sensitivity. Figure 4 studies the preceding window $\mathcal{W}_t = \{t-w, \ldots, t-1\}$, takeover threshold κ, and gate slope β. Performance is stable over a reasonably broad range of settings. For the history window, performance improves as more preceding steps are included and peaks around $w = 6$, while longer windows provide no further benefit. The detector is also relatively stable for moderate values of κ, with the best results around $\kappa = 2.5$. For β, intermediate values provide the best trade-off, while very small or large values make the gate respectively too aggressive or too gradual. We use $w = 6$, $\kappa = 2.5$, and $\beta = 1.0$ as the default configuration.

5.6 Complexity Analysis

TRACE operates within the original forward pass and does not require an additional model evaluation. During autoregressive decoding, provenance is propagated only for the current decoding position. Since the number of sources is fixed to three, updating the source state requires $O(T_t)$ additional operations per layer, while the provenance-weighted restoration in equation 12 requires a weighted aggregation over the T_t incoming contributions. The provenance state itself contains only

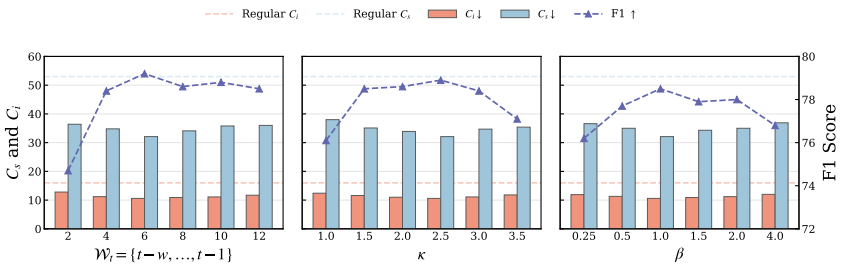


Figure 4: Hyperparameter sensitivity of TRACE on CHAIR with LLaVA-1.5-7B. We vary the history window size w, takeover threshold κ, and gate slope β. Lower C_s and C_i are better, while higher F1 is better.

Table 4: Runtime and peak GPU memory overhead under the same decoding setup. Values in parentheses denote the relative cost compared with regular decoding.

Method	Avg. Latency ↓	GPU Memory ↓
Regular	88.6 ms (×1.00)	14945 MB (×1.00)
VCD	181.5 ms (×2.05)	15749 MB (×1.05)
OPERA	618.9 ms (×6.99)	22706 MB (×1.52)
CRG	139.2 ms (×2.27)	14950 MB (×1.00)
TRACE	120.8 ms (×1.36)	15120 MB (×1.01)

three values per token and can be updated online, resulting in little additional memory beyond the model activations and KV cache.

Table 4 compares the practical inference cost with other training-free methods. TRACE introduces moderate latency relative to regular decoding while keeping peak GPU memory nearly unchanged. In contrast to methods that require additional decoding passes, the additional computation in TRACE consists only of contribution aggregation, provenance propagation, and gated restoration within the same forward pass.

6 Conclusion

We introduced TRACE, a training-free method for reducing hallucination in large vision-language models through source-level provenance tracking. By propagating visual, textual, and generated-source contributions across layers, TRACE identifies *provenance takeover*, where generation history becomes unusually dominant over visual evidence before a hallucinated mention. TRACE detects this shift online and selectively restores visual-origin contributions during decoding. Across CHAIR, AMBER-Generative, and MMHal-Bench, TRACE consistently reduces hallucination across multiple LVLM backbones while preserving visual-content coverage and response quality.

Limitations. TRACE relies on a window of preceding decoding steps to estimate provenance deviation, making it better suited to free-form generation. Very short outputs, such as yes/no answers, may not provide enough history for reliable detection. Extending the method to these settings may require prompt-based or learned reference statistics instead of generation history.

AI use statement

In this work, we used generative AI tools solely for polishing the manuscript for grammar, clarity, and readability. We have not used generative AI tools for research ideation, methodology development, experimental design, data analysis, interpretation of experimental results, or formulation of scientific conclusions, and the remaining required disclosure tasks are not applicable to this work. AI-assisted writing was carefully reviewed and revised before inclusion in the manuscript. We take responsibility for the final content of this work, including all text and claims.

