TRACING THE EVIDENCE: FAITHFUL TOKEN ATTRIBUTION THROUGH VISION-LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) exhibit strong reasoning capabilities, yet the visual and textual evidence supporting the generated responses remains difficult to identify. Faithful token attribution explains an LVLM's response by assigning scores that rank image and prompt tokens by how much the model relies on them, such that removing higher-ranked tokens causes the likelihood of the generated response to drop more rapidly. However, existing token-attribution methods have been developed mainly for text-based language models, and our empirical study reveals two challenges when complex multimodal sources are involved. First, the joint image-text attribution can underrepresent visual evidence relative to text, obscuring the image regions supporting the response. Second, visual evidence may influence the generated response through multiple intermediate reasoning paths, while existing methods trace only a limited subset of these paths, causing important visual contributions to be underestimated. Motivated by these insights, we introduce VTRACE, a multimodal token-attribution framework that traces input contributions through intermediate reasoning and calibrates attribution scores across modalities. VTRACE constructs pairwise attributions that highlight token-specific contributions and aggregates all forward attribution paths in closed form to account for both direct and indirect contributions. Cross-modal calibration then rescales image and text attribution scores using modality contributions estimated from response-likelihood changes, enabling a unified ranking of input tokens. Evaluations against seven baselines across six visual reasoning benchmarks demonstrate the superior attribution faithfulness. A demo [1] and a reference code [2] are available.

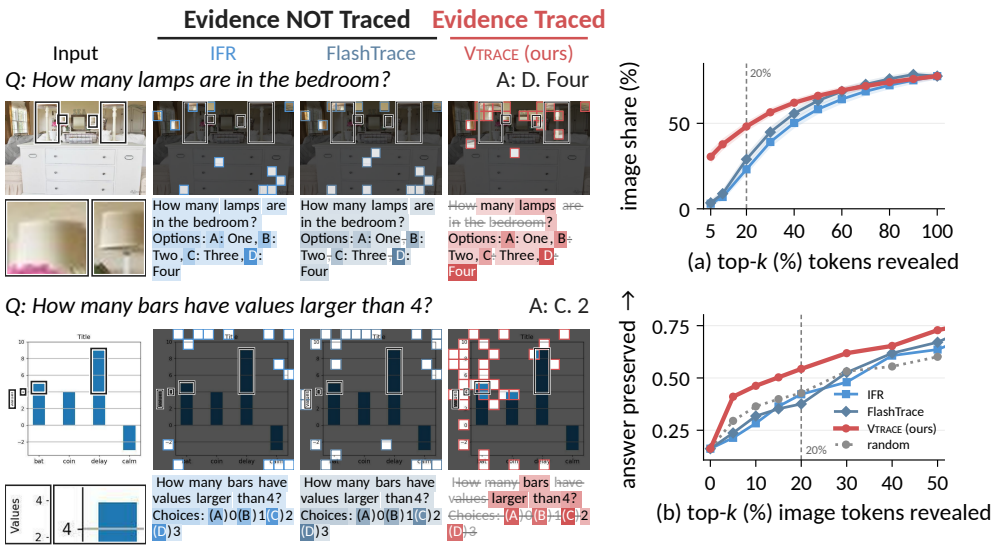


Figure 1: **Relevant visual evidence can be under-ranked in multimodal attribution. Left:** Top-20% attributed tokens in two examples, with color intensity indicating attribution strength. Black boxes mark the question-relevant region, enlarged below. **Right:** (a) percentage of attributed image tokens and (b) normalized answer probability recovery across different selected token budgets.

[1] https://huggingface.co/spaces/VTrace/vtrace-demo
[2] https://anonymous.4open.science/r/VTRACE-CB12

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.

1 INTRODUCTION

Large vision-language models (LVLMs) are increasingly capable of solving complex visual reasoning tasks by generating intermediate reasoning before arriving at a final answer (Xu et al., 2025; Chen et al., 2024b; Zhang et al., 2024b). However, observing the answer alone does not reveal which visual and textual evidence contributes to the answer, or how such evidence is used throughout the reasoning process (Stan et al., 2024; Shen, 2025). Explaining these predictions requires identifying how input evidence contributes to predictions throughout the reasoning (Pan et al., 2026; Uppaal et al., 2026; Deng et al., 2025). Token attribution provides a way to estimate these contributions: for a generated token or a target span, token attribution assigns preceding tokens an *attribution score* reflecting the contribution to that output (Abnar & Zuidema, 2020; Ferrando et al., 2022b; Achtibat et al., 2024). A faithful attribution should accurately reflect how input evidence contributes to the selected tokens.

Existing attribution methods typically quantify token contributions using either model-internal signals or behavioral changes under intervention. Internal-signal-based methods trace information propagation through transformer attention or token interactions to attribute predictions to preceding tokens (Abnar & Zuidema, 2020; Ferrando et al., 2022b; Achtibat et al., 2024), whereas perturbation-based methods measure how modifying input tokens changes the model's output distribution (Zhao & Shan, 2024). More recent methods extend attribution to the reasoning process, with FlashTrace (Pan et al., 2026) tracing influence through reasoning tokens and FlowTracer (Dong et al., 2026) modeling answer-directed information flow across the reasoning trace.

Despite the advances, attribution in LVLM reasoning presents two challenges. **1) Visual evidence generally receives lower attribution scores when image and text tokens are scored together.** Our empirical analysis reveals an imbalance in joint image-text attribution, where visual evidence are underrepresented relative to text. Illustrated in Figure 1, textual tokens such as answer options often receive higher attribution scores than image patches with question-relevant visual cues. Consequently, relevant image patches can be ranked below less relevant text tokens, obscuring critical visual evidence that supports the model's reasoning. **2) Visual evidence can be underestimated when only some of the reasoning paths to the target are traced.** Relevant visual evidence may support the target through multiple intermediate reasoning tokens. Shown in Figure 2, critical visual evidence from the officer's cap propagates through tokens such as *wearing* and *badge*. Direct attribution (a) ranks the cap only 164th among 176 image patches, while tracing the highest-attribution path (b) improves it to 51st. Aggregating all reasoning paths (c), however, raises it to 2nd, showing that limited path tracing substantially under-ranks evidence by capturing only part of its propagated contribution.

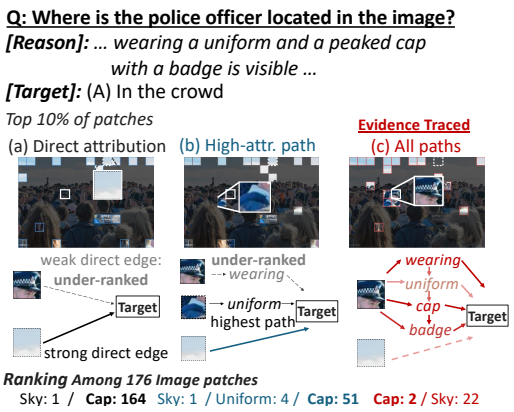


Figure 2: Tracing only highly-attributed paths misses evidence that many weaker paths carry.

To address these challenges, we introduce VTRACE, a multimodal token-attribution framework that traces visual and textual contributions through intermediate reasoning and calibrates attribution scores across modalities. To reduce the underrepresentation of visual evidence in joint image-text ranking, VTRACE isolates token-specific contributions by removing shared components within each modality, and calibrates aggregated image and text attributions according to each modality's effect on response likelihood. To recover visual contributions distributed across intermediate reasoning, VTRACE constructs pairwise token attributions over the full sequence and aggregates all direct and indirect reasoning paths in closed form, tracing the accumulated contributions back to the input tokens. Together, our contributions are threefold:

- We empirically characterize two critical limitations when extending token attribution to LVLMs: incomplete recovery of visual contributions through intermediate reasoning and underrepresentation of visual evidence in joint image-text rankings.
- We propose VTRACE, which combines modality-aware pairwise attribution, closed-form path aggregation, and cross-modal calibration to trace source evidence through reasoning while making image and text attribution scores comparable.

- We evaluate against seven baselines across six visual reasoning benchmarks. In joint image-text evaluation, VTRACE improves mean RISE insertion AUC by 7.1% and reduces deletion AUC by 18.0%, demonstrating more faithful attribution with competitive computational efficiency. The attribution-guided post-training experiment improves model's visual reasoning performance, suggesting the potential of VTRACE as a learning signal for LVLM reasoning.

2 Related Work

Our work builds on three lines of research. **1) Token attribution:** Existing methods estimate which source tokens influence a target token or output span using attention aggregation (Attention Rollout, ALTI) (Abnar & Zuidema, 2020; Ferrando et al., 2022b), relevance propagation (AttnLRP) (Achtibat et al., 2024), Hessian-based sensitivity (HETA) (Pramanik et al., 2026), or perturbation-based attribution (ReAGent) (Zhao & Shan, 2024). Graph-based approaches instead trace contribution paths, where IFR builds a graph over token states and components (Ferrando & Voita, 2024), FlashTrace recursively tracks attribution through generated tokens (Pan et al., 2026), and FlowTracer models conserved flow over an attention graph (Dong et al., 2026). **2) Multimodal attribution:** Chefer et al. (Chefer et al., 2021) propagate relevance through self- and cross-attention to attribute multimodal predictions to their inputs. LVLM-Interpret (Stan et al., 2024) extends attention- and relevance-based analysis to autoregressive LVLMs, while GLIMPSE (Shen, 2025) aggregates gradient-weighted attention across layers and generated tokens for response-level attribution. **3) Attribution through LVLM reasoning:** Multimodal-CoT and LLaVA-CoT generate intermediate reasoning from image and text before answering (Zhang et al., 2024b; Xu et al., 2025), motivating studies of whether such reasoning faithfully reflects the evidence used by the model (Chen et al., 2024b; Balasubramanian et al., 2025). Existing attribution methods, however, only partially capture how multimodal evidence propagates through intermediate reasoning and do not address attribution-scale differences between image and text tokens. VTRACE addresses both by aggregating direct and indirect influence across the reasoning trace and calibrating cross-modal attribution scores. Detailed review is in Appendix B.

3 Critical Challenges in Multimodal Attribution

3.1 Preliminaries

Autoregressive LVLM Generation. Given an image I and a text prompt q, an LVLM with parameters θ generates a response $y = (y_1, \ldots, y_n)$ containing intermediate reasoning and a final answer:

$$p_\theta(y \mid I, q) = \prod_{t=1}^{n} p_\theta(y_t \mid I, q, y_{<t}). \tag{1}$$

Under this autoregressive process, input evidence can contribute to later predictions both directly and through intermediate reasoning tokens. Our goal is to attribute a selected output token or span, ranging from an individual answer token to the complete response, to its supporting input evidence by accounting for these direct and indirect contributions.

Pairwise Token Attribution. Let $X = (x_1, \ldots, x_T)$ denote the complete sequence of input and generated tokens, with $\mathcal{I}$ and $\mathcal{T}$ indexing image and prompt-text tokens, respectively. We estimate direct token-to-token contributions using a pairwise attribution matrix $W \in \mathbb{R}_{\geq 0}^{T \times T}$, where W_{ij} scores the contribution from an earlier source token x_i to a later receiver token x_j. To trace contributions forward through the sequence, we retain only entries with $i < j$ and set $W_{ij} = 0$ otherwise. Thus, W is strictly upper triangular. A generated token can receive contributions from earlier tokens and subsequently contribute to later tokens, allowing input evidence to propagate through intermediate reasoning. Section 4.1 defines how the entries of W are computed.

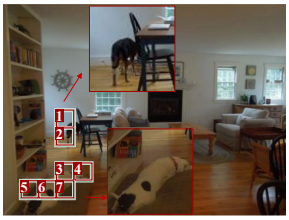


Figure 3: Existing methods overlook visual cues.

3.2 Visual Evidence Is Obscured by Textual Attribution

In multimodal attribution, image and text tokens are jointly ranked according to their contributions, yet existing methods exhibit a strong attribution imbalanced toward text tokens. As shown in Figure 1 Right, existing methods select only a small fraction of image tokens among their highest-ranked tokens. This imbalance is also evident in Figure 3: when the model correctly counts two dogs, neither IFR nor FlashTrace includes an image patch in its top-10 tokens, while option letters, punctuation, and question tokens occupy much of the ranking. These observations reveal attribution imbalance that pushes important visual evidence below textual tokens in attribution, causing the resulting attribution to underrepresent the visual information in multimodal reasoning. In contrast, VTrace aligns image and text scores by their effect on response likelihood, bringing seven image patches into the top 10, producing a ranking that better reflects the visual evidence required for the answer (Section 4.3).

3.3 Limited Reasoning-Path Tracing Marginalizes Visual Evidence

Existing attribution methods trace only direct or limited-hop contributions, making it difficult to trace distributed evidence back to its original sources. For instance, IFR measures direct input to answer contributions (Ferrando & Voita, 2024), while FlashTrace recursively propagates attribution through a limited number of reweighted reasoning steps (Pan et al., 2026). However, evidence from input tokens can be progressively integrated into subsequent generated tokens during reasoning, which in turn contribute to later predictions. As a result, the final prediction receives strong attribution from intermediate reasoning tokens while assigning weak attribution to the upstream visual evidence from which they originated. This is depicted in the aforementioned Figure 2. VTrace instead aggregates evidence over all direct and indirect paths (Section 4.2), allowing information propagated through the reasoning trace to be traced back to the crucial image and text tokens.

4 Methodology

To address these issues, VTrace proceeds in three stages. It first constructs a modality-aware pairwise attribution matrix, then aggregates direct and indirect contributions through intermediate reasoning, and finally calibrates image and text attribution scores for joint ranking. Figure 4 presents the framework, and Algorithm is in Appendix Algorithm 1.

4.1 Modality-Aware Pairwise Token Attribution

Tracing token-specific information. Consider layer ℓ of an LVLM with H attention heads and hidden dimension d. For a receiver token x_j, the attention block produces the update $\Delta_j^{(\ell)}$:

$$\Delta_j^{(\ell)} = \sum_{h=1}^{H} \sum_{i \le j} \alpha_{ji}^{(\ell,h)} f_i^{(\ell,h)}, \tag{2}$$

where $\alpha_{ji}^{(\ell,h)}$ is the attention weight assigned by receiver token x_j to source token x_i, and $f_i^{(\ell,h)} = v_i^{(\ell,h)} O^{(\ell,h)} \in \mathbb{R}^d$ is the output-projected value vector of x_i at head h.

As shown in Section 3.2, directly transferring token-attribution signals across modalities leads to biased multimodal attribution. We instead characterize a source by the information it contributes over the common component of its own modality. For a source x_i in either modality, we define the modality-aware write $\widetilde{f}_i^{(\ell,h)}$ [3] by subtracting its modality's mean output-projected value:

$$\widetilde{f}_i^{(\ell,h)} = f_i^{(\ell,h)} - \frac{1}{|\mathcal{M}(i)|} \sum\nolimits_{r \in \mathcal{M}(i)} f_r^{(\ell,h)}, \tag{3}$$

where $\mathcal{M}(i) = \mathcal{I}$ for $i \in \mathcal{I}$ and $\mathcal{M}(i) = \mathcal{T}$ for $i \in \mathcal{T}$, denoting the set of input tokens belonging to the same modality as token x_i. This removes the modality-wise mean component from each source write, emphasizing the component that distinguishes x_i from other tokens within the same modality.

