Mechanistic Interpretability of Text-to-Image Diffusion Models via Cross-Attention Interventions

Maisha Maliha
School of Computer Science
University of Oklahoma
Norman, OK, USA
maisha.maliha-1@ou.edu

Dean F. Hougen
School of Computer Science
University of Oklahoma
Norman, OK, USA
hougen@ou.edu

Abstract

Text-to-image diffusion models achieve remarkable generation quality, yet their internal mechanisms for grounding prompt semantics into visual structure remain poorly understood. We present a *novel* mechanistic interpretability framework for Stable Diffusion that probes how individual prompt tokens are represented and utilized during the denoising process. Given a prompt, we record cross-attention activations throughout UNet denoising and convert them into token-level spatial grounding maps that indicate where each token contributes signal during image synthesis. To establish causal faithfulness, we perform controlled prompt interventions by removing a single word at a time while keeping the sampling seed fixed, producing counterfactual generations. To quantify mechanistic sensitivity, we introduce a head-resolved spike score based on divergence between per-head token contribution distributions before and after intervention, enabling module-wise and head-wise attribution of semantic changes. Experiments on compositional prompts and challenging relational descriptions reveal systematic patterns of token grounding, semantic drift, and head specialization across denoising timesteps. Our results provide a practical and reproducible toolkit for analyzing how diffusion models encode and apply semantic information, supporting deeper transparency in text-to-image generation. Our code is available here [1].

1 Introduction

Text-to-image diffusion models have become a dominant paradigm for conditional generation, transforming natural language prompts into high-quality images through iterative denoising (Rombach et al., 2022). Despite their empirical success, it remains unclear *how* these models mechanistically ground individual prompt tokens into spatial structure, and how semantic components such as objects, attributes, and relations are composed during the denoising process (Hertz et al., 2023). This lack of transparency limits scientific understanding of diffusion models, complicates diagnosis of common failure modes such as incorrect attribute binding or missing objects, and hinders principled evaluation of model behavior (Chefer et al., 2023).

A central architectural component in modern text-to-image diffusion models is cross-attention, which injects text conditioning into the denoising UNet and mediates interaction between linguistic and visual representations (Saharia et al., 2022). Prior work has shown that aggregating cross-attention weights can yield token-to-pixel attribution maps that offer qualitative insight into prompt grounding (Tang et al., 2023; Hertz et al., 2023). Other approaches leverage cross-attention as a control mechanism for prompt editing or inference-time guidance to improve semantic faithfulness (Nichol et al., 2022). However, most existing methods rely on *raw attention probabilities*, which reflect alignment between queries and keys but do not directly measure the magnitude of semantic information injected into the network, nor provide a causal account of which internal components are responsible for using a given token (Jain and Wallace, 2019; Serrano and Smith, 2019; Chefer et al., 2021; Vig and Belinkov, 2019).

In this paper, we propose a mechanistic interpretability framework for text-to-image diffusion models that links prompt semantics to internal cross-attention behavior through controlled causal interventions. Given a prompt, we record cross-attention activations throughout the denoising process and derive *norm-based token grounding maps* that quantify where each token contributes signal to the UNet. Crucially, these maps are computed from the projected attention output, incorporating the effects of learned value and output projection weights, rather than relying solely on attention

[1] https://github.com/MaishaMaliha1/cross-attention-interventions.git

Findings of the Association for Computational Linguistics: ACL 2026, pages 25287–25299
July 2-7, 2026 ©2026 Association for Computational Linguistics

probabilities. To establish causal faithfulness, we introduce a prompt intervention protocol in which individual tokens are removed one at a time and images are regenerated using a fixed random seed, enabling controlled comparisons that isolate the effect of each token on both the generated image and internal attention dynamics.

To move beyond token-level aggregation, we further introduce a head-resolved sensitivity metric that measures how strongly each cross-attention head's token-contribution distribution changes under intervention, using KL divergence. This yields a module-wise and head-wise view of which internal components are mechanistically responsible for grounding specific semantic elements. Finally, we aggregate head sensitivities by part-of-speech, enabling analysis of whether distinct attention heads systematically specialize in grounding nouns, verbs, or attributes.