Ethics statement

This work aims to reduce hallucination in large vision-language models by improving the interpretability and controllability of information provenance during generation. While the proposed method is intended to improve the reliability of model outputs, it does not eliminate hallucinations and may have limitations across different models and application domains. We therefore encourage responsible use, particularly in high-stakes settings where incorrect model outputs may lead to harmful consequences. All experiments use publicly available datasets and pretrained models and do not involve human subjects, personally identifiable information, or sensitive data.

Reproducibility statement

We provide the information necessary to reproduce our experiments throughout the paper and supplementary materials. Detailed implementation settings are given in Appendix A, while the complete experimental configurations and evaluation protocols are described in Appendix A.2. Additional implementation details for the ablation studies are provided in Appendix A.3. We also include the full source code in the supplementary materials to facilitate reproduction of the reported results.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *CoRR*, abs/2502.13923, 2025. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13590–13618. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.807. URL https://doi.org/10.18653/v1/2024.findings-acl.807.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 397–406. IEEE, October 2021. doi: 10.1109/ICCV48922.2021.00045. URL https://doi.org/10.1109/ICCV48922.2021.00045.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating. *arXiv preprint arXiv:2605.24024*, 2026.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14303–14312. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01356. URL https://doi.org/10.1109/CVPR52733.2024.01356.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8698–8714, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.595. URL https://aclanthology.org/2022.emnlp-main.595/.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=DgH9YCsqWm. Spotlight.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13418–13427. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01274. URL https://doi.org/10.1109/CVPR52733.2024.01274.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 25004–25014. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02328.

Mingi Jung, Saehyung Lee, Eunji Kim, and Sungroh Yoon. Visual attention never fades: Selective progressive attention ReCalibration for detailed image captioning in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28527–28551. PMLR, 2025. URL https://proceedings.mlr.press/v267/jung25c.html.

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

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7057–7075, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.574. URL https://aclanthology.org/2020.emnlp-main.574/.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Incorporating residual and normalization layers into analysis of masked language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 4547–4568, 2021.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13872–13882. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01316. URL https://doi.org/10.1109/CVPR52733.2024.01316.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253, 2024a. doi: 10.48550/ARXIV.2402.00253. URL https://doi.org/10.48550/arXiv.2402.00253.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=LBl7Hez0fF.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *Computer Vision – ECCV 2024*, pp. 125–140. Springer, 2024b. doi: 10.1007/978-3-031-73010-8_8. URL https://doi.org/10.1007/978-3-031-73010-8_8.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. Quantifying context mixing in transformers. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3378–3400, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.245. URL https://aclanthology.org/2023.eacl-main.245/.

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, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/V1/D18-1437. URL https://doi.org/10.18653/v1/d18-1437.

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

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13088–13110. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.775. URL https://doi.org/10.18653/v1/2024.findings-acl.775.

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. *CoRR*, abs/2311.07397, 2023. doi: 10.48550/arXiv.2311.07397. URL https://doi.org/10.48550/arXiv.2311.07397.

Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, Junjie Ao, Enyan Dai, Lei Feng, Xiang Zhang, and Suhang Wang. LanP: Rethinking the impact of language priors in large vision-language models. *CoRR*, abs/2502.12359, 2025. doi: 10.48550/ARXIV.2502.12359. URL https://doi.org/10.48550/arXiv.2502.12359.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. doi: 10.1093/nsr/nwae403. URL https://doi.org/10.1093/nsr/nwae403.

Mert Yüksekgönül, Varun Chandrasekaran, Erik Jones, Suriya Gunasekar, Ranjita Naik, Hamid Palangi, Ece Kamar, and Besmira Nushi. Attention satisfies: A constraint-satisfaction lens on factual errors of language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=gfFVATffPd.

Haohan Zheng and Zhenguo Zhang. Modality bias in LVLMs: Analyzing and mitigating object hallucination via attention lens. *CoRR*, abs/2508.02419, 2025. doi: 10.48550/ARXIV.2508.02419. URL https://doi.org/10.48550/arXiv.2508.02419.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=AV7OXVlAyi.

APPENDIX

A Implementation Details

A.1 Benchmarks and Evaluation

CHAIR. We evaluate free-form image descriptions using CHAIR (Rohrbach et al., 2018). Object mentions are mapped to the MS-COCO vocabulary using the official CHAIR synonym list and evaluated with the standard instance-level CHAIR_I and sentence-level CHAIR_S metrics. We additionally report object recall and average response length (*Len*) to verify that reductions in hallucination are not caused by suppressing visual content or producing substantially shorter responses.

