Visual Attention Faithfulness in Vision-Language Models is Heterogeneous

Xurui Song[1,2*], **Weishi Wang**[1†], **Zhongqi Yue**[3], **Kuluhan Binici**[1]
Tao Bai[1], **Hongxin Shao**[1], **Daniel Dahlmeier**[1], **Jun Luo**[2]

[1]SAP
[2]College of Computing and Data Science, Nanyang Technological University, Singapore
[3]Microsoft

Abstract

Whether attention weights faithfully reflect model reasoning has been actively debated in NLP, yet this question remains largely unexplored for the visual modality in Vision-Language Models (VLMs). We address this gap through causal perturbation analysis on current VLMs, evaluating both the comprehensiveness and sufficiency gap of attention-ranked visual tokens. Our analysis reveals that visual attention faithfulness is heterogeneous, manifesting in three distinct processing modes: *Faithful-Sufficient*, where top-k attention tokens are both necessary and sufficient for prediction; *Faithful-Distributed*, where they are necessary but broader visual context remains required; and *Non-Focal*, where no localized attention region is individually necessary while visual information remains an essential trigger for prediction. Furthermore, human-annotated ground-truth regions satisfy comprehensiveness in only $\sim 60\%$ of cases compared with model attention rankings, revealing systematic divergence between model visual reliance and human intuition. We demonstrate these patterns across both general VQA on VQAv2 and document tasks on VRDU and ChartQA, showing that visual attention faithfulness varies systematically with processing demands and model architectures rather than being uniformly faithful or unfaithful.

1 Introduction

Despite the widespread use of attention weights as a window into model reasoning (Clark et al., 2019; Chefer et al., 2021), whether attention provides a faithful explanation of how models arrive at their predictions remains an open question. In text-based Natural Language Processing (NLP), this question has sparked sustained debate: evidence suggests that attention distributions can be permuted without changing predictions (Jain and Wallace, 2019), while counter-evidence shows that models with constrained attention behave meaningfully differently (Wiegreffe and Pinter, 2019). A key insight from this line of work is that faithfulness should not be treated as a binary property of attention but rather a graded, intervention-specific property that has to be measured by sample (Jacovi and Goldberg, 2020; DeYoung et al., 2020). Yet this faithfulness debate has been developed in text-only settings, leaving open whether its conclusions transfer to architectures that jointly process visual and textual information.

Vision-Language Models (VLMs) (Liu et al., 2023; Chen et al., 2024b; Bai et al., 2025; OpenAI, 2024) consisting of a vision transformer encoder (Dosovitskiy et al., 2021) and a language backbone (Radford et al., 2018) represent precisely such an architecture. By encoding image patches as visual tokens processed alongside text within a unified transformer (Vaswani et al., 2017), VLMs create a setting where attention operates over inputs with 2D spatial structure and high inter-token redundancy among neighboring patches. However, existing work on VLM attention has proceeded without first establishing whether attention faithfully reflects the model's reliance on visual information. Recent studies either assume that low-attention visual tokens are dispensable for inference acceleration (Chen et al., 2024a; Shang et al., 2025), or analyze attention patterns to characterize internal mechanisms (Golovanevsky et al., 2025; Jiang et al., 2025), yet the fundamental question of whether visual attention rankings correspond to actual model dependencies on specific visual regions remains unexplored.

We address this question through token-level causal perturbation analysis on representative open-source VLMs, Qwen2.5-VL and InternVL2.5 (Bai et al., 2025; Chen et al., 2024b). As illustrated in Figure 1, we ablate visual token embeddings at

[*]Work done during internship at SAP.
[†]Corresponding author.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 26472–26486
October 24-29, 2026 ©2026 Association for Computational Linguistics

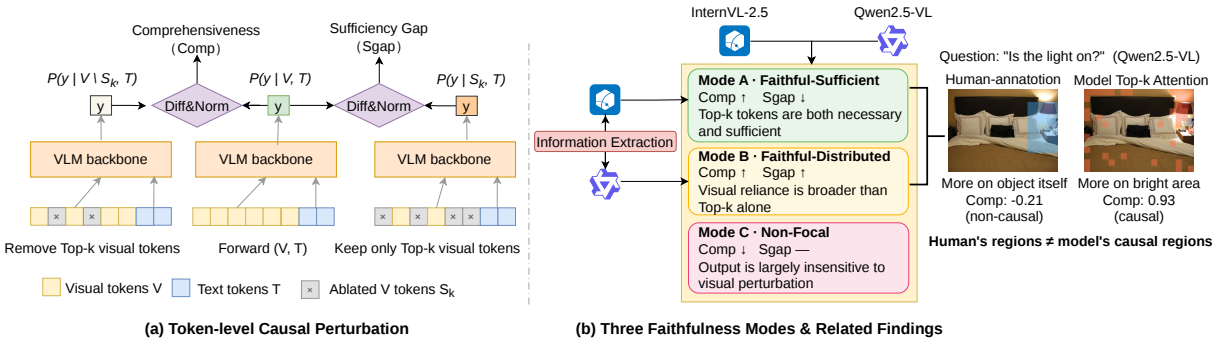


Figure 1: Overview of the proposed evaluation framework and main findings. (Left) Visual attention faithfulness is evaluated using comprehensiveness and the sufficiency gap under token-level perturbations, by respectively removing or retaining top-k attention-ranked visual tokens. (Right) Three faithfulness modes of visual attention of the VLM backbone. Different VLM architectures may handle the same task in different modes, for example, the IE task. Furthermore, what humans perceive as causal regions is not always the focus of the model's attention.

the first decoder layer, either removing or retaining tokens according to their attention ranking, and measure the resulting change in prediction probability. We evaluate both comprehensiveness and sufficiency gap of attention-ranked visual tokens on VQAv2 (Goyal et al., 2017). This evaluation captures whether attention identifies visual tokens that are necessary for the prediction, or whether they alone are sufficient to recover it.

Our analysis reveals that visual attention faithfulness is not uniform but heterogeneous in VLMs, manifesting in three distinct processing modes. In the *Faithful-Sufficient* mode , attention top-k tokens are both necessary and sufficient, indicating that attention captures the visual information the model relies on for these predictions. In the *Faithful-Distributed* mode , top-k tokens are necessary but not sufficient, as the model draws on broader visual context beyond its focal attention region. In the *Non-Focal* , no concentrated attention region is necessary for the output, yet visual information overall remains an essential trigger for prediction, indicating diffusely distributed visual dependencies that attention rankings cannot localize.

Next, we compare model visual attention against human-annotated ground-truth regions indicating which areas humans consider task-relevant. This comparison reveals systematic divergence: ground-truth regions satisfy comprehensiveness in only $\sim$60% of cases , indicating that models develop visual processing strategies that and diverge from the regions humans deem important.