To quantify the attribution from source x_i to receiver x_j, we measure how strongly the source-specific write aligns with the update at x_j. For $i < j$, the corresponding entry of W is:

[3] We use *write* in the residual-stream sense: a component reads from the residual stream and contributes an update back to it (Elhage et al., 2021).

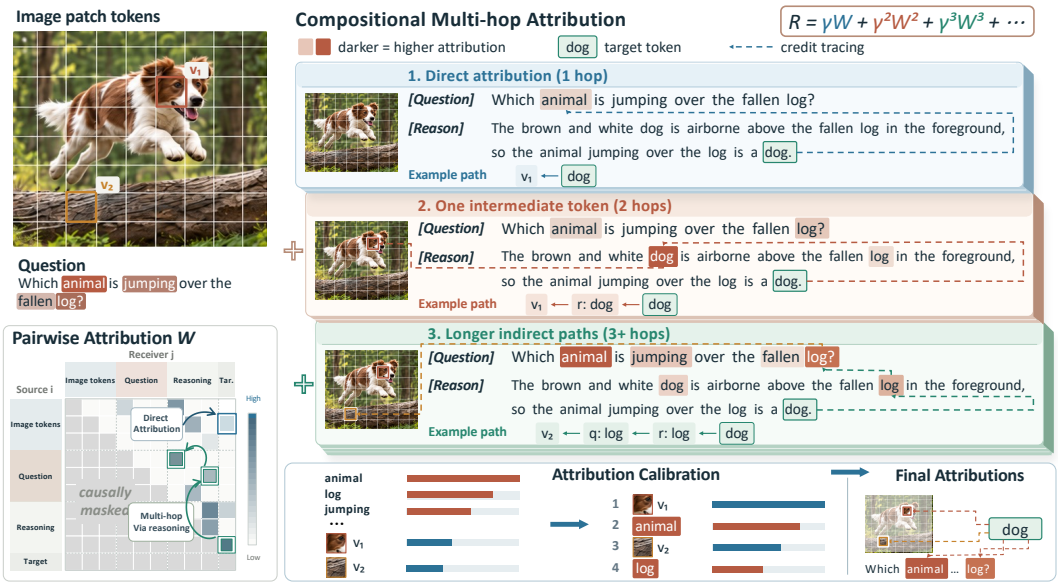


Figure 4: Overview of VTRACE. VTRACE (1) measures each pair of tokens against the average contribution of its own modality and (2) aggregates all direct and indirect paths in closed form, (3) enabling image patches and words to be faithfully ranked together.

$$W_{ij} = \frac{1}{L}\sum_{\ell=1}^{L} \frac{\left[\sum_{h=1}^{H} \alpha_{ji}^{(\ell,h)} \left\langle \tilde{f}_i^{(\ell,h)}, \Delta_j^{(\ell)} \right\rangle\right]_+}{\left\|\Delta_j^{(\ell)}\right\|_2}, \qquad i < j, \tag{4}$$

and $W_{ij} = 0$ otherwise. The inner product $\langle \cdot, \cdot \rangle$ measures alignment between the source vector and the receiver-token update, while $[\cdot]_+$ retains positive contributions. Thus, $W \in \mathbb{R}_{\geq 0}^{T \times T}$ is a strictly upper-triangular matrix containing the direct pairwise token attribution weights.

4.2 Compositional Multi-Hop Attribution: Direct and Indirect Contributions

The pairwise attribution matrix W quantifies direct contributions between source and receiver tokens. A source token x_i, however, can contribute to a later target x_j either directly or through intermediate tokens, including generated reasoning tokens. We refer to aggregating all such direct and indirect paths as *compositional multi-hop* (CMH) attribution. VTRACE computes CMH attribution over W using a Katz-based formulation (Katz, 1953):

$$R_{ij} = \gamma \widehat{W}_{ij} + \gamma^2 (\widehat{W}^2)_{ij} + \gamma^3 (\widehat{W}^3)_{ij} + \cdots = \sum_{\tau=1}^{T-1} \gamma^\tau (\widehat{W}^\tau)_{ij}, \tag{5}$$

where $\widehat{W} = \frac{W}{\max_{i,j} W_{ij} + \epsilon}$ is the normalized direct-attribution matrix, and $\gamma \geq 0$ weights paths by length. Here, $\widehat{W}_{ij}$ captures direct attribution, while $(\widehat{W}^\tau)_{ij}$ aggregates attribution over all length-τ paths from x_i to x_j. The finite path expansion admits the exact closed form:

$$R = \sum_{\tau=1}^{T-1} \gamma^\tau \widehat{W}^\tau = (\mathbf{I} - \gamma \widehat{W})^{-1} - \mathbf{I} \tag{6}$$

The proof is in Appendix C. Each R_{ij} aggregates the direct and indirect attribution from source x_i to receiver x_j represented by $\widehat{W}_{ij}$.

Scoring tokens by incoming and outgoing attribution. The matrix R assigns attribution to pairs of tokens. To obtain one score for each token, we consider every token x_r that precedes the selected receiver positions $x_j \in \mathcal{A}$. Its incoming attribution is the total path attribution reaching x_r from earlier tokens, and its outgoing attribution is the total path attribution from x_r to the selected receivers:

$$\mathrm{In}(x_r) = \sum_{i<r} R_{ir}, \qquad \mathrm{Out}_{\mathcal{A}}(x_r) = \sum_{j \in \mathcal{A}} R_{rj}. \tag{7}$$

We define the **uncalibrated attribution score** of x_r as: $u_r = \big(1 + \mathrm{In}(x_r)\big)\,\mathrm{Out}_{\mathcal{A}}(x_r)$.

This score includes paths that begin at x_r and paths that pass through x_r before reaching a selected receiver x_j. The added 1 preserves the contribution of a token with no incoming attribution.

4.3 Aligning Image and Text Attribution Scores

The multi-hop stage produces an uncalibrated attribution score u_r for each upstream token. To jointly attribute image and textual tokens, we calibrate their total attribution according to each modality's contribution to the generated output. For a modality set $\mathcal{S}$, we measure this contribution by the decrease in teacher-forced log-probability when its tokens are masked:

$$D_{\mathcal{S}} = \log p_\theta(y \mid X) - \log p_\theta\big(y \mid \mathrm{pad}(\mathcal{S})\big), \qquad \mathcal{S} \in \{\mathcal{I}, \mathcal{T}, \mathcal{I} \cup \mathcal{T}\}. \tag{8}$$

Here, $\mathrm{pad}(\mathcal{S})$ replaces positions in $\mathcal{S}$ with the model's pad-token embedding, so $D_{\mathcal{S}}$ measures the contribution of $\mathcal{S}$ to the output. To account for image-text interactions, we use exact two-player Shapley values to divide their joint contribution:

$$\phi_{\mathcal{I}} = \frac{1}{2}\left(D_{\mathcal{I}} + D_{\mathcal{I}\cup\mathcal{T}} - D_{\mathcal{T}}\right), \qquad \phi_{\mathcal{T}} = \frac{1}{2}\left(D_{\mathcal{T}} + D_{\mathcal{I}\cup\mathcal{T}} - D_{\mathcal{I}}\right). \tag{9}$$

Let $s = \phi_{\mathcal{I}}/(\phi_{\mathcal{I}} + \phi_{\mathcal{T}})$ be the image share. We then rescale image-token scores by:

$$\lambda = \frac{s}{1-s}\frac{\sum_{j\in\mathcal{T}} u_j}{\sum_{i\in\mathcal{I}} u_i}, \tag{10}$$

and leave user-text scores unchanged. This matches the total image-text attribution ratio to their estimated modality contributions while preserving the ranking within each modality.

Final VTrace attribution score. For each input token r, the final VTrace attribution score is:

$$u_{\text{VTrace}} \begin{cases} \lambda u_r, & r \in \mathcal{I}, \\ u_r, & r \in \mathcal{T}, \end{cases} \qquad r \in \mathcal{I} \cup \mathcal{T}. \tag{11}$$

Therefore, VTrace uses R_{ij} to aggregate direct and indirect attribution from source x_i to target x_j, and u_{VTrace} as the final cross-modal attribution score to jointly rank image and prompt-text tokens.

5 Experiments

Benchmark Datasets and Baseline Methods (Appendix E). We conduct experiments on 6 visual reasoning benchmarks to comphrehensively evaluate our method's capability: MMStar (Chen et al., 2024a), MathVista (Lu et al., 2024), MMMU (Yue et al., 2024), MMMU-pro (Yue et al., 2025), MathVerse (Zhang et al., 2024a) and VisualPuzzles (Song et al., 2025). They cover visually dependent understanding, mathematical reasoning, and knowledge-based visual reasoning, providing diverse settings for evaluating attribution across multimodal reasoning processes. We compare our method against existing attribution methods, including ReAGent (Zhao & Shan, 2024), HETA (Pramanik et al., 2026), FlowTracer (Dong et al., 2026), IFR (Ferrando & Voita, 2024), Attn Rollout (Abnar & Zuidema, 2020), AttnLRP (Achtibat et al., 2024), FlashTrace (Pan et al., 2026).

Implementation Details. Experiments are conducted on multiple scales of Qwen3-VL (Bai et al., 2025) and InternVL3.5 (Wang et al., 2025) (default Qwen3-VL-8B). For each sample, we generate one response containing intermediate reasoning and a final answer, and keep the response fixed across all attribution methods to ensure fairness. The main evaluation attributes the complete response to the input tokens. Following prior work on attribution for reasoning models (Pan et al., 2026), we evaluate attribution faithfulness using RISE (Petsiuk et al., 2018) and MAS (Chase Walker et al., 2024) insertion and deletion metrics, where tokens are ranked by attribution score and progressively removed or restored. Deletion measures how quickly the response degrades when highly attributed tokens are removed, while insertion measures how quickly it recovers when they are restored. We report both metrics under **image-only** and **joint** settings, ranking visual tokens alone or visual and textual tokens together, respectively. For an input perturbed at step k, we score the model using the normalized likelihood of the original generated trace y: $f(X)_k = \exp\Big(\frac{1}{n_{\text{gen}}}\sum_{t=1}^{n_{\text{gen}}} \log p_\theta\big(y_t \mid X^{(k)}, y_{<t}\big)\Big)$, where $X^{(k)}$ denotes the perturbed context at step k. Detailed implementation and evaluation setups are provided in Appendices F and G.

Table 1: **RISE attribution faithfulness** on Qwen3-VL-8B across six benchmarks under **Image** and **Joint** perturbation settings. We report RISE insertion (Ins.↑, higher is better) and deletion (Del.↓, lower is better) AUC. The Image variant perturbs image patch tokens only, while Joint perturbs both image and text tokens. **Best** and Runner-up are highlighted.

Dataset	Setting	RISE	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
MMStar	Image	Ins.↑	0.505	0.497	0.532	0.539	0.539	0.563	0.555	**0.600**
		Del.↓	0.458	0.463	0.412	0.409	0.439	0.384	0.394	**0.352**
	Joint	Ins.↑	0.371	0.489	0.506	0.507	0.501	0.529	0.554	**0.581**
		Del.↓	0.325	0.245	0.226	0.222	0.273	0.219	0.204	**0.180**
MathVista	Image	Ins.↑	0.500	0.514	0.579	0.591	0.584	0.599	0.600	**0.662**
		Del.↓	0.447	0.444	0.368	0.363	0.394	0.358	0.354	**0.311**
	Joint	Ins.↑	0.382	0.467	0.517	0.516	0.501	0.512	0.574	**0.647**
		Del.↓	0.345	0.307	0.274	0.269	0.331	0.274	0.241	**0.178**
MMMU	Image	Ins.↑	0.543	0.564	0.595	0.610	0.627	0.612	0.618	**0.665**
		Del.↓	0.482	0.480	0.438	0.425	0.444	0.419	0.417	**0.379**
	Joint	Ins.↑	0.368	0.589	0.583	0.608	0.604	0.615	0.629	**0.660**
		Del.↓	0.315	0.216	0.204	0.195	0.241	0.191	0.178	**0.159**
MMMU-Pro	Image	Ins.↑	0.527	0.542	0.570	0.572	0.604	0.593	0.583	**0.645**
		Del.↓	0.472	0.472	0.442	0.435	0.444	0.413	0.429	**0.378**
	Joint	Ins.↑	0.371	0.558	0.532	0.539	0.554	0.559	0.584	**0.619**
		Del.↓	0.322	0.241	0.242	0.242	0.269	0.229	0.212	**0.181**
MathVerse	Image	Ins.↑	0.523	0.565	0.636	0.640	0.624	0.610	0.644	**0.684**
		Del.↓	0.454	0.430	0.341	0.345	0.384	0.360	0.341	**0.296**
	Joint	Ins.↑	0.383	0.495	0.539	0.540	0.513	0.547	0.568	**0.606**
		Del.↓	0.330	0.314	0.279	0.283	0.342	0.274	0.241	**0.187**
VisualPuzzles	Image	Ins.↑	0.456	0.461	0.468	0.511	0.525	0.531	0.510	**0.557**
		Del.↓	0.409	0.405	0.379	0.354	0.380	0.324	0.358	**0.309**
	Joint	Ins.↑	0.376	0.488	0.477	0.504	0.523	0.529	0.530	**0.571**
		Del.↓	0.348	0.288	0.291	0.283	0.283	0.255	0.268	**0.207**

5.1 QUANTITATIVE ANALYSIS

VTRACE consistently outperforms existing attribution methods under both Image and Joint evaluation. As shown in Table 1, VTRACE outperforms all baselines across six benchmarks. Under the Image setting, VTRACE improves average RISE insertion by 6.8% and reduces deletion by 9.3%. Under the Joint setting, it improves insertion by 7.1% and reduces deletion by 18.0%. These consistent gains indicate that VTRACE more faithfully identifies and ranks the visual and textual tokens contributing to the model response. Moreover, improvements in both insertion and deletion further show that highly ranked tokens are more effective at recovering the response when restored and disrupting it when removed. For clarity, the main table reports RISE only, with complete RISE and MAS results provided in Appendix H.1.

Table 2: Ablation study on different components

Method	Image		Joint	
	Ins. ↑	Del. ↓	Ins. ↑	Del. ↓
VTRACE	**0.631**	**0.327**	**0.610**	**0.187**
w/o centering	0.625	0.340	0.599	0.200
w/o multi-hop	0.578	0.362	0.470	0.273
w/o calibration	0.631	0.327	0.602	0.200

Ablation study. We conduct ablation study on each component of VTRACE, including modality centering, compositional multi-hop aggregation and cross-modal calibration. As reported in Table 2, removing each component lowers attribution faithfulness. Specifically, modality centering and multi-hop aggregation improve faithfulness in both variants. Calibration further improves Joint attribution while preserving the relative ranking of image tokens, complementing the token level attribution with cross-modal calibration. The sensitivity study and is in Appendix H.2.