AMBER-Generative. We use the generative split of AMBER (Wang et al., 2023), consisting of 1,004 annotated images. Generated responses are evaluated with the official AMBER evaluation pipeline. We report CHAIR, Cover, Hal, and Cog following the benchmark definitions. We use only the generative setting because TRACE operates on autoregressive generation trajectories.

MMHal-Bench. We evaluate open-ended visual question answering using MMHal-Bench (Sun et al., 2024). The benchmark contains 96 image–question pairs from eight categories: ATTR, ADV, COMP, COUNT, REL, ENV, HOL, and OTHER. We follow the official judge-based evaluation protocol and report the overall score, hallucination rate, and per-category scores.

A.2 Models and Inference Configuration

Backbones. We evaluate TRACE on LLaVA-1.5-7B (Liu et al., 2023), Qwen-VL-Chat-7B (Bai et al., 2023), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). All models use their publicly released pretrained checkpoints without fine-tuning or parameter updates. The tokenizer, conversation template, vision preprocessing, multimodal projector, and language-model parameters are kept unchanged. Model details are shown in Table 5.

Generation. Unless otherwise stated, all methods use greedy decoding. For each model–benchmark pair, all methods receive identical images and prompts and share the same generation configuration. CHAIR generations are limited to 512 new tokens. For AMBER-Generative and MMHal-Bench, we follow the corresponding benchmark generation protocols. Generation stops when an end-of-sequence token is produced or the maximum generation length is reached.

TRACE configuration. TRACE is implemented according to Sec. 4. Unless otherwise specified, we use a history-window length of

$$w = 6,$$

a takeover threshold of

$$\kappa = 2.5,$$

and a gate slope of

$$\beta = 1.0.$$

The same configuration is used across all evaluated examples without per-image or per-token tuning. Intervention is disabled during the first w generated tokens because the history window has not yet been populated. The numerical constant used in the MAD denominator is $\epsilon = 10^{-8}$.

Evaluation. We use the official evaluation implementations for CHAIR, AMBER-Generative, and MMHal-Bench. All methods are evaluated on the same benchmark examples for each backbone.

Table 5: Architectures of the LVLM backbones used in our experiments.

Model	Vision encoder	Language model
LLaVA-1.5-7B	CLIP ViT-L/14 (336px)	Vicuna-v1.5-7B
Qwen-VL-Chat-7B	OpenCLIP ViT-bigG	Qwen-7B
Qwen2.5-VL-7B-Instruct	ViT with 2D-RoPE and window attention	Qwen2.5-7B

A.3 Ablation Study Implementation Details

We conduct all ablation experiments on CHAIR using LLaVA-1.5-7B. All variants use the same pre-trained checkpoint, image preprocessing, prompt, decoding configuration, and evaluation protocol. We use greedy decoding with at most 512 newly generated tokens. Unless otherwise specified, the history-window length, takeover threshold, and adaptive gate slope are fixed to

$$w = 6, \qquad \kappa = 2.5, \qquad \beta = 1.0.$$

No model parameters are updated in any ablation.

The ablation in Table 3a changes two components independently: (1) the quantity used to detect takeover and (2) the visual component used for restoration. We denote a configuration by Detection $\rightarrow$ Restoration, where Pos. uses the current token positions and Prov. uses the propagated source provenance defined in Sec. 3.2.

Position-based takeover detection. Let i_t denote the current decoding position at step t, and let $\mathcal{V}$, $\mathcal{Q}$, and $\mathcal{G}_t$ denote the positions corresponding to visual tokens, textual-context tokens, and previously generated tokens, respectively. Recall that $C_{i,j}^{(\ell)}$ is the normalized contribution transmitted from position j at layer $\ell - 1$ to position i at layer ℓ.

For the position-based detector, we do not propagate source identity across layers. Instead, the visual and generation-history contributions at the current decoding position are obtained directly by summing the incoming contribution mass according to the *current positions*:

$$p_{t,V}^{(\ell),\mathrm{pos}} = \sum_{j \in \mathcal{V}} C_{i_t,j}^{(\ell)}, \qquad p_{t,G}^{(\ell),\mathrm{pos}} = \sum_{j \in \mathcal{G}_t} C_{i_t,j}^{(\ell)}. \tag{14}$$

The corresponding history-to-visual ratio is

$$r_{t,\mathrm{pos}}^{(\ell)} = \log \frac{p_{t,G}^{(\ell),\mathrm{pos}} + \epsilon}{p_{t,V}^{(\ell),\mathrm{pos}} + \epsilon}. \tag{15}$$