We further validate these patterns on document information extraction (IE) using VRDU (Wang et al., 2023) dataset and document VQA using ChartQA (Masry et al., 2022), practical tasks where VLMs must locate, extract or understand field values from visually structured documents. On these tasks, both Qwen2.5-VL and InternVL2.5 reliably attend to answer-relevant regions, yet exhibit different processing modes. Qwen2.5-VL behaves as an extreme instance of the *Faithful-Distributed* mode, where evidence regions are highly necessary but insufficient without broader layout context. In contrast, InternVL2.5 behaves closer to the *Faithful-Sufficient* mode, where attended regions are both necessary and largely sufficient for prediction. These findings show that the three processing modes extend beyond general VQA to structurally different visual tasks, while also suggesting that different VLM architectures may adopt different visual processing strategies for the same task. Our contributions are as follows:

- We present the first systematic evaluation of visual attention faithfulness in VLMs, measuring both comprehensiveness and sufficiency via causal perturbation.
- We identify three visual attention processing modes that characterize how VLMs rely on visual information, showing that faithfulness is heterogeneous rather than uniform.
- We reveal systematic divergence between model visual attention and human-annotated ground-truth regions, showing that human annotations are an incomplete proxy for model visual dependencies.
- We further show that these modes generalize beyond VQA to document understanding tasks and vary across VLM architectures.

2 Related Work

Attention as Explanation. Prior work on attention interpretability in NLP has produced conflicting conclusions. While Jain and Wallace (2019) and Serrano and Smith (2019) argue that attention weights show weak correspondence with feature importance and can be manipulated without affecting predictions, Wiegreffe and Pinter (2019) contend that these findings do not invalidate attention as an explanation, since constraining attention during training leads to measurably different model behavior. Bastings and Filippova (2020) further questions attention's role by demonstrating that gradient-based saliency methods are better at identifying causal tokens. Meanwhile, attention patterns have been productively used to probe model internals, revealing that specific heads encode syntactic and semantic relations (Clark et al., 2019; Vig, 2019; Michel et al., 2019). These works establish that attention occupies an ambiguous position between faithful explanation and mere computational byproduct, with conclusions drawn exclusively from text-only architectures.

Faithfulness Evaluation. To move beyond qualitative debates, several frameworks have been proposed for quantitatively measuring explanation faithfulness. Jacovi and Goldberg (2020) formalize faithfulness as a graded property requiring empirical verification rather than assumption. DeYoung et al. (2020) operationalize this through comprehensiveness and sufficiency metrics that measure whether highlighted tokens are necessary or sufficient for the prediction, and Madsen et al. (2022) propose recursive masking and retraining for faithfulness assessments. For attention aggregation specifically, Abnar and Zuidema (2020) propose attention rollout and attention flow to aggregate raw attention across layers, addressing that single-layer attention may not reflect overall information routing. These evaluation paradigms provide the methodological foundation for systematically probing attention faithfulness, though they have been developed and validated exclusively on text inputs.

Interpretability in VLMs. Interpretability research on Vision-Language Models has pursued several directions. Visual explanation methods such as Grad-CAM (Selvaraju et al., 2017) and attention-based approaches (Chefer et al., 2021; Aflalo et al., 2022) generate spatial heatmaps highlighting image regions relevant to predictions, but do not verify whether these highlighted regions causally drive model outputs. On the mechanistic side, Golovanevsky et al. (2025) identifies universal cross-attention heads responsible for specific functions like object detection, while Jiang et al. (2025) link attention patterns in middle layers to hallucination phenomena. Recent work has also identified attention sink phenomena in both language models (Xiao et al., 2024) and vision transformers (Darcet et al., 2024), where certain tokens accumulate disproportionate attention regardless of semantic content. A separate line of work leverages attention sparsity for efficiency, assuming that visual tokens with low attention scores can be safely pruned or merged (Chen et al., 2024a; Shang et al., 2025; Zhang et al., 2025; Bolya et al., 2023). However, no prior work has systematically measured whether attention-ranked visual tokens are causally necessary or sufficient for VLM predictions. Our work fills this gap by applying comprehensiveness and sufficiency evaluation to VLM visual attention, revealing heterogeneous faithfulness patterns that these prior approaches do not account for.

3 Method

We formalize the evaluation of visual attention faithfulness in VLMs as a causal measurement problem. Given a VLM that processes both visual and textual tokens, we ask: do the visual tokens that receive the highest attention weights causally correspond to the tokens the model relies on for its predictions? We operationalize this question through three components: attention-based token ranking (§3.2), causal perturbation (§3.3), and faithfulness metrics (§3.4).

3.1 Problem Formulation

Consider a VLM that takes an image I and a text query Q as input. The image is encoded into a sequence of visual tokens $\mathbf{V} = \{v_1, v_2, \ldots, v_N\}$, and the query into text tokens $\mathbf{T} = \{t_1, t_2, \ldots, t_M\}$. The model autoregressively generates an answer $\mathbf{y} = (y_1, \ldots, y_{|\mathbf{y}|})$, with $P(\mathbf{y} \mid \mathbf{V}, \mathbf{T})$ denoting the mean probability assigned to the generated answer tokens. Our goal is to determine whether the attention-derived importance ranking over visual tokens faithfully reflects their causal contribution to $P(\mathbf{y} \mid \mathbf{V}, \mathbf{T})$.

3.2 Attention-based Token Ranking

To obtain a single importance score for each visual token, we aggregate attention weights across all

layers and heads. In a transformer with L layers and H attention heads per layer, let $\alpha_{l,h}(i,j)$ denote the attention weight from token j to token i at layer l, head h. We compute the aggregated attention score for each visual token v_i as:

$$a_i = \sum_{l=1}^{L} \frac{1}{H} \sum_{h=1}^{H} \frac{1}{|\mathbf{y}|} \sum_{s=1}^{|\mathbf{y}|} \alpha_{l,h}(i, y_s). \quad (1)$$

We average over heads as an unbiased aggregation since the dominant routing heads may vary across layers and inputs (Michel et al., 2019), and summing across layers captures the native attention distribution across the network without additional propagation assumptions, making it suitable for assessing visual attention faithfulness. Sorting visual tokens by a_i in descending order and selecting the top-k yields $\mathcal{S}_k = \{v_{(1)}, v_{(2)}, \ldots, v_{(k)}\}$, where (i) denotes the index with the i-th largest score.