Efficiency Analysis. VTRACE efficiently scales to full-trace attribution because it constructs the attribution matrix once and then reuses it to attribute every generated token. For all methods, we measure the full attribution runtime, from processing the fixed response to producing the final attribution scores. As a result, moving from one target span to all generated tokens adds only minimal overhead, increasing medium runtime from 0.24 s to 0.26 s (Figure 6(a)). Notably, this efficiency is achieved without sacrificing faithfulness, where Figure 6(b) shows that VTRACE attains the best faithfulness at substantially lower attribution time than baselines. Moreover, **VTRACE scales favorably in memory and runtime**, using 31 GB at around 3,000 tokens versus 87 GB for AttnLRP (Figure 6(c)) and remaining faster than baselines as sequence length grows (Figure 6(d)).

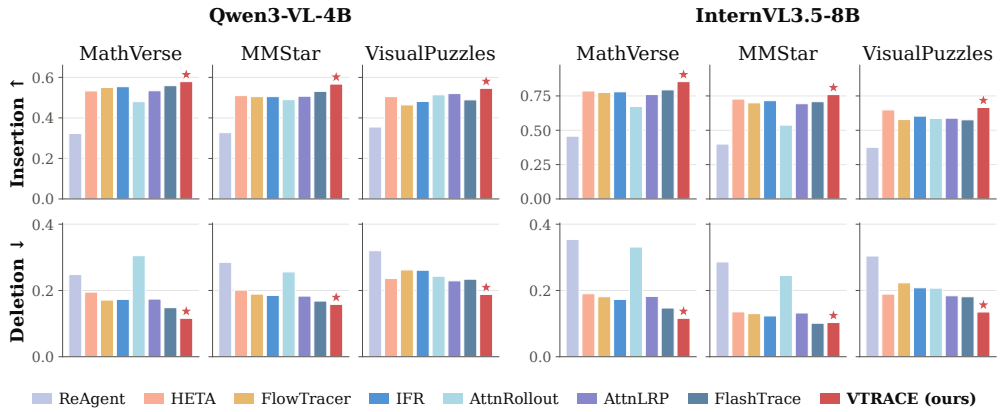


Figure 5: Attribution faithfulness across different LVLM sizes and families.

Table 3: Attribution faithfulness on correct and incorrect model predictions over benchmarks. We compare against the strongest baseline for each metric independently.

		Correct Predictions				Incorrect Predictions			
Variant	Method	Del. RISE ↓	Del. MAS ↓	Ins. RISE ↑	Ins. MAS ↑	Del. RISE ↓	Del. MAS ↓	Ins. RISE ↑	Ins. MAS ↑
Image	**Best Baseline**	0.377	0.508	0.590	0.451	0.424	0.579	0.603	0.460
	VTRACE	**0.333**	**0.468**	**0.644**	**0.514**	**0.398**	**0.550**	**0.636**	**0.496**
Joint	**Best Baseline**	0.209	0.338	0.581	0.381	0.200	0.323	0.605	0.410
	VTRACE	**0.171**	**0.235**	**0.626**	**0.463**	**0.176**	**0.241**	**0.640**	**0.474**

Table 4: Attribution-guided learning with VTRACE on Qwen3-VL-4B. Incorporating VTRACE attribution into GRPO improves the average performance across benchmarks.

Method	MMStar	MathVista	MMMU	MathVerse	MMMU-Pro	HallusionBench	RealWorldQA	MathVision	Avg.
Base	63.53	73.30	55.33	56.55	47.57	70.87	72.55	41.78	60.19
GRPO	69.53	**77.40**	61.89	62.64	**51.68**	72.77	72.29	43.75	63.99
+ VTRACE	**70.20**	77.20	**65.00**	**65.13**	51.45	**73.29**	**73.20**	**46.71**	**65.27**

VTRACE consistently improves attribution faithfulness across different LVLM architectures and sizes. Shown in Figure 5, VTRACE achieves the highest RISE insertion and lowest deletion scores on both Qwen3-VL-4B and InternVL3.5-8B across diverse benchmarks, indicating strong generalization across model families and sizes. Full results are in Appendix H.3.

VTRACE improves attribution faithfulness regardless of answer correctness. Shown in Table 3, VTRACE consistently outperforms the strongest baselines across different evaluation settings. This indicates that VTRACE can faithfully identify influential token relationships regardless of the model's prediction correctness. Full results are in Appendix H.4.

VTRACE can also provide a meaningful learning signal for improving LVLM reasoning. Following (Dong et al., 2026), we use its token-level attribution for credit assignment in Group Relative Policy Optimization (GRPO), directly integrating traced reasoning flow into post-training. Shown in Table 4, attribution-guided GRPO improves the average performance of Qwen3-VL-4B from 63.99 to 65.27. This demonstrates that VTRACE captures token relationships that are useful for optimization, extending its value beyond interpretability to downstream reinforcement learning and post-training.

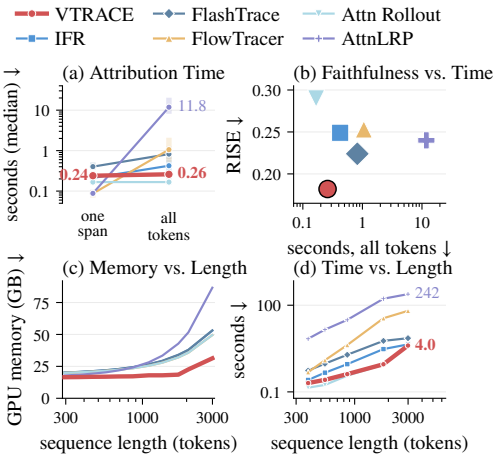


Figure 6: Efficiency analysis.

5.2 Qualitative Analysis

VTRACE achieves faster recovery under insertion and sharper degradation under deletion. Elaborated in Appendix H.5 Figure 11, VTRACE achieves faster recovery under insertion and sharper degradation under deletion.

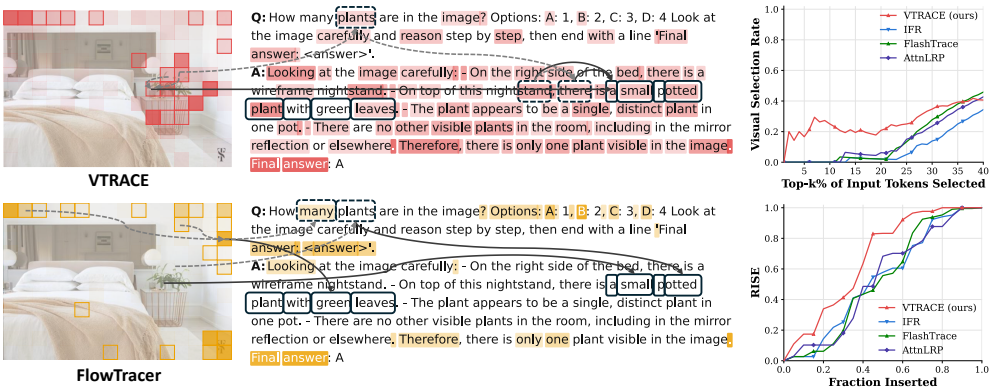


Figure 7: Qualitative study: VTRACE better traces reasoning tokens to key visual evidence.

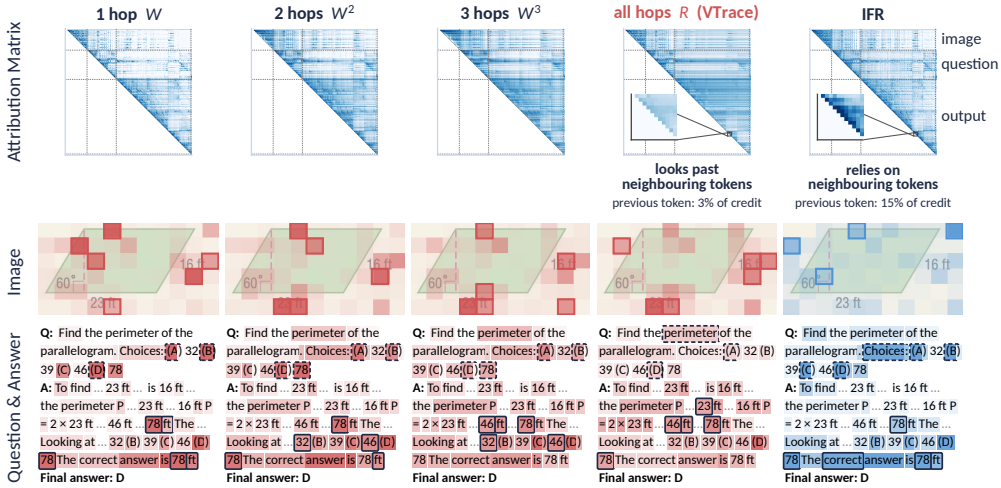


Figure 8: Qualitative study: Attribution comparison between diverse hops and IFR.

Case study. VTRACE more clearly traces visual evidence through intermediate reasoning to the final response. Evident in Figure 7, VTRACE assigns stronger attribution to plant-related image patches, while the baseline focuses more on background regions. VTRACE's token-level paths further connect plant-related reasoning tokens back to the corresponding visual patches, which the baseline largely misses. Notably, when top 10% tokens are recovered, image tokens account for around a quarter of VTRACE's selected tokens, while baselines select almost none. The RISE insertion curve (bottom-right) further shows that cross-modal calibration helps VTRACE recover response likelihood with fewer inserted tokens. Moreover, Figure 8 shows that limited-hop attribution concentrates credit on nearby reasoning tokens and misses key visual evidence (e.g., IFR), reiterating the challenge we identified in Sec 3.3. By aggregating all reasoning paths, VTRACE traces this credit back to the image patches that support the answer. An extended analysis of the diverse hops and VTRACE is in Appendix H.6. Additional case studies are in Appendix I.

6 CONCLUSION

In this work, we investigate faithful multimodal token attribution for LVLM reasoning, with the aim of tracing generated predictions back to the faithful visual and textual evidence that supports the generation. Through empirical analysis, we find that existing methods struggle with contributions propagated through intermediate reasoning and underrepresent relevant visual evidence. To address these challenges, we propose VTRACE, which represents token attributions in a pairwise attribution matrix and considers both direct and indirect paths. VTRACE further calibrates image and text attribution to provide an unbiased cross-modal ranking. Experiments across six visual reasoning benchmarks consistently outperforms baseline methods, demonstrating that VTRACE provides more faithful multimodal attribution while maintaining efficiency.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language polishing, literature survey, and the development of the demonstration dashboard. For example, generative AI was used to check whether related methods had been previously explored, and help with structuring the interface layout and rendering dashboard content.

We have not used generative AI tools to design the core algorithm and methodology proposed in this work, generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine hypotheses, design research methodology or experiments, translate content, clean or reformat datasets, support qualitative or thematic data analysis. All AI-assisted content is manually reviewed and verified. We take responsibility for the final content of this work, including text, claims, code and any artifacts produced with the aid of generative AI.

ETHICS STATEMENT

As a foundational study on faithful token attributions, we do not identify any direct ethics concern. Our research relies on publicly available datasets and models, all of which are appropriately acknowledged.

REPRODUCIBILITY STATEMENT

To support reproducible research, we summarize our efforts as below:

1. **Baselines & Datasets.** All six benchmarks and all three models are publicly available. Appendix E describes each of the seven baselines and lists each benchmark.
2. **Evaluation Metrics.** Section 5 summarizes the RISE and MAS protocol, and Appendix G gives the full definitions, the perturbation schedule and the system prompt used to generate the reasoning traces.
3. **Implementation.** Section 4 defines every step of VTRACE, Appendix C proves the closed form, and Algorithm 1 gives the full procedure. Appendix F describes the experimental details, the models used, the decoding settings, the hardware, and the setup of the attribution-guided GRPO experiment.
4. **Source Code & Demo.** We provide an anonymized reference code at https://anonymous.4open.science/r/VTRACE-CB12, and a demo for visualizing token attributions at https://huggingface.co/spaces/VTrace/vtrace-demo.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *ACL*, 2020.

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. Attnlrp: Attention-aware layer-wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*, 2024.

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.

Sriram Balasubramanian, Samyadeep Basu, and Soheil Feizi. A closer look at bias and chain-of-thought faithfulness of large (vision) language models. In *EMNLP*, 2025.

Dominic Simon Chase Walker, Kenny Chen, and Rickard Ewetz. Attribution quality metrics with magnitude alignment. In *IJCAI*, 2024.

Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, 2021.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024a.

Shuang Chen, Yue Guo, Zhaochen Su, Yafu Li, Yulun Wu, Jiacheng Chen, Jiayu Chen, Weijie Wang, Xiaoye Qu, and Yu Cheng. Advancing multimodal reasoning: From optimized cold start to staged reinforcement learning. *arXiv preprint arXiv:2506.04207*, 2025.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. Measuring and improving chain-of-thought reasoning in vision-language models. In *NAACL*, 2024b.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *CVPR*, pp. 3867–3876. IEEE, 2025.

Zhichen Dong, Yang Li, Yuhan Sun, Weixun Wang, Yijia Luo, Zinian Peng, Taiheng Ye, Chao Yang, Wenbo Su, Yu Cheng, et al. How does reasoning flow? tracing attention-induced information flow for targeted rl in llms. *arXiv preprint arXiv:2606.10646*, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. In *EMNLP*, 2024.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *EMNLP*, 2022a.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. Measuring the mixing of contextual information in the transformer. In *EMNLP*, 2022b.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *CVPR*, 2024.

Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953. doi: 10.1007/BF02289025.

Sicong Leng, Jing Wang, Jiaxi Li, Hao Zhang, Zhiqiang Hu, Boqiang Zhang, Yuming Jiang, Hang Zhang, Xin Li, Lidong Bing, Deli Zhao, Wei Lu, Yu Rong, Aixin Sun, and Shijian Lu. Mmr1: Enhancing multimodal reasoning with variance-aware sampling and open resources, 2025. URL https://arxiv.org/abs/2509.21268.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Wenbo Pan, Zhichao Liu, Xianlong Wang, Haining Yu, and Xiaohua Jia. Towards long-horizon interpretability: Efficient and faithful multi-token attribution for reasoning llms. *arXiv preprint arXiv:2602.01914*, 2026.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Vishal Pramanik, Maisha Maliha, Nathaniel Bastian, and Sumit Jha. Hessian-enhanced token attribution (heta): Interpreting autoregressive llms. In *ICLR*, 2026.

Guanxi Shen. Glimpse: Holistic cross-modal explainability for large vision-language models. *arXiv preprint arXiv:2506.18985*, 2025.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Yueqi Song, Tianyue Ou, Yibo Kong, Zecheng Li, Graham Neubig, and Xiang Yue. Visualpuzzles: Decoupling multimodal reasoning evaluation from domain knowledge, 2025. URL https://arxiv.org/abs/2504.10342.