As in the full TRACE detector, this ratio is normalized relative to the preceding decoding trajectory. For

$$\mathcal{W}_t = \{t - w, \ldots, t - 1\},$$

we first compute

$$m_{t,\mathrm{pos}}^{(\ell)} = \mathrm{Median}_{k \in \mathcal{W}_t} \, r_{k,\mathrm{pos}}^{(\ell)}, \tag{16}$$

and

$$d_{t,\mathrm{pos}}^{(\ell)} = \mathrm{Median}_{k \in \mathcal{W}_t} \left| r_{k,\mathrm{pos}}^{(\ell)} - m_{t,\mathrm{pos}}^{(\ell)} \right|. \tag{17}$$

The position-based takeover score is then

$$z_{t,\mathrm{pos}}^{(\ell)} = \frac{r_{t,\mathrm{pos}}^{(\ell)} - m_{t,\mathrm{pos}}^{(\ell)}}{d_{t,\mathrm{pos}}^{(\ell)} + \epsilon}. \tag{18}$$

This detector therefore attributes information according to the positions that currently transmit it. For example, information arriving through a generated-token position is counted as generation-history influence even if that position had previously absorbed visual information.

Provenance-based takeover detection. The provenance-based detector instead uses the propagated source state $S^{(\ell)}$. At the current decoding position, we define

$$p_{t,V}^{(\ell),\mathrm{prov}} = S_{i_t,V}^{(\ell)}, \qquad p_{t,G}^{(\ell),\mathrm{prov}} = S_{i_t,G}^{(\ell)}. \tag{19}$$

The corresponding ratio is

$$r_{t,\mathrm{prov}}^{(\ell)} = \log \frac{p_{t,G}^{(\ell),\mathrm{prov}} + \epsilon}{p_{t,V}^{(\ell),\mathrm{prov}} + \epsilon}, \tag{20}$$

and the normalized takeover score is

$$z^{(\ell)}_{t,\mathrm{prov}} = \frac{r^{(\ell)}_{t,\mathrm{prov}} - \mathrm{Median}_{k\in\mathcal{W}_t}\, r^{(\ell)}_{k,\mathrm{prov}}}{\mathrm{MAD}_{k\in\mathcal{W}_t}\left(r^{(\ell)}_{k,\mathrm{prov}}\right) + \epsilon}. \tag{21}$$

Unlike equation 14, this formulation retains the original information source after information has moved between token positions.

Position-based visual restoration. Let $a^{(\ell)}_{j\to i_t}$ denote the vector contribution transmitted from position j to the current decoding position i_t at layer ℓ. For position-based restoration, we aggregate only contributions arriving directly from visual-token positions:

$$u^{(\ell),\mathrm{pos}}_{t,V} = \sum_{j\in\mathcal{V}} a^{(\ell)}_{j\to i_t}. \tag{22}$$

Given a gate $\gamma^{(\ell)}_t$, the layer output is modified as

$$\widetilde{y}^{(\ell)}_{i_t} = y^{(\ell)}_{i_t} + \gamma^{(\ell)}_t u^{(\ell),\mathrm{pos}}_{t,V}. \tag{23}$$

Thus, only information currently transmitted from image-token positions is restored.

Provenance-based visual restoration. For provenance-based restoration, each incoming vector contribution is weighted by the amount of visual-origin information carried by its source position:

$$u^{(\ell),\mathrm{prov}}_{t,V} = \sum_{j=1}^{T_t} S^{(\ell-1)}_{j,V} a^{(\ell)}_{j\to i_t}. \tag{24}$$

The intervention becomes

$$\widetilde{y}^{(\ell)}_{i_t} = y^{(\ell)}_{i_t} + \gamma^{(\ell)}_t u^{(\ell),\mathrm{prov}}_{t,V}. \tag{25}$$

Consequently, visual-origin information can be restored even when it is currently transmitted through textual or previously generated token positions.