3.3 Causal Perturbation

To establish a causal rather than merely correlational link between attention rankings and model predictions, we adopt an interventionist approach: selectively ablating visual tokens and observing the effect on prediction probability. Specifically, for a given token set $\mathcal{S}$, we zero-ablate their hidden states $\mathbf{h}_i^{(0)}$ at the entry of the first decoder layer:

$$\mathbf{h}_i^{(0)} \leftarrow \mathbf{0}, \quad \forall\, v_i \in \mathcal{S}. \quad (2)$$

This zero-ablation introduces no additional information into the representation space, removing the contribution of the ablated tokens from all subsequent computations.

3.4 Faithfulness Metrics

We evaluate attention faithfulness along two complementary dimensions adapted from DeYoung et al. (2020): comprehensiveness measures whether attention-highlighted tokens are *necessary*, while sufficiency gap measures whether they are *sufficient*. Since different samples have varying baseline prediction probabilities $P(\mathbf{y} \mid \mathbf{V}, \mathbf{T})$, we normalize both metrics by this value to ensure comparability across samples.

Comprehensiveness. Removing the top-k attention-ranked tokens should degrade prediction if they are causally important:

$$\text{Comp}(\mathcal{S}_k) = \frac{P(\mathbf{y} \mid \mathbf{V}, \mathbf{T}) - P(\mathbf{y} \mid \mathbf{V} \setminus \mathcal{S}_k, \mathbf{T})}{P(\mathbf{y} \mid \mathbf{V}, \mathbf{T})}. \quad (3)$$

High comprehensiveness indicates that the top-k tokens are necessary for the prediction.

Sufficiency gap. Retaining only the top-k tokens should preserve prediction if they capture all necessary visual information. We define the sufficiency gap (Sgap) as the normalized probability drop after this restriction:

$$\text{Sgap}(\mathcal{S}_k) = \frac{P(\mathbf{y} \mid \mathbf{V}, \mathbf{T}) - P(\mathbf{y} \mid \mathcal{S}_k, \mathbf{T})}{P(\mathbf{y} \mid \mathbf{V}, \mathbf{T})}. \quad (4)$$

A low (or negative) sufficiency gap indicates that the top-k tokens alone are sufficient for the prediction. Together, these two metrics locate each sample in a comprehensiveness–Sgap space that characterizes the degree and nature of attention faithfulness.

4 Experiments and Analysis

4.1 Experiment Setup

Model and Inference. We conduct our main experiments on Qwen2.5-VL-7B, evaluate the generalizability of our findings on InternVL2.5-8B, and further perform scaling experiments with Qwen2.5-VL-3B and InternVL2.5-4B. We use greedy decoding (temperature = 0) throughout to ensure deterministic outputs, so that any changes in $P(\mathbf{y} \mid \mathbf{V}, \mathbf{T})$ can be directly attributed to token ablations rather than sampling variance.

Datasets. We evaluate on two benchmarks that exercise different aspects of visual understanding. For general visual question answering (VQA), we sample 90 image-question pairs from VQAv2 (Goyal et al., 2017), balanced across three question categories (30 yes/no, 30 number, 30 other) to capture a range of reasoning demands from binary judgments to open-ended descriptions. Each sample is accompanied by a manually verified ground-truth (GT) region indicating the image area that humans consider task-relevant (details in Appendix A). For document IE task, we randomly choose 60 samples answered correctly from the VRDU (Wang et al., 2023) spanning two document types (advertisement and registration forms), and we select 100 samples from the ChartQA (Masry et al., 2022) test set for the document VQA task.

Comparison Strategies. To isolate the causal contribution of attention-ranked tokens, for VQA we compare four perturbation strategies applied to the same number of tokens k:

- **Overall Top-K**: the k visual tokens receiving the highest aggregated attention scores a_i (§3.2), representing the model's own importance ranking.
- **GT Object**: the k tokens corresponding to the human-annotated GT region, representing human judgment of task-relevant areas.
- **Non-GT Top-K**: the k highest-attention tokens that fall *outside* the GT region, capturing attention to non-obvious regions.
- **Random**: k tokens randomly sampled at fixed seed, serving as an uninformed baseline.

The token count k is set equal to the number of visual tokens covered by the GT region for each sample, ensuring that all four strategies ablate the same proportion of visual information. This alignment makes comprehensiveness and sufficiency gap scores directly comparable across strategies: any observed differences reflect the importance of the *selected* tokens rather than an artifact of varying ablation size. For IE, instead of GT-aligned comparison, we ablate the top 1%, 3%, 5%, and 10% of visual tokens by attention score to trace the progression of causal effect as coverage grows from the most focal to broader regions. This is because model attention on these documents is already highly localized to the answer regions (§4.5), making a GT-aligned comparison largely redundant, and document layouts contain redundant copies of the same information across regions, which a single GT mask cannot capture. For the CharQA task, we adopt a similar proportion perturbation as IE for comparison.

Evaluation Protocol. For each strategy and each sample, we compute comprehensiveness by zero-ablating the selected tokens and measuring the normalized probability drop, and the sufficiency gap by retaining only the selected tokens while ablating all others (§3.4). Further implementation details and visual strategies of the tested models are described in Appendix A.

4.2 Three Processing Modes

Excluding 5 samples that need whole image understanding (Appendix C), Table 1 reports the results of the four perturbations on the remaining valid samples. Among the four strategies, overall top-k achieves the highest mean comprehensiveness (0.813) and the lowest mean sufficiency gap (0.469), suggesting that attention-ranked tokens are both highly necessary and sufficient for the model's predictions, and that visual attention partially reflects the model's reliance on visual input.

Strategy	Mean Comp	Mean Sgap
Overall Top-k	**0.813**	**0.469**
GT Object	0.505	0.846
Non-GT Top-k	0.659	0.653
Random	0.502	0.599

Table 1: Mean normalized comprehensiveness (Comp) and sufficiency gap (Sgap) of the four perturbation strategies on the valid VQA samples. Each strategy ablates the same number of visual tokens k, equal to the GT region size for the corresponding sample.

We then plot valid samples' overall top-k attention tokens in the (Comp,Sgap) space in Figure 2. The distribution exhibits a bimodal structure along the comprehensiveness axis, with the high-Comp region further separating into two bands of sufficiency gaps. Since comprehensiveness identifies whether a localized causal contributor exists, whereas sufficiency is only meaningful within such focal samples, we apply Otsu's method (Otsu, 1979) sequentially: first on Comp over all samples, and then on Sgap within the high-Comp subset. This produces the three-way partition shown in Figure 2b. We further inspect samples near the decision boundaries and manually refine a small number of assignments based on their attention maps and prediction behavior. The resulting partition (Figure 2c) defines the three groups used throughout the remainder of the analysis: *Faithful-Sufficient* (A), *Faithful-Distributed* (B), and *Non-Focal* (C). Across three modes, the model maintains consistently high answer accuracy (Table 2), suggesting the observed partition reflects genuine differences in the model's reliance on visual attention rather than being affected by incorrectness.