Gabriela Ben Melech Stan, Estelle Aflalo, Raanan Yehezkel Rohekar, Anahita Bhiwandiwalla, Shao-Yen Tseng, Matthew Lyle Olson, Yaniv Gurwicz, Chenfei Wu, Nan Duan, and Vasudev Lal. Lvlm-intrepret: An interpretability tool for large vision-language models. In *XAI4CV Workshop CVPR*, 2024.

Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. Journey before destination: On the importance of visual faithfulness in slow thinking. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4147–4168, 2026.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. In *NeurIPS*, 2024.

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.

xAI. Realworldqa: A benchmark for real-world spatial understanding. https://huggingface.co/datasets/xai-org/RealworldQA, 2024. Accessed: 2026-09-18.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step. In *ICCV*, 2025.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *CVPR*, 2024.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark, 2025. *URL https://arxiv. org/abs/2409.02813*, 2025.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MATHVERSE: does your multi-modal LLM truly see the diagrams in visual math problems? In *ECCV*, 2024a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *TMLR*, 2024b.

Zhixue Zhao and Boxuan Shan. Reagent: A model-agnostic feature attribution method for generative language models. *arXiv preprint arXiv:2402.00794*, 2024.

APPENDIX

We summarize the key structure of the Appendix as follow:

- **Section A:** Limitations
- **Section B:** Extended Related Work
- **Section C:** Proof of Equation (6)
- **Section D:** VTRACE Algorithm
- **Section E:** Baseline & Benchmark Descriptions
- **Section F:** Experimental Details
- **Section G:** Evaluation Details
- **Section H:** Extended Experiments
- **Section I:** Case Studies

A LIMITATIONS

While VTRACE demonstrates strong effectiveness, several limitations remain. Our current evaluation focuses primarily on image-based visual reasoning, while modern LVLMs also operate in broader settings, including multi-image reasoning, long-context multimodal understanding, and longer-horizon agentic interaction. Extending VTRACE to these settings would further establish the generality of the proposed attribution framework. Another direction is to integrate attribution into model training to improve model visual reasoning behavior. Our preliminary attribution-guided learning experiments suggest that this direction is promising.

B EXTENDED RELATED WORK

B.1 TOKEN ATTRIBUTION AND INFORMATION FLOW

Token attribution traces how source tokens contribute to the generation of a target token or output span. Existing methods differ mainly in the model signals used to define this contribution. Attention Rollout composes attention matrices across layers to capture how information is mixed between token positions throughout the network (Abnar & Zuidema, 2020). ALTI similarly aggregates token interactions across layers, but derives them from the attention block while accounting for residual connections and layer normalization (Ferrando et al., 2022b). Chefer et al. (Chefer et al., 2021) propose to propagate relevance through attention and residual operations, while AttnLRP extends the layer-wise relevance propagation to Transformer attention and supports attribution to both input tokens and intermediate representations (Achtibat et al., 2024). HETA incorporates attention and value information together with Hessian-based sensitivity and KL divergence under token masking (Pramanik et al., 2026).

For generative models, attribution must additionally account for the role of previously generated tokens. A source token may affect the final prediction directly or indirectly through intermediate tokens in the generated sequence. ALTI+ study this setting in machine translation by tracing contributions from both the source sentence and the generated prefix (Ferrando et al., 2022a). IFR represents each prediction as a computation graph whose nodes correspond to token states and model components, allowing attribution to follow routes through internal representations (Ferrando & Voita, 2024).

Several recent approaches **explicitly model indirect contribution** through generated tokens. Flash-Trace aggregates attribution over output spans and recursively follows attribution absorbed by intermediate generated tokens through limited steps (Pan et al., 2026). This captures cases in which an early source influences the selected output through later reasoning tokens rather than through a single direct interaction. FlowTracer constructs an attention-based flow graph, assigns capacities to edges, reweights them according to their ability to reach the target region, and imposes local flow conservation (Dong et al., 2026). The resulting token-level flow scores are further used as

feedback for reinforcement learning. Both methods therefore move beyond purely local attribution by considering routes through intermediate tokens. VTRACE instead aggregates multiple forward paths between source and receiver positions, allowing direct and multi-step effects to be represented within a single computation. The resulting incoming and outgoing contributions are then used to score earlier tokens with respect to selected receiver tokens.

A **complementary family** of methods **estimates importance through perturbation or intervention** rather than tracing internal connections. ReAGent repeatedly replaces subsets of input tokens and updates their importance according to changes in the model's next-token prediction (Zhao & Shan, 2024). This requires only forward evaluations and does not depend on gradients or explicit access to internal attribution signals. VTRACE separates this role from its token-level attribution computation: the pairwise attribution matrix is obtained once from cached model states, while additional forward evaluations with replaced image or prompt embeddings are used only to calibrate the relative contribution of the two input modalities. In our experiments, we further evaluate attribution faithfulness by perturbing the input tokens ranked by each method and measuring the resulting change in model prediction.

B.2 Attribution Across Image and Text

Attribution in large vision-language models introduces an additional challenge because evidence is distributed across heterogeneous visual and textual representations. A meaningful attribution must therefore identify important evidence within each modality fairly. Chefer et al. (Chefer et al., 2021) extend Transformer explanation methods to multimodal architectures containing self-attention, co-attention, and encoder-decoder attention, propagating relevance through the attention interactions that connect visual and textual representations. LVLM-Interpret provides attention maps, relevance maps, and causal visualizations for inspecting how large vision-language models generate responses (Stan et al., 2024). GLIMPSE combines gradient-weighted attention with propagation across layers and aggregation over generated tokens to identify visual and textual evidence supporting a generated response (Shen, 2025).

B.3 Attribution for Reasoning LVLM

Recent LVLMs increasingly generate explicit intermediate reasoning before producing a final answer. Multimodal-CoT generates a textual rationale conditioned on both image and language inputs and then uses this rationale to predict the response (Zhang et al., 2024b). LLaVA-CoT structures multimodal reasoning into stages including summarization, visual interpretation, logical reasoning, and conclusion, and performs search over intermediate stages during inference (Xu et al., 2025). These approaches make intermediate generated tokens an explicit part of the inference process rather than treating the model as producing only a final answer.

Prior work has also examined whether multimodal reasoning faithfully reflects the evidence used by the model. Chen et al. (Chen et al., 2024b) introduce a benchmark and metrics for evaluating reasoning consistency in vision-language models. Balasubramanian et al. (Balasubramanian et al., 2025) study whether generated reasoning reflects visual and textual biases introduced into the input. Their results show that models are less likely to acknowledge subtle visual cues than explicit textual cues, highlighting a potential mismatch between generated explanations and the evidence that actually affects the prediction. Such findings motivate attribution methods that inspect model-internal contributions rather than relying only on the content of the generated rationale.

VTRACE targets this setting by estimating how information propagates through multimodal generations. Given selected receiver tokens, which may correspond to reasoning steps, answer tokens, or an output span, it traces contributions from earlier image, prompt, and generated tokens through both direct and indirect paths. This provides a way to inspect which input tokens supports an intermediate reasoning step, how earlier reasoning tokens influence later ones, and which sources ultimately contribute to the final answer.

C Proof of Eq. 6

Theorem 1 (Closed-form multi-path aggregation)**.** *Let $\widehat{W} \in \mathbb{R}^{T\times T}$ be the strictly upper-triangular matrix. Define:*

$$R = \sum_{\tau=1}^{T-1} \gamma^{\tau} \widehat{W}^{\tau}. \tag{12}$$

Then:

$$R = (\mathbf{I} - \gamma\widehat{W})^{-1} - \mathbf{I}. \tag{13}$$

Proof. Because $\widehat{W}$ is strictly upper triangular, $\widehat{W}^{T} = \mathbf{0}$. Now consider the finite matrix geometric series and its expansion:

$$S = \sum_{\tau=0}^{T-1} (\gamma\widehat{W})^{\tau} = \mathbf{I} + \gamma\widehat{W} + \gamma^{2}\widehat{W}^{2} + \cdots + \gamma^{T-1}\widehat{W}^{T-1} \tag{14}$$

$$= \mathbf{I} + \sum_{\tau=1}^{T-1} \gamma^{\tau} \widehat{W}^{\tau}. \tag{15}$$

Multiplying S by $\mathbf{I} - \gamma\widehat{W}$ gives:

$$(\mathbf{I} - \gamma\widehat{W})S = S - \gamma\widehat{W}S. \tag{16}$$

Substituting the expression for S, we have:

$$(\mathbf{I} - \gamma\widehat{W})S = \left(\mathbf{I} + \gamma\widehat{W} + \gamma^{2}\widehat{W}^{2} + \cdots + \gamma^{T-1}\widehat{W}^{T-1}\right) - \left(\gamma\widehat{W} + \gamma^{2}\widehat{W}^{2} + \cdots + \gamma^{T}\widehat{W}^{T}\right). \tag{17}$$

Note that all the matching intermediate terms cancel, which leaves us with:

$$(\mathbf{I} - \gamma\widehat{W})S = \mathbf{I} - \gamma^{T}\widehat{W}^{T}. \tag{18}$$

Since $\widehat{W}^{T} = \mathbf{0}$, this reduces to:

$$(\mathbf{I} - \gamma\widehat{W})S = \mathbf{I}. \tag{19}$$

Hence, we have:

$$S = \mathbf{I} + \sum_{\tau=1}^{T-1} \gamma^{\tau} \widehat{W}^{\tau} = (\mathbf{I} - \gamma\widehat{W})^{-1}. \tag{20}$$

Finally, subtracting $\mathbf{I}$ from both sides gives:

$$R = \sum_{\tau=1}^{T-1} \gamma^{\tau} \widehat{W}^{\tau} = (\mathbf{I} - \gamma\widehat{W})^{-1} - \mathbf{I}, \tag{21}$$

as required. □

D VTrace Algorithm

Algorithm 1 gives the full procedure in pseudocode. Its input is the cached state of a single forward pass over the frozen trace: the value vectors, the attention weights, and the residual stream before and after each attention block.

- Step 1 computes the modality-aware pairwise attribution matrix W from the cached model states, as described in Section 4.1.

- Step 2 normalizes W to obtain $\widehat{W}$, computes the multi-hop attribution matrix R in closed form, and scores each source token using its incoming attribution and outgoing attribution to the selected receivers, as described in Section 4.2.

- Step 3 is the modality calibration of Section 4.3: three further forward passes measure the loss increase when the image, the text, or both are replaced by padding, the two modality weights are the Shapley values of this two player game, and the image scores are rescaled so that their share of the total matches the image weight while every within-modality order is left unchanged. This ensures the image and text tokens are well calibrated to be ranked fairly and faithfully.

Calibration boundary conditions. Calibration is applied over image and query token positions when used in perturbation evaluation. Calibration is applied only when both modality contributions from Equation (9) and both sums of token attribution scores are positive. If any of the quantities is zero or negative, the original scores are maintained. When applied, calibration multiplies image scores by the positive factor and leaves text scores unchanged. The operation preserves the ranking within each modality, and matches their total attribution ratio to the estimated contribution ratio.

Algorithm 1 VTRACE

```
# Input:  cached states of one forward pass (V, a, x_attn, x_pre),
#         attention output projections Wo, receiver positions recv_pos,
#         image positions img_pos, user-text positions txt_pos, gamma
# Output: VTrace attribution u_VTrace

# 1. Compute the modality-aware pairwise attribution matrix
F[l,i,h] = V[l,i,h] @ Wo[l,h]        # per-head output of i
d[l,j]   = x_attn[l,j] - x_pre[l,j]

# center within modality for input tokens (image or prompt-text tokens)
F[l,i,h] -= mean_k(F[l,k,h], k in modality(i)), i in img_pos or txt_pos

e[l,i,j] = sum_h(a[l,h,j,i] * dot(F[l,i,h], d[l,j]))
W[i,j]   = mean_l(relu(e[l,i,j]) / (norm(d[l,j]) + eps)), i < j

# 2. Katz-based multi-hop attribution
W_hat = W / (max(W) + eps)
R = inv(I - gamma * W_hat) - I
inflow[r] = sum_i(R[i,r])
outflow[r] = sum(R[r,j] for j in recv_pos)
u[r] = (1 + inflow[r]) * outflow[r]
u[min(recv_pos):] = 0                # sources must precede the receivers

# 3. Align image and text attribution scores
# damage(pos) = logp(gen | clean) - logp(gen | pad at pos)
D_img, D_txt = damage(img_pos), damage(txt_pos)
D_both       = damage(img_pos + txt_pos)
phi_img = (D_img + D_both - D_txt) / 2
phi_txt = (D_txt + D_both - D_img) / 2
s = phi_img / (phi_img + phi_txt)     # image share
lam = (s / (1 - s)) * (sum(u[txt_pos]) / sum(u[img_pos]))

# calibrate the image contribution of the attribution
u_VTrace = u
u_VTrace[img_pos] = lam * u[img_pos]
return u_VTrace
```

E BASELINE & BENCHMARK DESCRIPTIONS

E.1 BASELINE METHODS

We compare VTRACE with seven attribution methods:

- **ReAGent** (Zhao & Shan, 2024) is a perturbation method, which repeatedly replaces input tokens with plausible alternatives drawn from a masked language model and scores each token by the change this causes in the probability of the generated text.
- **HETA** (Pramanik et al., 2026) is a gradient method that adds second order terms from the Hessian to first order gradient attribution, so that interactions between input tokens contribute to their scores. We follow the released implementation.
- **Attention Rollout** (Abnar & Zuidema, 2020) composes head-averaged attention across layers. At each layer, attention is combined with the identity matrix as $0.5A^{(\ell)} + 0.5I$ and row-normalized. The resulting layer products provide cumulative input-to-answer attribution.
- **AttnLRP** (Achtibat et al., 2024) applies layer-wise relevance propagation through attention blocks with relevance-conserving rules for softmax and attention matrix multiplication. It requires one backward pass per attributed target token, so its cost increases with response length.
- **IFR** (Ferrando & Voita, 2024) builds on ALTI (Ferrando et al., 2022b) to decompose the Transformer into token-level information flows. It measures how each token contributes to subsequent representations through attention and MLP blocks, tracing these contributions layer by layer.
- **FlowTracer** (Dong et al., 2026) constructs a flow network from head-averaged attention over a selected layer range and scores each input token by the flow it carries to the answer tokens.
- **FlashTrace** (Pan et al., 2026) uses span-wise aggregation and recursive attribution to trace importance through intermediate reasoning. At each hop, important reasoning tokens become weighted targets for the next attribution step, propagating influence backward toward the original input. Attribution across hops is then aggregated into the final input scores.

E.2 BENCHMARK STATISTICS

We evaluate our method on diverse multimodal benchmarks. These benchmarks cover a broad range of multimodal capabilities, ranging from general visual perception to knowledge understanding and mathematical problem solving. A brief introduction for each benchmark is provided below.