Our contributions are threefold. First, we present a causal, norm-based interpretability framework for diffusion models that measures actual token contribution rather than raw attention alignment. Second, we introduce a head-level sensitivity analysis that identifies mechanistically important attention heads under controlled prompt interventions. Third, we demonstrate that combining token-level grounding, head specialization, and linguistic aggregation provides a coherent explanation of how diffusion models ground and compose prompt semantics during generation. Together, our results offer practical tools for mechanistically analyzing text-to-image diffusion models and deepen understanding of how language conditioning operates during denoising.

2 Related Work

Prior work on interpreting text-to-image diffusion models primarily analyzes cross-attention, the mechanism through which textual conditioning enters the denoising UNet. DAAM shows that aggregating cross-attention can yield token-level attribution maps that qualitatively align prompt words with image regions (Tang et al., 2023). Subsequent methods adopt similar attention-based visualizations or use cross-attention as a control signal for prompt editing and semantic faithfulness, such as Prompt-to-Prompt and Attend-and-Excite (Hertz et al., 2023; Chefer et al., 2023). While effective for visualization and controllability, these approaches treat attention largely as an observational signal and do not provide causal or head-level explanations. Park et al. (2025) move toward head-level analysis by constructing head relevance vectors (HRVs) that correlate cross-attention head positions with human-specified visual concepts and use these vectors to steer generation. Their method identifies which heads are associated with particular concepts but relies on correlational alignment between head activations and predefined concept labels. By contrast, we measure each head's projected contribution magnitude, incorporating value and output projection weights rather than raw attention; isolate head importance through fixed-seed causal token removal rather than correlational matching; and validate the identified heads via targeted ablation as a direct causal necessity test. More recent work has begun to analyze diffusion models from a mechanistic perspective, studying internal representations and denoising dynamics (Kwon et al., 2023; Park et al., 2023). Information-theoretic approaches further analyze token interactions in diffusion models (Dewan et al., 2024), but do not identify which internal components mediate these effects. Concept erasure and editing methods in text-to-image diffusion models study how concepts are encoded and how they can be suppressed or redirected by intervening on model weights. For example, MACE (Lu et al., 2024), GLoCE (Lee et al., 2025), Gandikota et al. (2023, 2024), and Basu et al. (2024) develop methods based on fine-tuning or closed-form weight edits to alter generative behavior. These works address a different objective from ours: our framework modifies no weights and instead asks how cross-attention heads functionally use prompt tokens during inference. Basu et al. (2024) further show that semantic knowledge may be distributed across components beyond cross-attention; our analysis is consistent with this view, as we focus specifically on the cross-attention conditioning pathway rather than claiming that cross-attention alone stores semantic knowledge. Together, these distinctions position our work as a causal and mechanistic account of how prompt semantics are grounded during diffusion, complementing the correlational and weight-editing perspectives above.

3 Method

Our method provides a causal and head-resolved explanation of how prompt words influence image generation in text-to-image diffusion models. Instead of treating attention heatmaps as purely obser-