Mode A: Faithful-Sufficient. Mode A accounts for 32.9% of samples and is characterized by near-maximal Comp (0.978) together with a slightly negative Sgap (-0.026) (Table 2), indicating that the top-k attention tokens are both causally necessary and individually sufficient for the prediction, while the remaining tokens contribute little or act as mild distractors. This mode is typically associated with questions answerable from localized object-centric evidence without requiring broader spatial reasoning. A representative example asks for the number of lanes on a road, where the top-

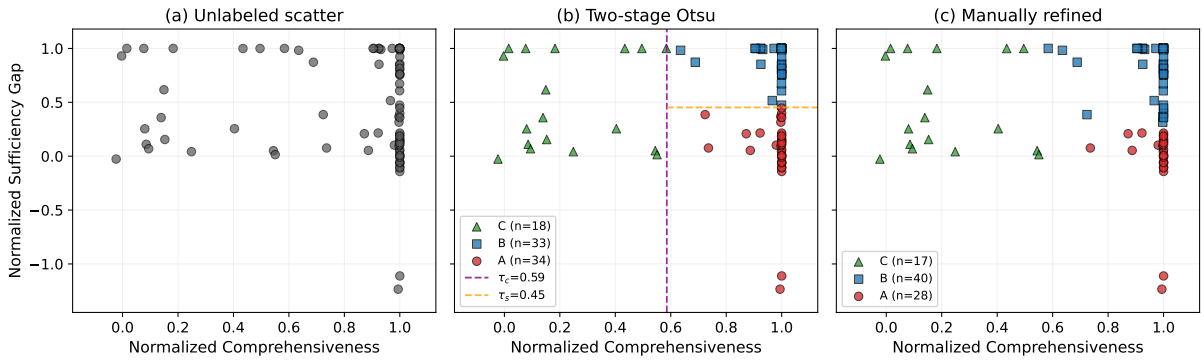


Figure 2: Per-sample distribution of normalized comprehensiveness and sufficiency gap on the valid VQA samples. (a) Unlabeled scatter. (b) Two-stage Otsu thresholding yields a three-way partition. (c) Manually refined partition used in the analysis.

k tokens align with the lane markings; ablating them collapses the prediction, whereas retaining only them preserves it (Figure 3, top row). In this regime, high-attention regions closely correspond to the evidence underlying the model's prediction.

Mode B: Faithful-Distributed. Mode B accounts for 47.1% of samples and is characterized by high Comp (0.952) together with a large Sgap (0.818) (Table 2), indicating that the top-k attention tokens are causally important but insufficient on their own to recover the prediction. Questions in this mode typically require integrating localized evidence with broader spatial or contextual information. For example, in the question *How many kites are flying?*, the top-k tokens concentrate on the kite cluster and remain highly necessary for the prediction, yet retaining only them leaves a substantial sufficiency gap, suggesting that accurate counting additionally depends on surrounding scene context (Figure 3, middle row). In this regime, attention faithfully identifies relevant evidence, but does not fully capture the distributed visual information underlying the model's decision.

Mode C: Non-Focal. Mode C accounts for 20.0% of samples and is characterized by low Comp (0.213) together with a moderate Sgap (0.460), indicating that the top-k attention tokens contribute little to the prediction despite the model still relying on visual input. Full-image ablation yields a Comp of 0.973 while removing all task-irrelevant text tokens such as format instructions only yields a Comp of 0.306, confirming that the image remains necessary even though no localized attention region is decisive. Questions in this mode exhibit limited reliance on specific visual evidence, with the image primarily serving as a contextual co-

Mode	%	Comp	Sgap	Acc
A: Faithful-Sufficient	32.9	0.978	−0.026	92.9
B: Faithful-Distributed	47.1	0.952	0.818	100.0
C: Non-Focal	20.0	0.213	0.460	94.1

Table 2: Per-mode share, mean overall top-k scores and model accuracy (%)[1] on the valid VQA samples.

Question Type	A	B	C
Yes/No	33.3	55.6	11.1
Number	48.3	37.9	13.8
Other	17.2	48.3	34.5

Table 3: Distribution (%) of each VQA question type across the three processing modes.

activation trigger. For example, in a question about laws against cellphone use, the image establishes the cellphone context that activates the model's relevant internal knowledge (Figure 3, bottom row). In this mode, visual attention is not faithful enough, but is indispensable context triggers.

Table 3 shows a clear association between VQAv2 question types and the three modes. Yes/no questions are concentrated in Mode B (55.6%), reflecting decisions that combine a localized target with broader contextual information. Number questions are split between Mode A (48.3%) and Mode B (37.9%): they fall into Mode A when counting can be resolved from localized instances, and into Mode B when it requires additional scene context beyond focal regions. Open-ended "other" questions are most frequently Mode B (48.3%) and exhibit the highest proportion of Mode C (34.5%), where answers are often driven by scene-level cues or image-conditioned knowledge rather than local-

[1]Accuracy per mode is assessed by VQAv2 ground truth.

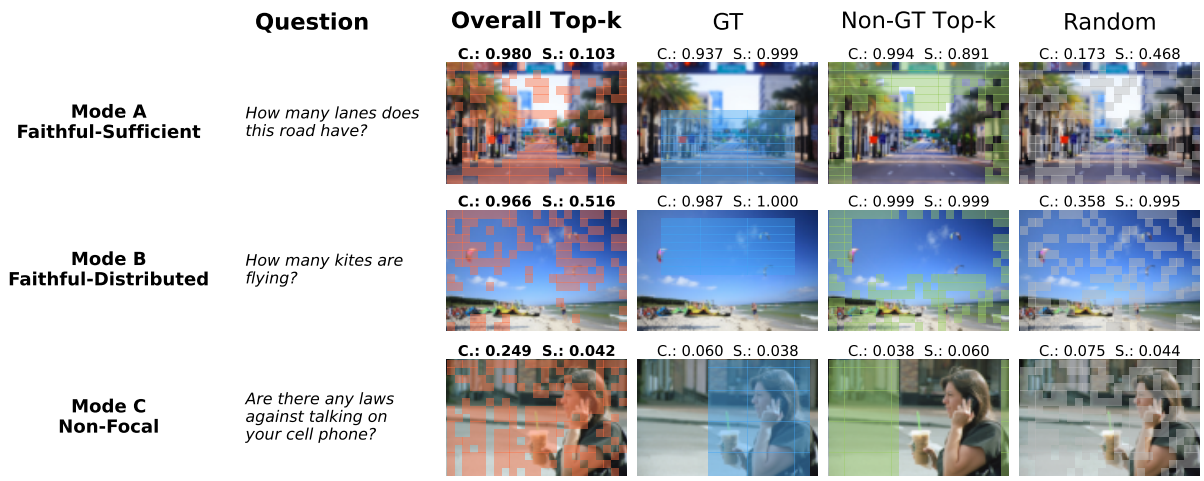


Figure 3: Representative samples for the three processing modes. Each row visualizes the four perturbation strategies on a representative case and reports the per-strategy normalized comprehensiveness (**C.**) and sufficiency gap (**S.**), illustrating why the top-k tokens are sufficient in Mode A, only necessary in Mode B, and almost inert in Mode C.