- **MMStar** (Chen et al., 2024a) is a vision-indispensable benchmark designed to evaluate multimodal capabilities of LVLMs while reducing the text-only shortcuts effect. It contains 1500 samples covering six core capabilities: coarse perception, fine-grained perception, instance reasoning, logical reasoning, science and technology, and mathematics.
- **MathVista** (Lu et al., 2024) is a benchmark for mathematical reasoning in visual contexts. Its questions cover figure question answering, geometry problem solving, math word problems, textbook question answering and visual question answering, and span seven reasoning types: algebraic, arithmetic, geometric, logical, numeric commonsense, scientific and statistical reasoning. We use the *testmini* split of 1,000 samples.
- **MathVerse** (Zhang et al., 2024a) evaluates visual mathematical reasoning using problems from plane geometry, solid geometry, and functions. Each problem is provided in six variants with different amounts of textual and visual information. We use the *testmini* split.
- **MMMU** (Yue et al., 2024) consists of college level multimodal questions collected from exams, quizzes and textbooks. It spans six disciplines: art and design, business, science, health and medicine, humanities and social science, and technology and engineering, across 30 subjects and 30 image types such as charts, diagrams, tables, chemical structures and medical images. We use the validation split of 900 samples.

- **MMMU-Pro** (Yue et al., 2025) extends MMMU to provide a more challenging evaluation of multimodal reasoning. It removes questions that text-only models answer correctly, expands the candidate options, and adds a vision-only setting in which the question is embedded in the image, so that answering requires reading the image.

- **VisualPuzzles** (Song et al., 2025) is a benchmark that decouples multimodal reasoning from domain knowledge. It contains 1,168 puzzles adapted from logical reasoning questions of the Chinese Civil Service Examination, covering five reasoning categories: algorithmic, analogical, deductive, inductive and spatial reasoning.

We additionally test three visual reasoning benchmarks for our post-training experiments.

- **HallusionBench** (Guan et al., 2024) evaluates hallucination and visual reasoning in LVLMs using 346 images and 1,129 human-designed questions. It particularly emphasis on failures caused by visual illusion and language hallucination.

- **RealWorldQA** (xAI, 2024) evaluates multimodal understanding in real-world scenes. The benchmark includes anonymized vehicle-view images and other natural scenes, with questions focusing particularly on spatial relationships, object properties, and everyday visual understanding.

- **MathVision** (Wang et al., 2024) evaluates visual mathematical reasoning using problems from real mathematics competitions. It spans 16 mathematical disciplines and five difficulty levels, requiring models to jointly interpret visual content and perform advanced mathematical reasoning.

F Experimental Details

Models. We evaluate on the publicly released Qwen3-VL (Bai et al., 2025) checkpoints at 8B and 4B parameters and on InternVL3.5-8B (Wang et al., 2025).

Reasoning Trace Generation. The benchmarks provide the image, the question and the reference answer, but not the reasoning trace that we attribute. We therefore run each model once per question with greedy decoding and a budget of 2,048 max tokens, and record the full context: the image tokens, the question, the generated reasoning and the final answer. This trace is then frozen. Every attribution method is evaluated on the same trace, and every perturbation in the evaluation measures the likelihood of the same trace, so this avoids any discrepancy caused by the generated reasoning traces. For VTrace, the decay rate $\gamma = 1$ by default. We highlight that we do not filter traces by answer correctness, and Table 3 reports the metrics separately for correct and incorrect predictions.

Hardware. All token attribution-based experiments run on a single NVIDIA RTX PRO 6000 Blackwell GPU with 96GB of memory. Post-training based experiments (e.g., GRPO) are conducted on AMD Instinct MI355X GPUs.

F.1 Attribution-guided Learning Details

We conduct the attribution-guided experiments Section 5 using the `verl` framework Sheng et al. (2024). Specifically, we fine-tune Qwen3-VL-4B-Instruct with Group Relative Policy Optimization (GRPO). For each training sample, we generate $n = 8$ rollout responses with a temperature of 0.7 and top-$p = 0.95$.

The training data are constructed by combining and filtering publicly available multimodal reasoning data from MMR1-RL Leng et al. (2025) and the reinforcement-learning split of ReVisual-R1 Chen et al. (2025). We follow the targeted-RL weighting strategy of Dong et al. (2026), where we use VTrace attribution scores to identify the most important response tokens. The top 40% of response tokens are assigned a weight of 1.5, while all remaining tokens retain the default weight of 1.0. All models are trained for 2 epochs under the same training configuration.

G Evaluation Details

Because there are no token-level ground-truth labels indicating which image regions or question tokens contribute to a generated reasoning trace, we evaluate attribution using perturbation-based metrics. Following prior attribution work, we use RISE (Petsiuk et al., 2018) and MAS (Chase Walker et al., 2024), considering both their deletion and insertion variants. These metrics progressively remove or restore input tokens according to their attribution scores and measure how the model's confidence in the generated trace changes.

G.1 Perturbation Protocol

Deletable Tokens. We perturb only tokens from the multimodal input. Let $\mathcal{I}$ and $\mathcal{T}$ denote the image-token and question-token positions, respectively. We consider two evaluation settings: **image**, where only tokens in $\mathcal{I}$ are perturbed, and **joint**, where tokens in both $\mathcal{I}$ and $\mathcal{T}$ are perturbed. Chat-template tokens are excluded, and generated tokens are kept fixed because they constitute the reasoning trace being explained.

Perturbation schedule. For image tokens, perturbation is performed in pixel space. We first construct a Gaussian-blurred version of the image and replace the image regions corresponding to selected visual tokens with their blurred counterparts. The modified image is re-encoded by the LVLM at each perturbation step. For question tokens, selected tokens are replaced with the padding token.

Moreover, perturbing one token at a time would require a separate forward pass for every token. Following FlashTrace (Pan et al., 2026), we therefore evaluate attribution in proportional steps. Given attribution scores u, the P deletable tokens are ranked by decreasing attribution magnitude $|u_i|$ and partitioned into $K = \min(20, P)$ approximately equal-sized groups. Each evaluation thus requires $K + 1$ forward passes.

- For **deletion**, evaluation starts from the original input and progressively perturbs groups from highest to lowest attribution. A faithful attribution should therefore cause the **model response to decrease rapidly**.
- For **insertion**, evaluation starts from the fully perturbed input and progressively restores the same groups in the same order. A faithful attribution should **recover the model response rapidly**.

G.2 Metric Definitions

Let

$$f(X) = \exp\left(\frac{1}{n_{\text{gen}}} \sum_{t=1}^{n_{\text{gen}}} \log p_\theta(y_t \mid X, y_{<t})\right)$$

denote the length-normalised likelihood of the original generated trace under context X, as defined in Section 5.

Let π denote the ordering of the P deletable tokens by decreasing attribution magnitude $|u_i|$, with attribution scores u. We define $X_{\text{del}}^{(k)}$ as the input after perturbing the first k groups under this ordering, and $X_{\text{ins}}^{(k)}$ as the fully perturbed input after restoring the first k groups, for $k = 0, \ldots, K$. We have:

$$f_k^{\text{del}} = f\left(X_{\text{del}}^{(k)}\right), \qquad f_k^{\text{ins}} = f\left(X_{\text{ins}}^{(k)}\right).$$

We normalize each response curve between its fully perturbed and clean endpoints and enforce its expected monotonic direction:

$$r_k^{\text{del}} = \min_{j \le k} \frac{f_j^{\text{del}} - f_K^{\text{del}}}{f_0^{\text{del}} - f_K^{\text{del}}}, \qquad r_k^{\text{ins}} = \max_{j \le k} \frac{f_j^{\text{ins}} - f_0^{\text{ins}}}{f_K^{\text{ins}} - f_0^{\text{ins}}}.$$

This allows the normalized responses to be bounded to $[0, 1]$. We can define our metrics as:

G.2.1 RISE

RISE (Petsiuk et al., 2018) measures the area under the normalized perturbation curve:

$$\mathrm{RISE}_{\mathrm{del}} = \mathrm{AUC}\left(r^{\mathrm{del}}\right)\downarrow, \qquad \mathrm{RISE}_{\mathrm{ins}} = \mathrm{AUC}\left(r^{\mathrm{ins}}\right)\uparrow .$$

A faithful attribution should remove important evidence early under deletion, causing the response to decrease rapidly, and restore it early under insertion, causing the response to recover rapidly. Thus, **lower deletion RISE** and **higher insertion RISE** indicate **better faithfulness**. Notably, RISE depends only on the attribution ranking and does not account for attribution magnitude.

G.2.2 MAS

Magnitude Aligned Scoring (MAS) (Chase Walker et al., 2024) additionally evaluates whether attribution magnitude agrees with the observed model response. Let $\mathcal{G}_k$ denote the token group modified at step k. The fraction of attribution mass restored after insertion step k is given by:

$$m_k^{\mathrm{ins}} = \frac{\sum_{j=1}^{k}\sum_{i\in\mathcal{G}_j}|u_i|}{\sum_i |u_i|},$$

while the mass remaining after deletion is $m_k^{\mathrm{del}} = 1 - m_k^{\mathrm{ins}}$. The mismatch between attribution mass and model response is thus given by:

$$d_k^{\mathrm{del}} = \left|r_k^{\mathrm{del}} - m_k^{\mathrm{del}}\right|, \qquad d_k^{\mathrm{ins}} = \left|r_k^{\mathrm{ins}} - m_k^{\mathrm{ins}}\right|.$$

MAS incorporates this mismatch into the corresponding perturbation curve:

$$\mathrm{MAS}_{\mathrm{del}} = \mathrm{AUC}\left(r^{\mathrm{del}} + d^{\mathrm{del}}\right)\downarrow, \qquad \mathrm{MAS}_{\mathrm{ins}} = \mathrm{AUC}\left(r^{\mathrm{ins}} - d^{\mathrm{ins}}\right)\uparrow .$$

MAS therefore considers both attribution ranking and magnitude, where a faithful attribution should assign attribution mass in proportion to the observed change in model response. As with RISE, **lower deletion** and **higher insertion scores** indicate **better faithfulness**. In implementation, the normalized response and MAS curves are clipped to $[0, 1]$, and AUC is computed using the trapezoidal rule.

G.3 System Prompt

A fixed system prompt is used to generate reasoning traces. Specifically, the system prompt instructs the model to reason step by step over the visual evidence before providing the final answer.

Prompt Template for Reasoning Trace Generation

System:

Look at the image carefully and reason step by step, then end with a line 'Final answer: <answer>'.

User:

`{image} {question}`

H Extended Experiments

H.1 Main Results

Table 5 and Table 6 report the complete Image-variant and Joint-variant faithfulness results with both RISE and MAS under insertion and deletion on Qwen3-VL-8B. VTrace consistently outperforms the baselines for both the image and joint variants. VTrace improves MAS insertion scores by 9.9% and 18.4% for image and join variant respectively, and reduces the deletion scores by 7.7% and 29.6%. The results turn out VTrace manages to identify more faithful tokens that are important for constructing the reasoning.

Table 5: Attribution faithfulness of the **Image variant** on Qwen3-VL-8B across six benchmarks. We report RISE and MAS insertion ($_{\text{ins}}\uparrow$, higher is better) and deletion ($_{\text{del}}\downarrow$, lower is better) AUC and misalignment scores. The Image variant perturbs image patch tokens only. Best results are highlighted in **red bold**, and second-best results are highlighted in blue underlining.

Dataset	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
MMStar	$\text{RISE}_{\text{ins}}\uparrow$	0.505	0.497	0.532	0.539	0.539	0.563	0.555	**0.600**
	$\text{MAS}_{\text{ins}}\uparrow$	0.336	0.329	0.368	0.384	0.399	0.392	0.410	**0.460**
	$\text{RISE}_{\text{del}}\downarrow$	0.458	0.463	0.412	0.409	0.439	0.384	0.394	**0.352**
	$\text{MAS}_{\text{del}}\downarrow$	0.620	0.625	0.563	0.551	0.566	0.547	0.528	**0.489**
MathVista	$\text{RISE}_{\text{ins}}\uparrow$	0.500	0.514	0.579	0.591	0.584	0.599	0.600	**0.662**
	$\text{MAS}_{\text{ins}}\uparrow$	0.334	0.340	0.421	0.449	0.452	0.427	0.464	**0.535**
	$\text{RISE}_{\text{del}}\downarrow$	0.447	0.444	0.368	0.363	0.394	0.358	0.354	**0.311**
	$\text{MAS}_{\text{del}}\downarrow$	0.615	0.607	0.513	0.499	0.518	0.521	0.482	**0.446**
MMMU	$\text{RISE}_{\text{ins}}\uparrow$	0.543	0.564	0.595	0.610	0.627	0.612	0.618	**0.665**
	$\text{MAS}_{\text{ins}}\uparrow$	0.381	0.389	0.431	0.461	0.500	0.444	0.479	**0.536**
	$\text{RISE}_{\text{del}}\downarrow$	0.482	0.480	0.438	0.425	0.444	0.419	0.417	**0.379**
	$\text{MAS}_{\text{del}}\downarrow$	0.639	0.643	0.606	0.576	0.580	0.582	0.561	**0.522**
MMMU-Pro	$\text{RISE}_{\text{ins}}\uparrow$	0.527	0.542	0.570	0.572	0.604	0.593	0.583	**0.645**
	$\text{MAS}_{\text{ins}}\uparrow$	0.363	0.359	0.399	0.422	0.468	0.428	0.432	**0.511**
	$\text{RISE}_{\text{del}}\downarrow$	0.472	0.472	0.442	0.435	0.444	0.413	0.429	**0.378**
	$\text{MAS}_{\text{del}}\downarrow$	0.631	0.639	0.609	0.580	0.579	0.569	0.575	**0.517**
MathVerse	$\text{RISE}_{\text{ins}}\uparrow$	0.523	0.565	0.636	0.640	0.624	0.610	0.644	**0.684**
	$\text{MAS}_{\text{ins}}\uparrow$	0.359	0.380	0.475	0.509	0.498	0.445	0.513	**0.562**
	$\text{RISE}_{\text{del}}\downarrow$	0.454	0.430	0.341	0.345	0.384	0.360	0.341	**0.296**
	$\text{MAS}_{\text{del}}\downarrow$	0.614	0.605	0.500	0.479	0.512	0.527	0.476	**0.435**
VisualPuzzles	$\text{RISE}_{\text{ins}}\uparrow$	0.456	0.461	0.468	0.511	0.525	0.531	0.510	**0.557**
	$\text{MAS}_{\text{ins}}\uparrow$	0.296	0.262	0.272	0.337	0.354	0.346	0.333	**0.374**
	$\text{RISE}_{\text{del}}\downarrow$	0.409	0.405	0.379	0.354	0.380	0.324	0.358	**0.309**
	$\text{MAS}_{\text{del}}\downarrow$	0.566	0.613	0.582	0.528	0.522	0.489	0.537	**0.456**

Table 6: Attribution faithfulness of the **Joint variant** on Qwen3-VL-8B across six benchmarks. The Joint variant perturbs both image and text tokens.