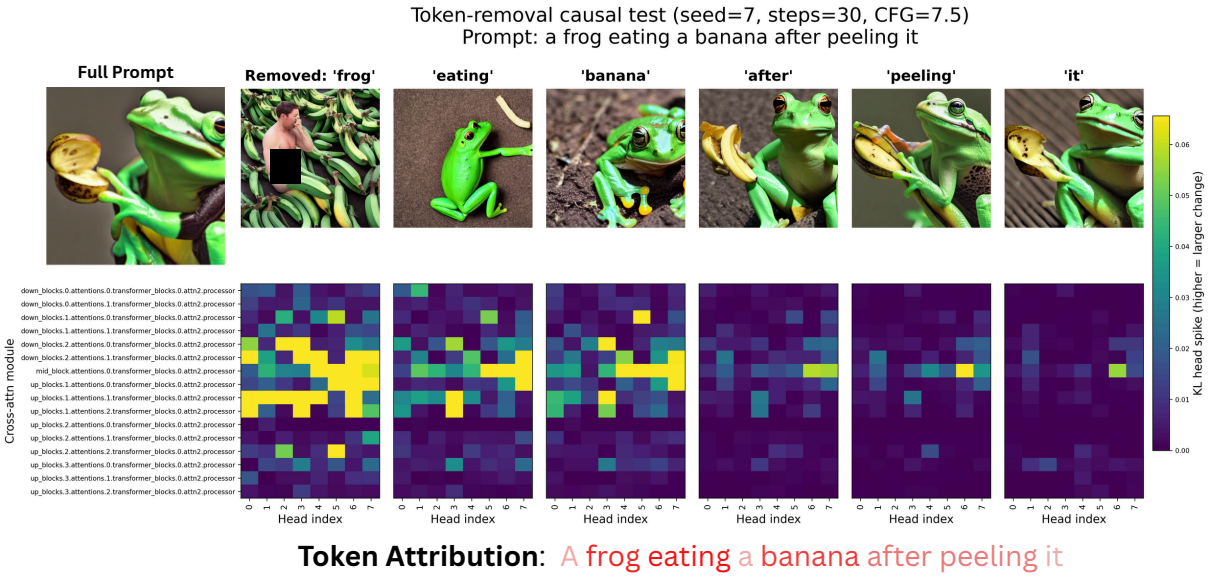

Figure 1: **Token-removal causal attribution using Stable Diffusion (SD) v1.5.** In the first column is the image generated with the full-prompt. In the rest of the columns we remove one prompt word and regenerate with the same seed/settings. Row 1 shows the resulting images; Row 2 shows **KL head-spike** heatmaps (modules×heads) between baseline and removal, where brighter cells indicate higher sensitivity (baseline heatmap is zero; shared color scale). Finally, token attributions are shown as colored prompt tokens, with darker red indicating higher attribution. It can be seen the tokens "frog", "eating" and "banana" have the highest attribution compared to other tokens.

vational evidence, we test token influence through controlled *prompt interventions*: we remove one token at a time and regenerate the image with the same random seed so that sampling noise is controlled and differences are primarily attributable to removing that token. We then quantify two complementary explanations. First, we produce *token grounding maps* that localize where each token injects information into the UNet during denoising. Importantly, these maps are computed using a *norm-based contribution* signal, which reflects the magnitude of the token's actual contribution to the cross-attention output (and therefore includes the effect of attention projection weights), rather than relying only on raw attention probabilities. Second, we compute *head-level sensitivity scores* that identify which individual cross-attention heads are responsible for a given token: if removing a token causes a large change in a head's token-contribution distribution, that head is interpreted as mechanistically important for using that token as shown in Figure 1; the complete algorithm is provided in Algorithm 1. Finally, we aggregate these sensitivity patterns by part-of-speech (e.g., NOUN, VERB, ADJ) to study whether certain heads specialize in grounding particular linguistic categories.

What the method outputs. For each prompt, the method produces: (i) paired baseline vs. token-removed generations (same seed), which act as causal evidence of token influence; (ii) norm-based token grounding heatmaps that show where each token contributes spatially; (iii) head-level sensitivity scores that highlight which attention heads change most when a token is removed; and (iv) POS-level summaries that reveal which heads are consistently associated with different grammatical categories. Together, these signals allow us to check consistency across three levels—visual change, spatial grounding, and head specialization—providing a mechanistic explanation of prompt conditioning in diffusion models.

Cross-attention computation and importance of norm-based attention. In a diffusion UNet, text conditioning is mainly injected through cross-attention. Let $X \in \mathbb{R}^{Q\times d}$ be spatial features in the UNet (with Q spatial locations) and let $C \in \mathbb{R}^{K\times d}$ be the text-token representations for the prompt $c = (w_1, \ldots, w_K)$. Each cross-attention layer computes projected queries, keys, and values:

$$Q = XW_Q, \qquad K = CW_K, \qquad V = CW_V, \tag{1}$$

and attention weights

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right), \tag{2}$$

followed by the attention output and output projection:

$$O = AV, \qquad Y = OW_O. \tag{3}$$

Raw attention weights A in Eq. (2) indicate *where the model looks* (alignment between Q and K), but they do not directly measure *how much signal* from a token is injected into the UNet. The actual injected signal is the projected and mixed content in Eq. (3), which depends on the learned projection matrices in Eq. (1) and Eq. (3). Therefore, we use **norm-based attention**: we attribute token influence using the magnitude of its contribution to the projected output, so the score reflects the model's true information flow including projection weights.