n	A (%)	B (%)	C (%)
50	33.0 ± 4.2	47.0 ± 4.6	20.1 ± 3.7
60	32.8 ± 3.2	47.1 ± 3.6	20.1 ± 2.8
70	33.0 ± 2.4	47.0 ± 2.6	20.0 ± 2.0
95% CI	[23.9, 43.5]	[36.8, 57.6]	[12.9, 29.7]

Table 4: Stability of the three-mode partition across sample sizes. Values are mean $\pm$ standard deviation; brackets denote the confidence intervals.

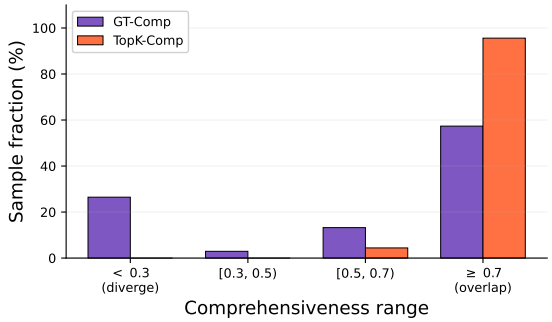


Figure 4: Sample distribution of GT-Comp and Top-k Comp in modes A and B.

ized visual evidence. Overall, the mode structure aligns with the underlying reasoning demands induced by different question types.

4.3 Stability Analyses for Three Modes

As shown in Figure 2 (b), the three-mode structure is already pronounced in the data. To quantify stability, we re-sample the valid samples without replacement at sizes at $n = 50$, 60, and 70, repeating each 1000 times, and report Wilson 95% confidence intervals on the full set in Table 4. The confidence intervals consistently exclude zero for all three modes, supporting the stability of the three-mode structure. Across sample sizes, repeated subsampling also yields highly consistent mode proportions, with the variance decreasing monotonically as n increases. These results indicate that the three-mode partition is a data-driven characterization of this underlying distribution, rather than one induced by a particular sample size.

4.4 GT versus Top-k Divergence

Table 1 shows a substantial gap in comprehensiveness between the GT region and the overall top-k region (0.505 vs. 0.813), suggesting a mismatch between human-annotated evidence and the regions causally used by the model. We further analyze samples from Modes A and B, where overall top-k comprehensiveness is consistently high. As shown in Figure 4, GT comprehensiveness exhibits a bimodal distribution rather than concentrating at high values. The bimodal GT-Comp distribution separates samples where human-annotated regions align with the model's causal evidence from those where they do not.

In the aligned subset, mean GT-Comp reaches 0.963, close to the mean Top-k Comp of 0.992, indicating that both regions capture the same causal signal, and the model answer accuracy is high at 97.4%. As shown in Figure 5 (top row), for the question *Are there red lights?*, both the GT mask and the top-k tokens concentrate on the lights, and ablating either region sharply reduces the answer probability. In contrast, the misaligned subset ex-

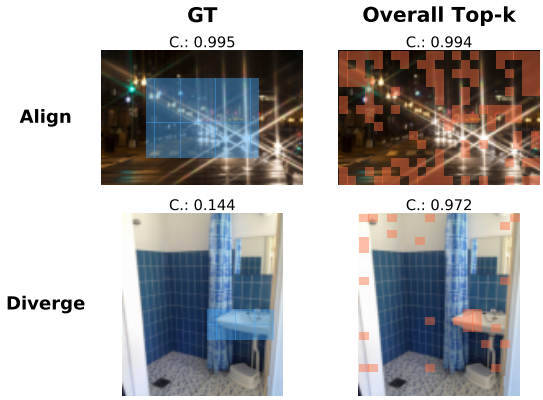


Figure 5: Representative samples for the two GT versus Top-k regimes. Top row: GT and Top-k overlap on the salient answer object, and either ablation collapses the prediction. Bottom row: GT covers the named object but ablating it leaves the prediction nearly unchanged, while ablating the Top-k region collapses it.

hibits a mean GT-Comp of -0.150 while Top-k Comp remains high at 0.945, indicating that the human-annotated region contributes little to the prediction even when strong causal evidence is captured by the model's top-k regions, and the model answer accuracy is also high at 94.4%. Figure 5 (bottom row) shows a representative example for the question *What color is the sink on the right?*: ablating the GT sink region leaves the prediction largely unchanged, whereas ablating the top-k tokens substantially reduces the answer probability. This suggests that the model resolves the color attribute from surrounding contextual regions carrying correlated chromatic cues rather than from the object region itself.

These results highlight a key aspect of the model's visual reasoning: while it exhibits partial alignment with human visual interpretation, it also relies on broader contextual dependencies beyond semantically corresponding regions. The model's own attention more faithfully reflects its internal visual reliance than human-annotated GT regions. Therefore, treating GT masks from human understanding as a faithfulness ground truth for VLMs is insufficient.

4.5 From VQA to Document Tasks

We further conduct the perturbation framework on document information extraction (IE) task, a practical application, where layout, repetition, and field structure replace natural-scene visual cues. Following the VQA formulation, each extraction target is cast as a question over the document, as illustrated

Strategy	Mask Ratio		
	5%	10%	30%
Top-k	**0.37 / 1.00**	**0.49 / 1.00**	**0.77 / 0.91**
Random	0.01 / 1.00	0.02 / 1.00	0.27 / 0.99

Table 5: ChartQA causal perturbation on Qwen2.5-VL-7B. Each cell reports Comp / Sgap

in Figure 6 (a). For example, for the query *Who is the advertiser?* in a campaign-finance document, the attention heatmap concentrates on the answer string "Committee to Elect Mike Carr", with minimal activation elsewhere on the page. This setting enables top-k-based perturbation analysis without relying on explicit GT region annotations.