Dataset	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
MMStar	$\text{RISE}_{\text{ins}}\uparrow$	0.371	0.489	0.506	0.507	0.501	0.529	0.554	**0.581**
	$\text{MAS}_{\text{ins}}\uparrow$	0.165	0.277	0.299	0.305	0.314	0.314	0.349	**0.397**
	$\text{RISE}_{\text{del}}\downarrow$	0.325	0.245	0.226	0.222	0.273	0.219	0.204	**0.180**
	$\text{MAS}_{\text{del}}\downarrow$	0.494	0.408	0.372	0.364	0.384	0.341	0.332	**0.246**
MathVista	$\text{RISE}_{\text{ins}}\uparrow$	0.382	0.467	0.517	0.516	0.501	0.512	0.574	**0.647**
	$\text{MAS}_{\text{ins}}\uparrow$	0.186	0.268	0.316	0.324	0.327	0.304	0.377	**0.497**
	$\text{RISE}_{\text{del}}\downarrow$	0.345	0.307	0.274	0.269	0.331	0.274	0.241	**0.178**
	$\text{MAS}_{\text{del}}\downarrow$	0.521	0.494	0.435	0.427	0.464	0.422	0.379	**0.246**
MMMU	$\text{RISE}_{\text{ins}}\uparrow$	0.368	0.589	0.583	0.608	0.604	0.615	0.629	**0.660**
	$\text{MAS}_{\text{ins}}\uparrow$	0.170	0.394	0.393	0.429	0.450	0.421	0.436	**0.503**
	$\text{RISE}_{\text{del}}\downarrow$	0.315	0.216	0.204	0.195	0.241	0.191	0.178	**0.159**
	$\text{MAS}_{\text{del}}\downarrow$	0.468	0.356	0.343	0.323	0.334	0.294	0.292	**0.215**
MMMU-Pro	$\text{RISE}_{\text{ins}}\uparrow$	0.371	0.558	0.532	0.539	0.554	0.559	0.584	**0.619**
	$\text{MAS}_{\text{ins}}\uparrow$	0.166	0.348	0.338	0.351	0.379	0.349	0.384	**0.446**
	$\text{RISE}_{\text{del}}\downarrow$	0.322	0.241	0.242	0.242	0.269	0.229	0.212	**0.181**
	$\text{MAS}_{\text{del}}\downarrow$	0.477	0.401	0.403	0.402	0.377	0.355	0.350	**0.246**
MathVerse	$\text{RISE}_{\text{ins}}\uparrow$	0.383	0.495	0.539	0.540	0.513	0.547	0.568	**0.606**
	$\text{MAS}_{\text{ins}}\uparrow$	0.193	0.297	0.337	0.345	0.345	0.331	0.350	**0.434**
	$\text{RISE}_{\text{del}}\downarrow$	0.330	0.314	0.279	0.283	0.342	0.274	0.241	**0.187**
	$\text{MAS}_{\text{del}}\downarrow$	0.489	0.508	0.438	0.447	0.485	0.416	0.375	**0.254**
VisualPuzzles	$\text{RISE}_{\text{ins}}\uparrow$	0.376	0.488	0.477	0.504	0.523	0.529	0.530	**0.571**
	$\text{MAS}_{\text{ins}}\uparrow$	0.187	0.264	0.260	0.303	0.323	0.321	0.319	**0.366**
	$\text{RISE}_{\text{del}}\downarrow$	0.348	0.288	0.291	0.283	0.283	0.255	0.268	**0.207**
	$\text{MAS}_{\text{del}}\downarrow$	0.494	0.487	0.479	0.459	0.430	0.400	0.434	**0.292**

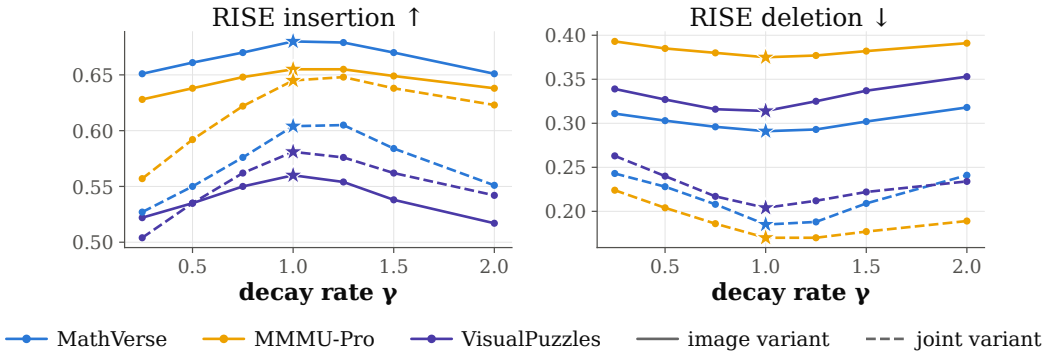


Figure 9: Sensitivity analysis for VTRACE to the decay rate γ.

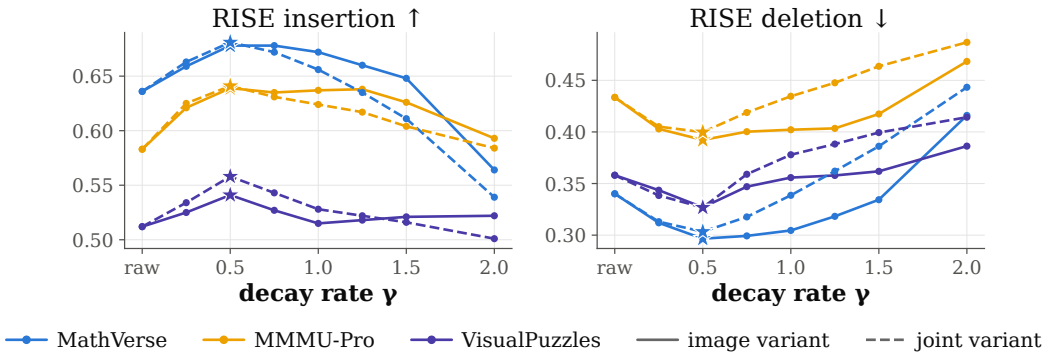


Figure 10: Sensitivity analysis for IFR to the decay rate γ.

H.2 Hyperparameter Sensitivity Study

We conduct sensitivity study for decay rate γ. As shown in Figure 9, $\gamma = 1.0$ achieves the strongest performance across benchmarks, indicating that propagating attribution through intermediate tokens helps recover source evidence missed by direct attribution. To examine whether the proposed multihop aggregation generalizes beyond VTRACE, we additionally apply it to the baseline method IFR Ferrando & Voita (2024). As shown in Figure 10, IFR exhibits a similar sensitivity pattern, with the strongest overall performance around $\gamma = 0.5$. The results indicate that the proposed multi-hop aggregation can also enhance existing attribution methods.

H.3 Full Results on Generalizability Across LVLM Families and Sizes

We report the complete attribution results on InternVL3.5-2B, Qwen3-VL-4B, and InternVL3.5-8B in Tables 7 to 9. The evaluation includes both RISE and MAS scores under the Image and Joint variant. Across different LVLM families and model scales, VTRACE consistently outperforms the baseline attribution methods and demonstrates strong generalizability.

Table 7: Attribution faithfulness on **Qwen3-VL-4B**.

Dataset	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
Image variant									
MMStar	$RISE_{ins}$ ↑	0.531	0.531	0.531	0.525	0.555	0.575	0.539	**0.614**
	MAS_{ins} ↑	0.382	0.364	0.368	0.376	0.421	0.419	0.399	**0.482**
	$RISE_{del}$ ↓	0.480	0.453	0.454	0.462	0.452	0.418	0.447	**0.386**
	MAS_{del} ↓	0.631	0.620	0.614	0.609	0.586	0.577	0.587	**0.531**
MathVerse	$RISE_{ins}$ ↑	0.490	0.600	0.612	0.611	0.618	0.596	0.617	**0.673**
	MAS_{ins} ↑	0.321	0.434	0.453	0.474	0.498	0.438	0.483	**0.553**
	$RISE_{del}$ ↓	0.411	0.294	0.288	0.285	0.313	0.290	0.277	**0.230**
	MAS_{del} ↓	0.570	0.425	0.415	0.398	0.417	0.422	0.388	**0.345**
VisualPuzzles	$RISE_{ins}$ ↑	0.470	0.463	0.441	0.445	0.514	0.521	0.457	**0.551**
	MAS_{ins} ↑	0.319	0.267	0.247	0.256	0.350	0.335	0.270	**0.379**
	$RISE_{del}$ ↓	0.425	0.381	0.406	0.405	0.379	0.345	0.398	**0.321**
	MAS_{del} ↓	0.556	0.599	0.624	0.616	0.528	0.519	0.602	**0.470**
Joint variant									
MMStar	$RISE_{ins}$ ↑	0.335	0.510	0.503	0.503	0.491	0.516	0.526	**0.548**
	MAS_{ins} ↑	0.149	0.310	0.300	0.301	0.326	0.314	0.317	**0.369**
	$RISE_{del}$ ↓	0.289	0.192	0.181	0.183	0.240	0.169	0.166	**0.157**
	MAS_{del} ↓	0.417	0.311	0.296	0.304	0.329	0.249	0.274	**0.214**
MathVerse	$RISE_{ins}$ ↑	0.323	0.533	0.550	0.554	0.480	0.534	0.559	**0.579**
	MAS_{ins} ↑	0.142	0.345	0.344	0.356	0.326	0.328	0.333	**0.398**
	$RISE_{del}$ ↓	0.248	0.195	0.171	0.173	0.305	0.174	0.148	**0.116**
	MAS_{del} ↓	0.403	0.310	0.260	0.266	0.414	0.245	0.229	**0.198**
VisualPuzzles	$RISE_{ins}$ ↑	0.355	0.505	0.464	0.481	0.514	0.520	0.489	**0.546**
	MAS_{ins} ↑	0.174	0.281	0.243	0.260	0.328	0.313	0.259	**0.344**
	$RISE_{del}$ ↓	0.320	0.236	0.262	0.261	0.243	0.229	0.234	**0.188**
	MAS_{del} ↓	0.446	0.405	0.446	0.443	0.360	0.361	0.395	**0.265**

Table 8: Attribution faithfulness on **InternVL3.5-2B**.

Dataset	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
Image variant									
MMStar	$RISE_{ins}$ ↑	0.523	0.639	0.592	0.614	0.659	0.618	0.635	**0.694**
	MAS_{ins} ↑	0.348	0.492	0.459	0.486	0.554	0.477	0.520	**0.588**
	$RISE_{del}$ ↓	0.478	0.381	0.411	0.388	0.387	0.384	0.372	**0.327**
	MAS_{del} ↓	0.650	0.533	0.547	0.524	0.510	0.532	0.499	**0.461**
MathVerse	$RISE_{ins}$ ↑	0.405	0.653	0.623	0.659	0.680	0.616	0.672	**0.709**
	MAS_{ins} ↑	0.245	0.512	0.487	0.540	0.581	0.478	0.561	**0.601**
	$RISE_{del}$ ↓	0.310	0.181	0.193	0.173	0.198	0.173	0.165	**0.139**
	MAS_{del} ↓	0.443	0.265	0.293	0.266	0.306	0.261	0.266	**0.239**
VisualPuzzles	$RISE_{ins}$ ↑	0.464	0.562	0.511	0.512	0.599	0.544	0.534	**0.641**
	MAS_{ins} ↑	0.260	0.368	0.328	0.322	0.449	0.356	0.355	**0.492**
	$RISE_{del}$ ↓	0.401	0.322	0.370	0.360	0.327	0.321	0.343	**0.256**
	MAS_{del} ↓	0.588	0.488	0.542	0.547	0.450	0.478	0.517	**0.372**
Joint variant									
MMStar	$RISE_{ins}$ ↑	0.469	**0.714**	0.677	0.682	0.686	0.694	0.706	0.707
	MAS_{ins} ↑	0.254	0.548	0.505	0.516	0.580	0.518	0.541	**0.585**
	$RISE_{del}$ ↓	0.372	0.191	0.204	0.199	0.241	0.192	0.194	**0.189**
	MAS_{del} ↓	0.587	0.321	0.339	0.333	0.316	0.308	0.327	**0.291**
MathVerse	$RISE_{ins}$ ↑	0.481	0.765	0.741	0.749	0.728	0.763	0.759	**0.799**
	MAS_{ins} ↑	0.252	0.626	0.603	0.618	0.642	0.627	0.610	**0.722**
	$RISE_{del}$ ↓	0.358	0.200	0.199	0.195	0.250	0.175	0.172	**0.153**
	MAS_{del} ↓	0.581	0.311	0.305	0.298	0.321	0.263	0.271	**0.209**
VisualPuzzles	$RISE_{ins}$ ↑	0.433	0.668	0.605	0.603	0.668	0.613	0.631	**0.690**
	MAS_{ins} ↑	0.217	0.477	0.409	0.413	0.550	0.417	0.434	**0.557**
	$RISE_{del}$ ↓	0.350	0.221	0.254	0.249	0.242	0.219	0.237	**0.193**
	MAS_{del} ↓	0.556	0.351	0.409	0.403	0.310	0.332	0.385	**0.256**

Table 9: Attribution faithfulness on **InternVL3.5-8B**.

Dataset	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
Image variant									
MMStar	$RISE_{ins}\uparrow$	0.529	0.684	0.637	0.679	0.681	0.646	0.694	**0.726**
	$MAS_{ins}\uparrow$	0.364	0.570	0.506	0.572	0.580	0.517	0.592	**0.633**
	$RISE_{del}\downarrow$	0.479	0.344	0.365	0.338	0.370	0.367	0.325	**0.303**
	$MAS_{del}\downarrow$	0.643	0.475	0.505	0.462	0.491	0.509	0.447	**0.426**
MathVerse	$RISE_{ins}\uparrow$	0.421	0.718	0.695	0.732	0.733	0.662	0.742	**0.753**
	$MAS_{ins}\uparrow$	0.298	0.620	0.579	0.632	0.624	0.545	0.648	**0.661**
	$RISE_{del}\downarrow$	0.322	0.170	0.174	0.157	0.174	0.189	0.154	**0.144**
	$MAS_{del}\downarrow$	0.456	0.260	**0.248**	0.284	0.331	0.269	0.270	0.251
VisualPuzzles	$RISE_{ins}\uparrow$	0.415	0.565	0.494	0.543	**0.622**	0.540	0.550	0.617
	$MAS_{ins}\uparrow$	0.256	0.414	0.330	0.399	**0.508**	0.372	0.408	0.487
	$RISE_{del}\downarrow$	0.379	0.278	0.345	0.304	0.270	0.278	0.296	**0.239**
	$MAS_{del}\downarrow$	0.529	0.404	0.512	0.455	0.378	0.404	0.438	**0.353**
Joint variant									
MMStar	$RISE_{ins}\uparrow$	0.400	0.727	0.700	0.715	0.537	0.693	0.708	**0.760**
	$MAS_{ins}\uparrow$	0.184	0.583	0.548	0.578	0.361	0.532	0.543	**0.657**
	$RISE_{del}\downarrow$	0.286	0.135	0.130	0.123	0.245	0.132	**0.101**	0.103
	$MAS_{del}\downarrow$	0.457	0.212	0.204	0.193	0.341	0.194	0.157	**0.149**
MathVerse	$RISE_{ins}\uparrow$	0.457	0.786	0.775	0.780	0.673	0.760	0.794	**0.854**
	$MAS_{ins}\uparrow$	0.272	0.675	0.675	0.685	0.540	0.639	0.681	**0.800**
	$RISE_{del}\downarrow$	0.354	0.190	0.181	0.173	0.331	0.182	0.147	**0.116**
	$MAS_{del}\downarrow$	0.513	0.278	0.266	0.254	0.456	0.266	0.217	**0.168**
VisualPuzzles	$RISE_{ins}\uparrow$	0.376	0.648	0.578	0.603	0.586	0.588	0.576	**0.666**
	$MAS_{ins}\uparrow$	0.181	0.470	0.407	0.447	0.420	0.401	0.359	**0.531**
	$RISE_{del}\downarrow$	0.304	0.189	0.223	0.208	0.207	0.184	0.181	**0.135**
	$MAS_{del}\downarrow$	0.452	0.295	0.362	0.337	0.280	0.274	0.299	**0.188**

Table 10: Attribution faithfulness on correctly and incorrectly answered samples for the **Joint variant**. RISE and MAS are reported under insertion ($_{ins}\uparrow$, higher is better) and deletion ($_{del}\downarrow$, lower is better). Best results are highlighted in **red bold**, and second-best results are highlighted in blue underlining.