Position/provenance combinations. Using the quantities above, the four variants in Table 3a are implemented as follows:

$$\text{Pos.} \to \text{Pos.}: \quad z^{(\ell)}_t = z^{(\ell)}_{t,\mathrm{pos}}, \qquad u^{(\ell)}_{t,V} = u^{(\ell),\mathrm{pos}}_{t,V}, \tag{26}$$

$$\text{Prov.} \to \text{Pos.}: \quad z^{(\ell)}_t = z^{(\ell)}_{t,\mathrm{prov}}, \qquad u^{(\ell)}_{t,V} = u^{(\ell),\mathrm{pos}}_{t,V}, \tag{27}$$

$$\text{Pos.} \to \text{Prov.}: \quad z^{(\ell)}_t = z^{(\ell)}_{t,\mathrm{pos}}, \qquad u^{(\ell)}_{t,V} = u^{(\ell),\mathrm{prov}}_{t,V}, \tag{28}$$

$$\text{Prov.} \to \text{Prov.}: \quad z^{(\ell)}_t = z^{(\ell)}_{t,\mathrm{prov}}, \qquad u^{(\ell)}_{t,V} = u^{(\ell),\mathrm{prov}}_{t,V}. \tag{29}$$

The final configuration corresponds to the complete TRACE method.

Takeover-gating ablation. For Table 3b, takeover detection and visual restoration are fixed to their provenance-based forms, $z^{(\ell)}_{t,\mathrm{prov}}$ and $u^{(\ell),\mathrm{prov}}_{t,V}$, respectively. We vary only the function that maps the takeover score to the intervention strength $\gamma^{(\ell)}_t$.

For all takeover-based variants, intervention is disabled until the history window contains w preceding decoding steps:

$$\gamma^{(\ell)}_t = 0, \qquad t \le w. \tag{30}$$

Always-on. To remove takeover-dependent selection, we apply the same provenance-based visual restoration with constant strength:

$$\gamma^{(\ell)}_{t,\mathrm{always}} = \begin{cases} 0, & t \le w, \\ 1, & t > w. \end{cases} \tag{31}$$

The resulting intervention is therefore

$$\widetilde{y}_{i_t}^{(\ell)} = y_{i_t}^{(\ell)} + u_{t,V}^{(\ell),\mathrm{prov}}, \qquad t > w. \tag{32}$$

This variant preserves the same restoration direction as TRACE but removes dependence on the magnitude of provenance takeover.

Hard gate. The hard-gating variant converts the takeover score into a binary decision:

$$\gamma_{t,\mathrm{hard}}^{(\ell)} = \begin{cases} 0, & t \le w, \\ \mathbb{I}\left[z_{t,\mathrm{prov}}^{(\ell)} > \kappa\right], & t > w. \end{cases} \tag{33}$$

Hence, once the takeover score crosses the threshold, the complete visual-origin component is restored:

$$\widetilde{y}_{i_t}^{(\ell)} = y_{i_t}^{(\ell)} + \mathbb{I}\left[z_{t,\mathrm{prov}}^{(\ell)} > \kappa\right] u_{t,V}^{(\ell),\mathrm{prov}}. \tag{34}$$

Adaptive gate. The full TRACE configuration uses a continuous intervention strength:

$$\gamma_{t,\mathrm{adapt}}^{(\ell)} = \begin{cases} 0, & t \le w, \\ \mathrm{clip}\left(\dfrac{z_{t,\mathrm{prov}}^{(\ell)} - \kappa}{\beta}, 0, 1\right), & t > w. \end{cases} \tag{35}$$

The corrected representation is

$$\widetilde{y}_{i_t}^{(\ell)} = y_{i_t}^{(\ell)} + \gamma_{t,\mathrm{adapt}}^{(\ell)} u_{t,V}^{(\ell),\mathrm{prov}}. \tag{36}$$

Unlike hard gating, this formulation distinguishes weak and strong takeover events: no intervention is applied for $z_t^{(\ell)} \le \kappa$, the intervention increases linearly for $\kappa < z_t^{(\ell)} < \kappa + \beta$, and full restoration is reached when $z_t^{(\ell)} \ge \kappa + \beta$.

Implementation procedure. At every decoding step t and Transformer layer ℓ, we first compute the incoming contribution vectors $a_{j \to i_t}^{(\ell)}$ and their normalized contribution weights $C_{i_t,j}^{(\ell)}$. The detector then computes either $z_{t,\mathrm{pos}}^{(\ell)}$ or $z_{t,\mathrm{prov}}^{(\ell)}$, depending on the ablation. The chosen gating rule converts this score to $\gamma_t^{(\ell)}$. Finally, the restoration module constructs either $u_{t,V}^{(\ell),\mathrm{pos}}$ or $u_{t,V}^{(\ell),\mathrm{prov}}$ and modifies the current representation before it is passed to the subsequent Transformer computation. All other components of the forward pass are unchanged.