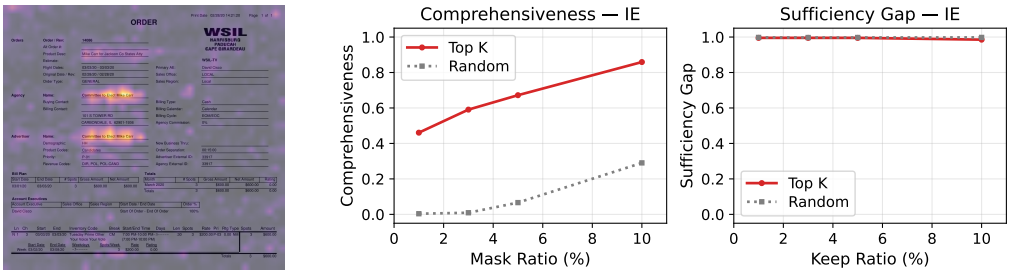
As shown in Figure 6 (b), ablating only the top 1% of visual tokens already reduces the answer probability by 43.8%, and comprehensiveness increases to 0.826 at the top 10%, indicating that a small focal region drives most of the prediction. In contrast, retaining only the top-k tokens recovers at most $\sim 5\%$ of the prediction probability (Sgap ~ 0.947), suggesting strong dependence on broader document context. Therefore, Qwen-VL-2.5 treats IE samples as an extreme instance of Mode B, where localized evidence is necessary but insufficient without layout-level context.

For ChartQA samples in Table 5, as the perturbation ratio increases, the top-k comprehensiveness rises steadily and stays well above random, while the sufficiency gap remains close to 1, so the ChartQA task also falls into Mode B.

These results extend the three-mode framework beyond VQA to document understanding, while suggesting practical implications for attention-aware token selection in VLM systems, where preserving broader layout context may remain essential.

4.6 Cross-Model Generalizability

We further replicate the perturbation pipeline on InternVL2.5-8B, a VLM with a different LM backbone and a vision tower employing dynamic-tile preprocessing. On VQA, the overall top-k visual tokens again attain the highest mean comprehensiveness (0.733) and the lowest mean sufficiency gap (0.468) among the four strategies, so attention-ranked tokens remain the most faithful selection; applying two-stage Otsu thresholding to InternVL's own distribution recovers the same three-mode partition, although the per-sample mode assignment



(a) Attention heatmap

(b) Comp and Sgap vs. perturbation ratio

Figure 6: (a) Attention concentrates on the answer string for the IE question *Who is the advertiser?*. (b) Top-k ablation (red) yields high Comp while a random baseline (grey) stays near zero; the Sgap remains near one for both, so the focal region is necessary but not sufficient for IE tasks on Qwen-VL-2.5.

	Comp		Mode (%)		
Model	Top-k	GT	A	B	C
Qwen2.5-VL-3B	0.739	0.432	31.8	37.6	30.6
Qwen2.5-VL-7B	0.813	0.505	32.9	47.1	20.0
InternVL2.5-4B	0.792	0.636	18.8	54.1	27.1
InternVL2.5-8B	0.733	0.609	29.4	45.9	24.7

Table 6: Attention faithfulness across model scales.

overlaps only partially with Qwen's. On IE, attention is again concentrated on the answer string and comprehensiveness grows monotonically with the mask ratio as on Qwen, but the sufficiency gap decreases steadily and reaches 0.189 at the top 10%, placing InternVL IE closer to mode A . On ChartQA, the trend is the same as the IE results and displays the mode A. These results suggest that heterogeneous faithfulness is a model-agnostic phenomenon, while which mode a particular task falls into depends on the underlying architecture, since different visual encoders and LM backbones may concentrate attention over different regions of the same image. Detailed results and discussion is provided in Appendix D.

4.7 Scaling Across Model Sizes

To assess the consistency of our findings across model scales, we replicate the VQA experiments on Qwen2.5-VL-3B and InternVL2.5-4B under the same protocol and compare them with their larger counterparts. As shown in Table 6, all three attention modes are observed across model sizes, and the attention-ranked top-k regions consistently achieve higher comprehensiveness than the human-annotated GT regions. These results suggest that the observed faithfulness patterns are consistent across model scales.

5 Conclusion

We systematically evaluate visual attention faithfulness in VLMs through token-level causal perturbation. Our results show that visual attention faithfulness is not binary, but instead manifests in three processing modes with different patterns of visual dependence, and different VLM architectures may process the same task under different modes. Furthermore, we find that human-annotated regions only align with the regions causally used by the model on some tasks, while the model relies more on hidden semantically corresponding regions for others. Experiments on document information extraction demonstrate that these patterns extend beyond general VQA to structured practical visual tasks. Together, these findings provide a unified perspective on how the visual attention faithfulness is manifested in VLMs.

Limitations

This work focuses on current VLMs with dynamic-resolution visual strategies, and may not directly extend to alternative schemes, such as models with fixed-resolution image inputs (e.g., LLaVA-style models). Second, we do not explicitly study reasoning-oriented VLMs that perform multi-step deliberation before answering, which may exhibit different visual attention behaviors, and we leave this direction for future work.

References

Samira Abnar and Willem H. Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4190–4197. Association for Computational Linguistics.

Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. 2022. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 21374–21383. IEEE.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2020, Online, November 2020*, pages 149–155. Association for Computational Linguistics.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your vit but faster. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 387–396. IEEE.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXI*, Lecture Notes in Computer Science, pages 19–35. Springer.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 21 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. 2024. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4443–4458. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Michal Golovanevsky, William Rudman, Vedant Palit, Carsten Eickhoff, and Ritambhara Singh. 2025. What do vlms notice? A mechanistic interpretability pipeline for gaussian-noise-free text-image corruption and evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 11462–11482. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *CoRR*, abs/2404.15255.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4198–4205. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

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

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, Lecture Notes in Computer Science, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Andreas Madsen, Nicholas Meade, Vaibhav Adlakha, and Siva Reddy. 2022. Evaluating the faithfulness of importance measures in NLP by recursively masking allegedly important tokens and retraining. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Findings of ACL, pages 1731–1751. Association for Computational Linguistics.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Paul Michel, Omer Levy, and Graham Neubig. 2019. Are sixteen heads really better than one? In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 14014–14024.

OpenAI. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Nobuyuki Otsu. 1979. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 618–626. IEEE Computer Society.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2931–2951. Association for Computational Linguistics.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2025, Honolulu, HI, USA, October 19-25, 2025*, pages 22857–22867. IEEE.

Rheeya Uppaal, Phu Mon Htut, Min Bai, Nikolaos Pappas, Zheng Qi, and Sandesh Swamy. 2026. 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)*, pages 4147–4168, Rabat, Morocco. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*, pages 37–42. Association for Computational Linguistics.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023. VRDU: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 5184–5193. ACM.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 11–20. Association for Computational Linguistics.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A. Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2025. Sparse-vlm: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Sample	10%	30%	50%	Mode
Is there a toy?	1.00/1.00	0.99/1.00	1.00/0.99	B
Is there rope around?	1.00/1.00	1.00/0.97	1.00/0.82	B
Are these beignets?	1.00/1.00	1.00/0.50	1.00/0.38	B
How many snowboards?	0.03/1.00	0.10/0.38	0.11/0.00	C
What room is this?	0.43/0.51	0.46/0.21	0.51/0.05	C

Table 7: Attention-ranked top-k ablation on the five excluded VQA samples. Each cell reports Comp/Sgap.