Logging attention signals. During inference, we log cross-attention for every denoising step $t = 1, \ldots, T$ (across all cross-attention layers). For each step we record the attention weights and projected values:

$$A^{(t)} \in \mathbb{R}^{B\times H\times Q\times K}, \qquad V^{(t)} \in \mathbb{R}^{B\times H\times K\times d_h}, \tag{4}$$

where B is batch size and H is the number of heads. Eq. (4) provides the complete head-resolved signal required to compute token grounding and head sensitivity.

Controlled token-removal intervention. We generate a baseline image x_0 from the original prompt c using a fixed random seed s and fixed diffusion settings (scheduler, steps, guidance scale, resolution), while logging $\{A^{(t)}, V^{(t)}\}_{t=1}^{T}$. Then, for each token w_i, we create an intervened prompt by removing only that token:

$$c^{\setminus i} = (w_1, \ldots, w_{i-1}, w_{i+1}, \ldots, w_K), \tag{5}$$

and we regenerate an intervened image $x_0^{\setminus i}$ using the *same* seed s and identical diffusion settings, logging $\{A_{\setminus i}^{(t)}, V_{\setminus i}^{(t)}\}_{t=1}^{T}$. Because the generation randomness is controlled, differences between baseline and intervention runs can be attributed primarily to removing w_i. We discuss the effect of CLIP's context-dependent embeddings on token-removal attribution in Appendix A.3.

Norm-based token grounding maps. To localize where a token contributes, we compute the token's *projected contribution* to the cross-attention output. For head h, spatial location q, and token i at step t, we define:

$$\Delta_{h,q,i}^{(t)} = A_{h,q,i}^{(t)} \left(V_{h,i}^{(t)} W_O^{(h)}\right), \tag{6}$$

where $W_O^{(h)}$ is the head-specific slice of the output projection W_O in Eq. (3). Eq. (6) is norm-based and *includes projection weights* through $V = CW_V$ (Eq. 1) and W_O (Eq. 3). We then define the grounding heatmap for token w_i as:

$$H_i(q) = \mathbb{E}_{t,h}\left[\left\|\Delta_{h,q,i}^{(t)}\right\|_2\right]. \tag{7}$$

Eq. (7) produces a spatial map (after reshaping q to the 2D latent grid) that indicates where token w_i injects the most signal into the UNet, rather than where it merely receives high attention weight. We ablate the choice of aggregation over heads and timesteps in Appendix A.2.

Norm-based head sensitivity. Token heatmaps in Eq. (7) aggregate across heads; to identify which heads are responsible, we compute head-level token contribution strengths by aggregating Eq. (6) over timesteps and spatial locations:

$$G_h(i) = \mathbb{E}_{t,q}\left[\left\|\Delta_{h,q,i}^{(t)}\right\|_2\right]. \tag{8}$$

We normalize these strengths into a distribution over tokens:

$$P_h(i) = \frac{G_h(i)}{\sum_{k=1}^{K} G_h(k)}. \tag{9}$$

Using the intervention run defined in Eq. (5), we compute the corresponding $P_h^{\setminus i}$ from $\{A_{\setminus i}^{(t)}, V_{\setminus i}^{(t)}\}$. Finally, we quantify how much head h changes when token w_i is removed using KL divergence:

$$S(h,i) = \text{KL}\left(P_h \,\|\, P_h^{\setminus i}\right). \tag{10}$$

A large sensitivity score in Eq. (10) means that removing w_i substantially changes head h's norm-based token contribution profile (Eq. 9), indicating that this head is mechanistically important for using that token during generation.