Dataset	Split	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
MMStar	Correct	$RISE_{ins}\uparrow$	0.371	0.482	0.504	0.504	0.497	0.521	0.552	**0.581**
		$MAS_{ins}\uparrow$	0.166	0.271	0.297	0.303	0.310	0.306	0.347	**0.400**
		$RISE_{del}\downarrow$	0.326	0.251	0.230	0.225	0.278	0.225	0.207	**0.181**
		$MAS_{del}\downarrow$	0.494	0.417	0.377	0.369	0.390	0.351	0.336	**0.248**
	Incorrect	$RISE_{ins}\uparrow$	0.371	0.511	0.514	0.514	0.513	0.551	0.559	**0.580**
		$MAS_{ins}\uparrow$	0.164	0.295	0.305	0.311	0.324	0.337	0.352	**0.389**
		$RISE_{del}\downarrow$	0.322	0.227	0.216	0.212	0.257	0.202	0.195	**0.177**
		$MAS_{del}\downarrow$	0.492	0.381	0.358	0.350	0.365	0.315	0.320	**0.241**
MathVista	Correct	$RISE_{ins}\uparrow$	0.376	0.460	0.510	0.507	0.492	0.501	0.567	**0.644**
		$MAS_{ins}\uparrow$	0.177	0.262	0.309	0.314	0.317	0.292	0.370	**0.495**
		$RISE_{del}\downarrow$	0.341	0.308	0.276	0.272	0.336	0.276	0.242	**0.174**
		$MAS_{del}\downarrow$	0.522	0.496	0.438	0.430	0.470	0.424	0.381	**0.241**
	Incorrect	$RISE_{ins}\uparrow$	0.410	0.501	0.551	0.560	0.542	0.565	0.606	**0.657**
		$MAS_{ins}\uparrow$	0.228	0.299	0.347	0.373	0.376	0.361	0.410	**0.502**
		$RISE_{del}\downarrow$	0.364	0.299	0.264	0.257	0.310	0.265	0.234	**0.194**
		$MAS_{del}\downarrow$	0.519	0.484	0.420	0.409	0.437	0.413	0.370	**0.269**
MMMU	Correct	$RISE_{ins}\uparrow$	0.368	0.583	0.576	0.599	0.593	0.606	0.623	**0.654**
		$MAS_{ins}\uparrow$	0.169	0.386	0.383	0.415	0.436	0.409	0.427	**0.494**
		$RISE_{del}\downarrow$	0.314	0.217	0.206	0.197	0.246	0.193	0.179	**0.159**
		$MAS_{del}\downarrow$	0.466	0.359	0.347	0.327	0.341	0.297	0.296	**0.216**
	Incorrect	$RISE_{ins}\uparrow$	0.370	0.613	0.607	0.637	0.644	0.645	0.650	**0.682**
		$MAS_{ins}\uparrow$	0.176	0.425	0.428	0.473	0.500	0.464	0.468	**0.532**
		$RISE_{del}\downarrow$	0.322	0.211	0.199	0.187	0.223	0.183	0.171	**0.156**
		$MAS_{del}\downarrow$	0.474	0.343	0.330	0.308	0.310	0.282	0.279	**0.212**

H.4 Full Results on Robustness to Prediction Correctness

We evaluate attribution faithfulness separately on correctly and incorrectly answered samples to examine whether the attribution quality depends on prediction correctness. Tables 10 and 11 show the

Table 11: Attribution faithfulness on correctly and incorrectly answered samples for the **Image variant**. RISE and MAS are reported under insertion ($_{\text{ins}}\uparrow$, higher is better) and deletion ($_{\text{del}}\downarrow$, lower is better). Best results are highlighted in **red bold**, and second-best results are highlighted in blue underlining.

Dataset	Split	Metric	ReAGent	HETA	FlowTracer	IFR	Attn Rollout	AttnLRP	FlashTrace	VTRACE
MMStar	Correct	$\text{RISE}_{\text{ins}}\uparrow$	0.499	0.493	0.534	0.542	0.541	0.560	0.556	**0.603**
		$\text{MAS}_{\text{ins}}\uparrow$	0.328	0.326	0.371	0.390	0.403	0.389	0.414	**0.466**
		$\text{RISE}_{\text{del}}\downarrow$	0.453	0.459	0.402	0.399	0.429	0.380	0.386	**0.344**
		$\text{MAS}_{\text{del}}\downarrow$	0.616	0.618	0.549	0.536	0.552	0.539	0.514	**0.478**
	Incorrect	$\text{RISE}_{\text{ins}}\uparrow$	0.525	0.509	0.526	0.531	0.531	0.572	0.550	**0.591**
		$\text{MAS}_{\text{ins}}\uparrow$	0.357	0.336	0.360	0.369	0.388	0.399	0.399	**0.442**
		$\text{RISE}_{\text{del}}\downarrow$	0.471	0.477	0.440	0.437	0.469	0.397	0.418	**0.374**
		$\text{MAS}_{\text{del}}\downarrow$	0.632	0.646	0.603	0.595	0.606	0.568	0.566	**0.520**
MathVista	Correct	$\text{RISE}_{\text{ins}}\uparrow$	0.496	0.509	0.576	0.586	0.580	0.597	0.597	**0.663**
		$\text{MAS}_{\text{ins}}\uparrow$	0.326	0.335	0.418	0.443	0.446	0.424	0.459	**0.537**
		$\text{RISE}_{\text{del}}\downarrow$	0.440	0.439	0.358	0.356	0.388	0.348	0.345	**0.297**
		$\text{MAS}_{\text{del}}\downarrow$	0.611	0.599	0.500	0.489	0.510	0.506	0.470	**0.429**
	Incorrect	$\text{RISE}_{\text{ins}}\uparrow$	0.523	0.539	0.590	0.614	0.603	0.611	0.617	**0.655**
		$\text{MAS}_{\text{ins}}\uparrow$	0.373	0.364	0.434	0.480	0.478	0.442	0.485	**0.523**
		$\text{RISE}_{\text{del}}\downarrow$	0.483	0.469	0.416	0.396	0.422	0.405	0.394	**0.373**
		$\text{MAS}_{\text{del}}\downarrow$	0.634	0.644	0.575	0.546	0.558	0.588	0.538	**0.525**
MMMU	Correct	$\text{RISE}_{\text{ins}}\uparrow$	0.533	0.562	0.591	0.606	0.624	0.607	0.616	**0.666**
		$\text{MAS}_{\text{ins}}\uparrow$	0.372	0.389	0.430	0.458	0.496	0.440	0.479	**0.540**
		$\text{RISE}_{\text{del}}\downarrow$	0.470	0.463	0.421	0.409	0.428	0.403	0.400	**0.359**
		$\text{MAS}_{\text{del}}\downarrow$	0.626	0.623	0.585	0.555	0.560	0.566	0.539	**0.497**
	Incorrect	$\text{RISE}_{\text{ins}}\uparrow$	0.575	0.567	0.608	0.625	0.638	0.627	0.627	**0.661**
		$\text{MAS}_{\text{ins}}\uparrow$	0.412	0.390	0.435	0.472	0.513	0.456	0.481	**0.523**
		$\text{RISE}_{\text{del}}\downarrow$	0.522	0.541	0.493	0.477	0.497	0.470	0.472	**0.448**
		$\text{MAS}_{\text{del}}\downarrow$	0.683	0.717	0.676	0.647	0.647	0.637	0.634	**0.606**

complete results for the Image and Joint Variants, where VTRACE achieves the best performance on both correct and incorrect predictions. The results indicate that the attribution gains arise from attribution method rather than answer correctness.

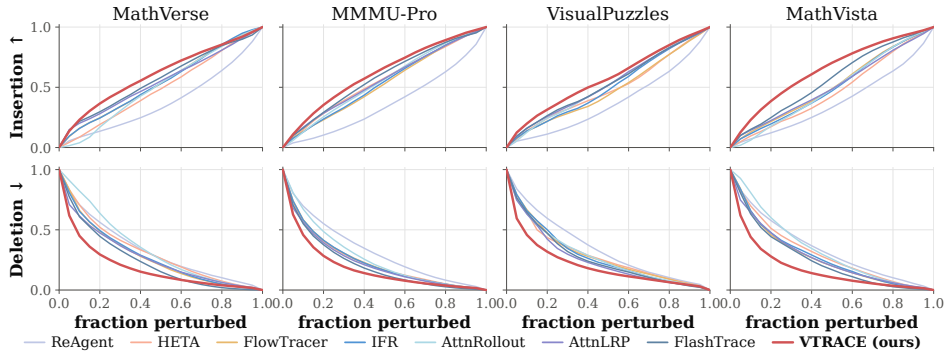


Figure 11: Fine-grained insertion and deletion perturbation curves across benchmarks.

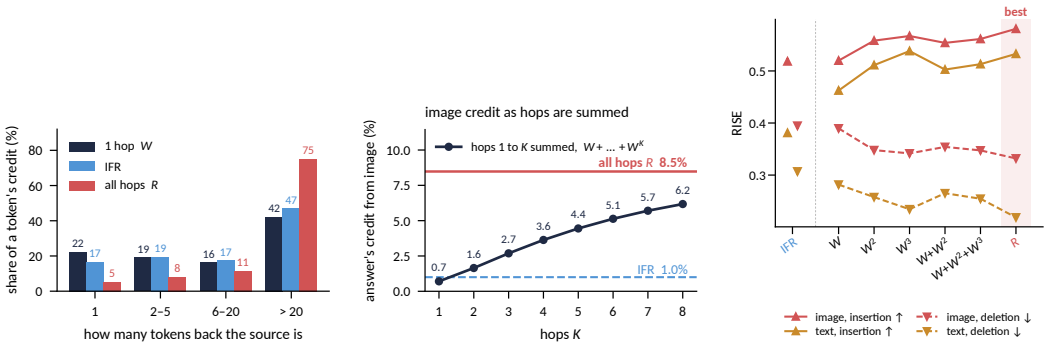
H.5 Fine-Grained Perturbation Curves

VTRACE produces more faithful rankings throughout the perturbation process, beyond what is captured by the AUC scores. Shown in Figure Figure 11, VTRACE achieves faster recovery under insertion and sharper degradation under deletion. The advantage is especially clear at early perturbation stages, where restoring only a small fraction of top-ranked tokens rapidly recovers the response, while removing them causes substantial degradation.

H.6 Summing paths carries credit back to the inputs.

Figure 12a shows how far back a generated token's credit comes from. Evidently, a single hop keeps credit near the receiver: under W, 22% of a token's credit falls on the token directly before it and

58% on the 20 tokens before it (IFR: 17% and 53%). Once VTRACE sums all direct and indirect paths (R), the token directly before keeps 5%, and 75% of the credit comes from sources more than 20 positions back. For a generated token, those sources are mostly the question, the image and the early reasoning. Figure 12b tracks how much of the answer's credit lands on the image as longer paths are added. The direct edge gives the image 0.7% and IFR gives it 1.0%. The share rises with every hop, to 1.6% with paths of up to two edges and 6.2% with paths of up to eight, and $\mathbb{R}$ reaches 8.5%. The direct edge therefore sees almost none of the image, and the image share is still rising after eight hops. Figure 12c tests whether this extra credit is faithful. We rank the image patches and the question tokens separately by each operator's score, then remove them in that order (deletion) or add them back to a fully masked input (insertion) and measure the RISE. Notably, longer paths help in all four settings, and VTRACE, which aggregates all paths performs the best, as expected.



(a) Source distance (b) Image credit vs. hops (c) RISE by operator

Figure 12: (a) How far back a generated token's credit comes from. (b) The answer's image credit as hops are summed. (c) Performance comparison between diverse hops, VTRACE (R) and IFR.

I CASE STUDIES

Image token attribution. Figure 13 compares the top 20% image patches selected by each method against a perturbation-based reference. For each patch, we blur it and measure the resulting drop in the model's trace likelihood, where a larger drop indicates greater reliance on that region. This reference map is shown in the second column. The reported score on the bottom right measures how much of the reference importance is captured by the selected patches, with higher values indicating better alignment. **We demonstrate two better performing and two slightly behind cases of VTRACE.** In the first two examples, VTRACE focuses more strongly on the relevant people and scene regions, scoring 0.50 and 0.42 compared with 0.46 and 0.35 for the strongest baseline. The third example is tied with FlowTracer. The fourth shows a slightly less aligned case, while both VTRACE and AttnLRP both captures the important balls, VTRACE also assigns attribution to the player and table edge. This suggests that VTRACE can miss important visual regions when they are not clearly reflected in the generated reasoning trace.

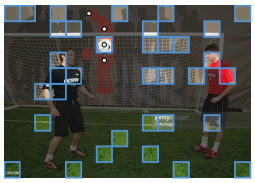
Joint image and text attribution. Figures 14 to 17 show four case studies, comparing IFR, a second baseline and VTRACE on the same fixed model response, with the boxed final answer (without loss of generality) selected as the target. In each row, the image on the left and the question text outlines the method's top 20% of the image tokens, and dims the rest. The text on the right shows the question and the full response, with each token shaded by its attribution score: the darker the token, the more it contributed to the answer.

A faithful attribution should point to the evidence the answer relies on: the image regions the question is about, the key words of the question, and the reasoning steps that carry what the model saw to its answer. Evidently, VTRACE does this more consistently than the baselines. Its top patches concentrate on the relevant objects and labels, while the baselines often spread their patches over the background and the image border. In the text, VTRACE highlights the key question words and the reasoning steps that use visual evidence, whereas the baselines focus mostly on the answer options and the final sentence. Figure 17 shows a harder case, where VTRACE's image selection is less focused. See the captions for detailed description.