A Details of Experiment Setup

Ground-Truth Region Annotation. For each of the 90 VQA samples, we define a ground-truth (GT) region as the set of visual tokens corresponding to the image area that a human would consider directly relevant to answering the question. Annotations are derived through a semi-automatic pipeline: we begin with COCO instance segmentation masks (Lin et al., 2014) overlaid on VQAv2 images, then manually select and refine bounding rectangles to cover exactly the task-relevant objects or areas (e.g., the road surface for "How many lanes?" or the traffic light for "Is there a red light?"). Each annotation is expressed as a set of grid coordinates over the visual token grid, enabling direct correspondence between annotated regions and token positions. For the annotation process, first, a human expert, referencing the VQAv2 questions, images, and answers, and using COCO instance segmentation masks as a reference, selects the visual regions that can ultimately contribute to the answer. Then, a second human expert reviews the annotated image data in conjunction with the VQAv2 questions and answers, adjusting any that were inaccurate.

Visual Token Grid. Qwen2.5-VL encodes each image into a spatial grid of visual tokens through patch embedding followed by 2×2 spatial merging, producing grid dimensions that vary with image resolution. Each language-visible token corresponds to a localized image region, and the flattened token ordering preserves the underlying 2D spatial arrangement through positional encoding. InternVL2.5 instead employs a dynamic tile-based preprocessor: the input image is partitioned into a variable number of 448×448 tiles together with a downscaled thumbnail tile. Each tile is independently encoded by the vision tower and transformed into a fixed-length token block via spatial downsampling, after which the per-tile blocks are concatenated and separated by tile-boundary tokens. The thumbnail tile provides an additional global-context representation alongside the locally constrained tile tokens. In both models, the language backbone receives a flat sequence of visual tokens whose ordering encodes the underlying spatial structure.

B Discussion

Perturbation design. We perturb visual tokens at the entry of the first decoder layer of the language-model backbone, rather than masking image pixels or intervening inside the vision encoder. Since the object of evaluation is the language model's attention over visual tokens, the intervention is placed at the position where the language model reads those tokens, so that the measured probability change reflects the language model's reliance on a specific token rather than the vision encoder's response to a corrupted input. Pixel- or encoder-side ablations characterize the input dependencies of the vision encoder and offer a complementary view. Our setting is also align with recent token-level processing of visual tokens in VLMs (Chen et al., 2024a; Shang et al., 2025; Zhang et al., 2025). For the token-level perturbation, we replace the targeted hidden state with the zero vector instead of removing the token from the sequence: visual tokens occupy positions on a 2D spatial grid, and deleting a token would shift the positions of the remaining tokens and thereby distort the encoded layout, conflating the loss of a token's content with the loss of spatial structure. Zero-ablation preserves sequence length and positional structure while suppressing the content at the targeted positions, and is one of the standard activation-level interventions discussed alongside mean and Gaussian-noise ablation in activation-patching (Heimersheim and Nanda, 2024).

Sample selection. We evaluate the framework on two task families that exercise different aspects of visual understanding. VQAv2 (Goyal et al., 2017) spans three question categories (yes/no, number, open-ended) associated with diverse visual-reliance patterns within a single benchmark. VRDU (Wang

et al., 2023) complements VQAv2 with a structurally different visual regime, where text-rich layouts, field-level evidence, and naturally occurring redundancy test whether the framework transfers beyond natural images. Despite the relatively small evaluation scale, the three-mode structure exhibits clear separation in the $(\mathrm{Comp}, \mathrm{Sgap})$ space (Table 2), is reproduced on InternVL2.5-8B, and remains consistent for the five excluded VQA samples under a separate ratio-based protocol (Appendix C). These observations suggest that the identified processing modes reflect stable behavioral patterns rather than dataset-specific artifacts.

Potential broader impact. The differences between Qwen2.5-VL (Mode B) and InternVL2.5 (closer to Mode A) on the same IE task suggest that different VLM backbones may adopt different visual processing preferences for the same task. This implies that visual attention faithfulness is not solely task-dependent but is partly model-dependent, which may affect the transferability of attention-aware token selection, compression, and interpretability methods across architectures. Our findings also suggest that human-annotated regions should not be treated as universal proxies for model visual reliance in real-world VLM applications.

Model selection rationale. For model backbones, the two kinds of models, Qwen2.5VL and InternVL2.5, are chosen to represent two widely adopted paradigms for current dynamic-resolution VLMs. Qwen2.5-VL adopts native-resolution encoding, where the entire image is processed without explicit tiling, producing a variable number of visual tokens that scales with image resolution while preserving the original aspect ratio. InternVL2.5 adopts dynamic tiling, decomposing an image into a variable number of fixed-resolution tiles that are encoded separately. These together cover the mainstream design space of dynamic-resolution VLMs.

Comparison with attention visualization. Existing attention visualization methods primarily characterize where a model attends by presenting attention maps, whereas our work evaluates whether the attended regions faithfully explain the model's prediction. Our goal is therefore not to find a visualization method, but to explore the explanatory faithfulness of existing visual attention. Our causal evaluation further reveals cases where model-attended regions differ from human-annotated evidence in their contribution to the prediction, and enables the identification of distinct faithfulness modes that cannot be determined from attention maps alone.

	Comp		**Sgap**	
Strategy	Qwen	InternVL	Qwen	InternVL
GT Object	0.505	0.609	0.846	0.555
Non-GT Top-k	0.659	0.474	0.653	0.659
Overall Top-k	**0.813**	**0.733**	**0.469**	**0.468**
Random	0.502	0.487	0.599	0.549

Table 8: Mean normalized comprehensiveness (Comp) and sufficiency gap (Sgap) of the four perturbation strategies on valid VQA samples. Overall top-k remains the strategy with the highest Comp and the lowest Sgap on both models.

Relation to gradient-based saliency. Gradient-based saliency and attention provide visual importance estimates from different signals: saliency is derived from gradients, whereas attention is obtained directly from the transformer's forward attention weights. Comparing their relative attribution quality addresses the question of attribution methods, whereas our focus is specifically on the faithfulness of visual attention itself in VLMs. We therefore directly evaluate attention through causal interventions, revealing heterogeneous faithfulness across samples and models. Extending the analysis to gradient-based attribution methods is complementary to the present study.

Comparison with other work. Uppaal et al. (2026) also studies faithfulness in VLMs, but focuses on whether the explicit reasoning is grounded in the input image, relying on black-box methods such as self-reflection. In contrast, we investigate a mechanism-level question: whether the model's internal attention over visual tokens is faithful. We evaluate visual attention faithfulness through token-level ablation and quantify both comprehensiveness and sufficiency. Thus, while both works study visual faithfulness, we address a different object of faithfulness at a different level: we study internal attention faithfulness rather than the faithfulness of explicit reasoning.

C Excluded VQA Samples

Five of the 90 collected VQA samples are excluded from the main analysis because their human-annotated GT regions cover almost the entire image, leaving little area for the GT-aligned strategies to be meaningfully different from full-image

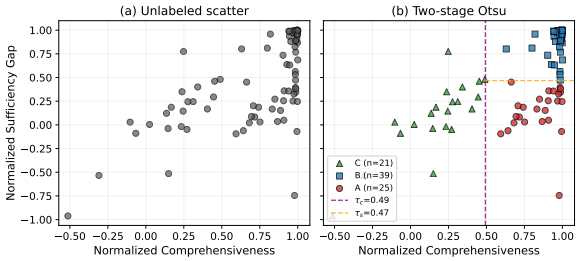


Figure 7: Processing-mode scatter on InternVL2.5-8B.

Strategy	Mask Ratio		
	5%	10%	30%
Top-k	**0.65 / 0.42**	**0.71 / 0.26**	**0.79 / 0.08**
Random	0.01 / 0.82	0.01 / 0.74	0.06 / 0.39

Table 9: ChartQA causal perturbation on InternVL2.5-8B. Each cell reports Comp / Sgap. Retaining only the top-30% attention tokens already suffices (NSgap drops to 0.08), placing chart understanding closer to the Faithful-Sufficient regime (mode A).

ablation. To further validate if our found faithfulness framework still works for these samples, we re-evaluate them on Qwen2.5-VL-7B using the ratio-based perturbation procedure: for $k \in \{10\%, 30\%, 50\%\}$ of all visual tokens, we mask the top-k ranked by attention and measure Comp and Sgap directly. From Table 7, three samples fall into Mode B (Faithful-Distributed): their Comp stays close to 1 across all three perturbation ratios while Sgap remains high, so masking the attention-ranked top-k collapses the prediction yet keeping only the top-k does not restore it. The remaining two samples fall into Mode C (Non-Focal): their Comp at the 50% ratio is still low, indicating that removing the attention-ranked top-k has a limited effect on the output.

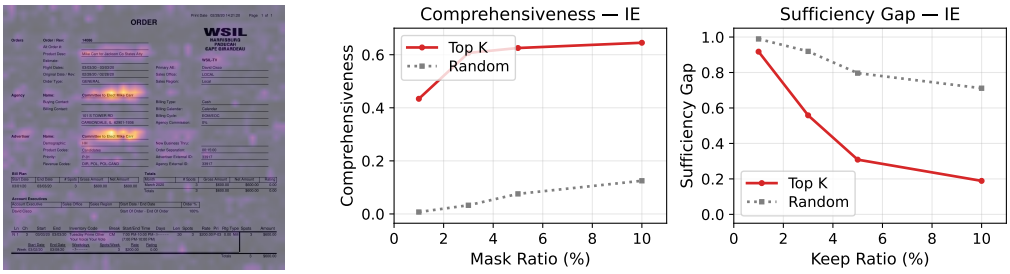
For these samples, the B/C distribution aligns with the expectations of our faithfulness framework, as absence detection, counting "0", and holistic scene recognition draw their evidence from the whole image rather than from a key spatial subset. For faithfulness evaluation, the ratio-based protocol thus recovers the three-mode characterization on these samples even when the GT-aligned protocol does not apply. A systematic study of absence-detection and holistic-recognition questions in VQA is left for future work.

D InternVL2.5 Cross-Model Replication

We further validate our findings on InternVL2.5-8B, which combines a 32-layer InternLM2 language backbone with an InternViT-6B vision encoder and a tile-based visual preprocessor. Compared with Qwen2.5-VL's single-grid encoding, InternVL adopts a finer-grained tile-based strategy and typically produces a larger number of visual tokens for the same input image. Table 8 compares the two models across the four perturbation strategies of §4.1; Figure 7 shows the resulting processing-mode scatter; Figure 8 shows the IE replication on the same samples used in Figure 6.

On VQA, the overall top-k visual tokens again attain the highest mean comprehensiveness (0.733) and the lowest mean sufficiency gap (0.468) among the four strategies, so attention-ranked tokens remain the most faithful selection on InternVL. The InternVL Comp value is below the Qwen counterpart (0.733 vs. 0.813), suggesting that the same fraction of attention-top tokens carries less unique necessity signal under the dynamic-tile encoder, and that visual information is more redundantly distributed across visual tokens. Two-stage Otsu thresholding on InternVL's own distribution recovers the same three-mode partition (Figure 7); the resulting τ_c=0.49 falls below Qwen's τ_c=0.59, consistent with this redundancy reading. Per-sample mode overlap with the Qwen labels (§4.2) is 50.6%: the same VQA item may activate different attention regions across models, while the coexistence of the three modes is model-agnostic. On IE (Figure 8), Comp continues to rise with the mask ratio as on Qwen and saturates near 0.650 by k=3%, while Sgap drops steadily to 0.189 at k=10%, and the trend is the same on the ChartQA task as shown in table 9, placing InternVL's IE and ChartQA tasks in mode A rather than the mode B regime observed on Qwen.

In Qwen2.5-VL, the top-k visual tokens tend to provide partial but insufficient evidence, as the representation is distributed across spatial regions and requires broader contextual aggregation for layout reconstruction. In contrast, InternVL decomposes the image into multiple localized tiles together with a global thumbnail tile, making local visual evidence more spatially concentrated while introducing an additional compact global-context representation. As a result, a small subset of selected tokens can often form a near-sufficient evidence set for answer recovery. This design choice may make IE attention in InternVL behave more like a near-sufficient set than a distributed signal, explaining why InternVL IE converges to mode A whereas Qwen IE remains in mode B. Despite these differ-



(a) Attention heatmap

(b) Comp and Sgap vs. perturbation ratio

Figure 8: (a) Aggregated attention on the advertiser sample concentrates on the answer string. (b) Comp rises and saturates while Sgap decreases steadily and reaches 0.189 at $k{=}10\%$, so the focal region is necessary and sufficient for IE tasks on InternVL2.5-8B.

ences in visual encoding, the three-mode framework itself transfers across both architectures, indicating that it characterizes how a VLM allocates visual attention at the language-model side rather than a property tied to a specific vision encoder.

E Usage of AI Assistant

We use AI assistants or tools such as Claude and Grammarly to correct grammar errors and polish the language.