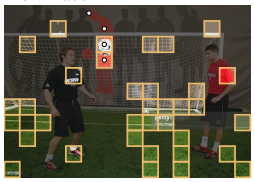
Figure 13: **Image token attribution on four MMStar questions.** Each method's top 20% of patches is outlined, with the share of the measured dependence it captures in the corner (higher is better). The measured dependence is the likelihood drop from blurring each patch in turn. VTRACE leads on the first two questions, ties FlowTracer on the third, and trails AttnLRP on the fourth.

IFR



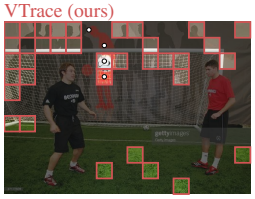
Q: What is the sport being played in the image? Options: A: Tennis, B: Soccer, C: Volleyball, D: Basketball
A: Step1: Identify key elements in the image. - There is a large net with a goal frame, which is characteristic of a soccer goal. - A soccer ball is visible, positioned in front of the goal. - Two men are standing on a green field that resembles artificial turf, commonly used in soccer. - The players are wearing athletic attire, including shorts and cleats, typical for soccer. - The background has silhouettes of people and a large red figure, possibly a mascot or part of a training setup, but this does not change the sport. Step2: Eliminate incorrect options. - Tennis: Played on a hard or grass court with rackets and a ball, no net with a goal frame. - Volleyball: Played on a court with a net, but the net is lower and there is no goal frame or soccer ball. - Basketball: Played on a court with a hoop and backboard, not a net with a goal frame. Step3: Confirm the sport. - The presence of a soccer goal, a soccer ball, and players on a field strongly indicates the sport is soccer (football). Final answer: B: Soccer

FlowTracer



Q: What is the sport being played in the image? Options: A: Tennis, B: Soccer, C: Volleyball, D: Basketball
A: Step1: Identify key elements in the image. - There is a large net with a goal frame, which is characteristic of a soccer goal. - A soccer ball is visible, positioned in front of the goal. - Two men are standing on a green field that resembles artificial turf, commonly used in soccer. - The players are wearing athletic attire, including shorts and cleats, typical for soccer. - The background has silhouettes of people and a large red figure, possibly a mascot or part of a training setup, but this does not change the sport. Step2: Eliminate incorrect options. - Tennis: Played on a hard or grass court with rackets and a ball, no net with a goal frame. - Volleyball: Played on a court with a net, but the net is lower and there is no goal frame or soccer ball. - Basketball: Played on a court with a hoop and backboard, not a net with a goal frame. Step3: Confirm the sport. - The presence of a soccer goal, a soccer ball, and players on a field strongly indicates the sport is soccer (football). Final answer: B: Soccer

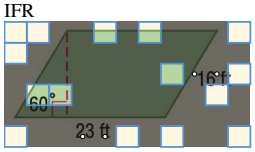
VTrace (ours)



Q: What is the sport being played in the image? Options: A: Tennis, B: Soccer, C: Volleyball, D: Basketball
A: Step1: Identify key elements in the image. - There is a large net with a goal frame, which is characteristic of a soccer goal. - A soccer ball is visible, positioned in front of the goal. - Two men are standing on a green field that resembles artificial turf, commonly used in soccer. - The players are wearing athletic attire, including shorts and cleats, typical for soccer. - The background has silhouettes of people and a large red figure, possibly a mascot or part of a training setup, but this does not change the sport. Step2: Eliminate incorrect options. - Tennis: Played on a hard or grass court with rackets and a ball, no net with a goal frame. - Volleyball: Played on a court with a net, but the net is lower and there is no goal frame or soccer ball. - Basketball: Played on a court with a hoop and backboard, not a net with a goal frame. Step3: Confirm the sport. - The presence of a soccer goal, a soccer ball, and players on a field strongly indicates the sport is soccer (football). Final answer: B: Soccer

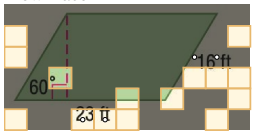
Figure 14: **Recognizing the sport.** The model answers "Soccer". VTRACE places most of its top patches on the goal net, the object that identifies the sport, while IFR and FlowTracer spread more patches over the players and the grass. In the text, VTRACE scores the option "Soccer" and the concluding "soccer" highest, and also shades the scene the model describes, such as "large net", "goal frame" and "soccer ball"; FlowTracer concentrates on the question wording.

IFR



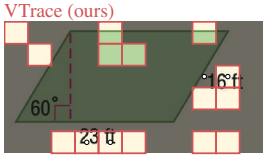
Q: Find the perimeter of the parallelogram. Choices: (A)32(B)39(C)46(D)78
A: To find the perimeter of the parallelogram, we need to add up the lengths of all four sides. From the image, we are given: - One side is23 ft. - The adjacent side is16 ft. In a parallelogram, opposite sides are equal in length. Therefore: - The side opposite the2 3 ft side is also23 ft. - The side opposite the16 ft side is also16 ft. So, the perimeter P is: P =23 ft +16 ft +23 ft +16 ftP =2 ×23 ft +2 ×16 ftP =46 ft +32 ftP =78 ftThe angle of60° and the dashed line (which appears to be the height) are not needed to calculate the perimeter, as perimeter only depends on the side lengths. Looking at the choices: (A)32(B)39(C)46(D)78The correct answer is78 ft. Final answer: D

FlowTracer



Q: Find the perimeter of the parallelogram. Choices: (A)32(B)39(C)46(D)78
A: To find the perimeter of the parallelogram, we need to add up the lengths of all four sides. From the image, we are given: - One side is23 ft. - The adjacent side is16 ft. In a parallelogram, opposite sides are equal in length. Therefore: - The side opposite the2 3 ft side is also23 ft. - The side opposite the16 ft side is also16 ft. So, the perimeter P is: P =23 ft +16 ft +23 ft +16 ftP =2 ×23 ft +2 ×16 ftP =46 ft +32 ftP =78 ftThe angle of60° and the dashed line (which appears to be the height) are not needed to calculate the perimeter, as perimeter only depends on the side lengths. Looking at the choices: (A)32(B)39(C)46(D)78The correct answer is78 ft. Final answer: D

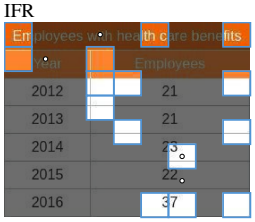
VTrace (ours)



Q: Find the perimeter of the parallelogram. Choices: (A)32(B)39(C)46(D)78
A: To find the perimeter of the parallelogram, we need to add up the lengths of all four sides. From the image, we are given: - One side is23 ft. - The adjacent side is16 ft. In a parallelogram, opposite sides are equal in length. Therefore: - The side opposite the2 3 ft side is also23 ft. - The side opposite the16 ft side is also16 ft. So, the perimeter P is: P =23 ft +16 ft +23 ft +16 ftP =2 ×23 ft +2 ×16 ftP =46 ft +32 ftP =78 ftThe angle of60° and the dashed line (which appears to be the height) are not needed to calculate the perimeter, as perimeter only depends on the side lengths. Looking at the choices: (A)32(B)39(C)46(D)78The correct answer is78 ft. Final answer: D

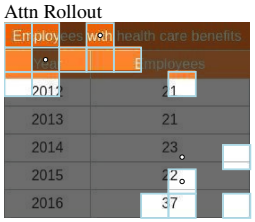
Figure 15: **Reading the side lengths of a parallelogram.** The model adds the side lengths shown in the image, $2 \times 23 + 2 \times 16 = 78$, and answers (D). VTRACE's top patches cover the "23 ft" and "16 ft" labels, while IFR's lie mostly along the image border. In the text, IFR highlights little beyond the answer options, whereas VTRACE highlights "perimeter" in the question and the side lengths and sums in the reasoning that lead to the answer.

IFR



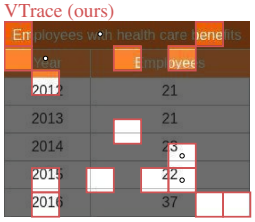
Q: At Bloomington Consulting, the head of human resources examined how the number of employees with health care benefits varied in response to policy changes. According to the table, what was the rate of change between2014 and2015? (Unit: employees per year) Choices: (A) -1(B)1(C)0(D) -2
A: To find the rate of change between2014 and2015, we need to calculate the change in the number of employees divided by the change in time. From the table: - In2014, the number of employees with health care benefits was23. - In2015, the number of employees with health care benefits was22. The change in employees = Employees in 2015 - Employees in2014 =22 -23 = -1. The change in time =2015 -2014 =1 year. Therefore, the rate of change = (Change in employees) / (Change in time) = -1 /1 = -1 employees per year. This means the number of employees with health care benefits decreased by1 per year between2014 and2015. Looking at the choices: (A) -1(B)1(C) 0(D) -2The correct answer is (A) -1. Final answer: A

Attn Rollout



Q: At Bloomington Consulting, the head of human resources examined how the number of employees with health care benefits varied in response to policy changes. According to the table, what was the rate of change between2014 and2015? (Unit: employees per year) Choices: (A) -1(B)1(C)0(D) -2
A: To find the rate of change between2014 and2015, we need to calculate the change in the number of employees divided by the change in time. From the table: - In2014, the number of employees with health care benefits was23. - In2015, the number of employees with health care benefits was22. The change in employees = Employees in 2015 - Employees in2014 =22 -23 = -1. The change in time =2015 -2014 =1 year. Therefore, the rate of change = (Change in employees) / (Change in time) = -1 /1 = -1 employees per year. This means the number of employees with health care benefits decreased by1 per year between2014 and2015. Looking at the choices: (A) -1(B)1(C) 0(D) -2The correct answer is (A) -1. Final answer: A

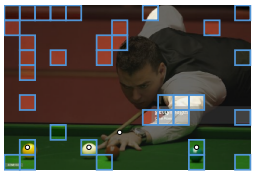
VTrace (ours)



Q: At Bloomington Consulting, the head of human resources examined how the number of employees with health care benefits varied in response to policy changes. According to the table, what was the rate of change between2014 and2015? (Unit: employees per year) Choices: (A) -1(B)1(C)0(D) -2
A: To find the rate of change between2014 and2015, we need to calculate the change in the number of employees divided by the change in time. From the table: - In2014, the number of employees with health care benefits was23. - In2015, the number of employees with health care benefits was22. The change in employees = Employees in 2015 - Employees in2014 =22 -23 = -1. The change in time =2015 -2014 =1 year. Therefore, the rate of change = (Change in employees) / (Change in time) = -1 /1 = -1 employees per year. This means the number of employees with health care benefits decreased by1 per year between2014 and2015. Looking at the choices: (A) -1(B)1(C) 0(D) -2The correct answer is (A) -1. Final answer: A

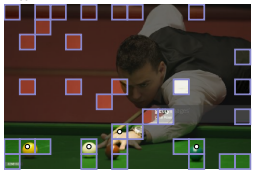
Figure 16: **Reading values from a table.** The model reads that the count fell from 23 in 2014 to 22 in 2015 and answers (A) -1. VTRACE's top patches fall on the year column and the 2014 and 2015 rows, while Attention Rollout concentrates on the table's title bar. In the text, Attention Rollout highlights setup words such as "Bloomington Consulting" and "calculate", whereas VTRACE highlights the years, the change in employees and the result -1.

IFR



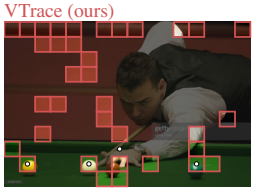
Q: What is the color of the ball closest to the man? Options: A: White, B: Red, C: Blue, D: Brown
A: Looking at the image, we can see a snooker player leaning over the table, preparing to take a shot. His cue stick is positioned near the cue ball. The balls on the table include: - A white cue ball (the one he is aiming to strike) - A red ball (which is the object ball he is likely targeting) - A yellow ball- A blue ball- A brown ballThe question asks for the color of the ball closest to the man. The man is leaning over the table, and his hand is near the cue ball. The cue ball is white. The red ball is positioned very close to the cue ball, and also very close to the man's hand and body. However, the cue ball is the one he is directly interacting with (holding the cue stick near it), and it is the ball he is about to strike. The red ball is also very close, but the cue ball is the one he is aiming at. Looking at the positions: the cue ball is directly in front of his hand, and the red ball is just to the right of the cue ball. Since the man is leaning over, his body is closest to the cue ball. The cue ball is the ball he is physically closest to, as he is about to strike it. Therefore, the ball closest to the man is the white cue ball. Final answer: A

AttnLRP



Q: What is the color of the ball closest to the man? Options: A: White, B: Red, C: Blue, D: Brown
A: Looking at the image, we can see a snooker player leaning over the table, preparing to take a shot. His cue stick is positioned near the cue ball. The balls on the table include: - A white cue ball (the one he is aiming to strike) - A red ball (which is the object ball he is likely targeting) - A yellow ball- A blue ball- A brown ballThe question asks for the color of the ball closest to the man. The man is leaning over the table, and his hand is near the cue ball. The cue ball is white. The red ball is positioned very close to the cue ball, and also very close to the man's hand and body. However, the cue ball is the one he is directly interacting with (holding the cue stick near it), and it is the ball he is about to strike. The red ball is also very close, but the cue ball is the one he is aiming at. Looking at the positions: the cue ball is directly in front of his hand, and the red ball is just to the right of the cue ball. Since the man is leaning over, his body is closest to the cue ball. The cue ball is the ball he is physically closest to, as he is about to strike it. Therefore, the ball closest to the man is the white cue ball. Final answer: A

VTrace (ours)



Q: What is the color of the ball closest to the man? Options: A: White, B: Red, C: Blue, D: Brown
A: Looking at the image, we can see a snooker player leaning over the table, preparing to take a shot. His cue stick is positioned near the cue ball. The balls on the table include: - A white cue ball (the one he is aiming to strike) - A red ball (which is the object ball he is likely targeting) - A yellow ball- A blue ball- A brown ballThe question asks for the color of the ball closest to the man. The man is leaning over the table, and his hand is near the cue ball. The cue ball is white. The red ball is positioned very close to the cue ball, and also very close to the man's hand and body. However, the cue ball is the one he is directly interacting with (holding the cue stick near it), and it is the ball he is about to strike. The red ball is also very close, but the cue ball is the one he is aiming at. Looking at the positions: the cue ball is directly in front of his hand, and the red ball is just to the right of the cue ball. Since the man is leaning over, his body is closest to the cue ball. The cue ball is the ball he is physically closest to, as he is about to strike it. Therefore, the ball closest to the man is the white cue ball. Final answer: A

Figure 17: **A harder case: the ball closest to the man.** The model answers (A) White. In the text, VTRACE highlights the key question words "ball closest" and "man" and the reasoning about the white cue ball. In the image, its top patches cover the balls on the table, including the white cue ball and the red ball at the man's hand, but many also fall on the dark background above the table.