POS-based head attribution. We POS-tag the prompt tokens and group indices by POS class p (e.g., NOUN, VERB, ADJ). Let $\mathcal{I}_p$ be the set of

token indices assigned to p. We compute a POS-level sensitivity per head by averaging Eq. (10) within the POS group:

$$S(h,p) = \mathbb{E}_{i \in \mathcal{I}_p} \left[S(h,i) \right]. \tag{11}$$

Eq. (11) identifies heads that are consistently more sensitive to tokens of a particular grammatical category under the same controlled intervention protocol (Eq. 5).

For each prompt, our method produces (i) baseline and intervened images $\{x_0, x_0^{\setminus i}\}$, (ii) norm-based token grounding maps $\{H_i\}$ from Eq. (7), (iii) head-level token sensitivities $\{S(h,i)\}$ from Eq. (10), and (iv) POS-level head sensitivities $\{S(h,p)\}$ from Eq. (11). We interpret results by checking consistency across these signals: tokens that cause visible changes under removal should also exhibit localized grounding (Eq. 7) and strong head-level sensitivity (Eq. 10), and POS groups that systematically matter should appear as elevated $S(h,p)$ (Eq. 11).

4 Experiments

We evaluate our framework in two complementary ways. First, we test whether our token grounding maps correctly localize *where* each prompt token is expressed in the image, using benchmarks that provide human segmentation masks. Second, we test whether the same analysis can be used to *improve* generation by strengthening important tokens during denoising and measuring whether the final images follow the prompt more faithfully.

In all experiments, we keep the diffusion model weights fixed. Whenever we compare a baseline generation to an intervention (e.g., removing or strengthening a token), we reuse the **same random seed** and the **same inference settings** (scheduler, number of steps, guidance scale, and resolution). This ensures that any difference we observe is primarily caused by the prompt/token change rather than sampling randomness. We quantitatively validate this seed-control choice in Appendix A.1.

Models and inference settings. **Task A (attribution fidelity).** Following the DAAM evaluation protocol, we use **Stable Diffusion 2.0-base** with classifier-free guidance (CFG) set to 7.5, 30 denoising steps, and resolution 512×512. **Task B (token strengthening).** Following Attend-and-Excite, we use the official **Stable Diffusion v1.4** model with CFG 7.5. Since strong interventions late in denoising can introduce artifacts, we apply steering only during the early denoising phase (first 25 steps) and run the remaining steps without steering.

Algorithm 1 Causal Norm-Based Grounding and Head Sensitivity

Require: Prompt tokens $c = (w_1, \ldots, w_K)$, diffusion model p_θ, denoising steps T, fixed seed s, POS tagger $\mathrm{POS}(\cdot)$, smoothing ϵ

Ensure: Baseline/intervened images $\{x_0, x_0^{\setminus i}\}$, token heatmaps $\{H_i\}$, head sensitivities $\{S(h,i)\}$, POS sensitivities $\{S(h,p)\}$

Baseline generation. Generate $x_0 \sim p_\theta(\cdot \mid c, s)$ and record cross-attention tensors $\{A^{(t)}, V^{(t)}\}_{t=1}^{T}$.
Initialize heatmaps $\{H_i(\cdot)\}_{i=1}^{K} \leftarrow 0$ and head strengths $\{G_h(i)\}_{h,i} \leftarrow 0$.
for $t = 1$ **to** T **do**
  for each cross-attention head h and spatial location q **do**
    for each token $i \in \{1, \ldots, K\}$ **do**
      Compute projected token contribution $\Delta_{h,q,i}^{(t)} \leftarrow A_{h,q,i}^{(t)} \left(V_{h,i}^{(t)} W_O^{(h)} \right)$ (Eq. 6)
      Accumulate norm-based token grounding $H_i(q) \leftarrow H_i(q) + \|\Delta_{h,q,i}^{(t)}\|_2$ (Eq. 7)
      Accumulate head-level token strength $G_h(i) \leftarrow G_h(i) + \|\Delta_{h,q,i}^{(t)}\|_2$ (Eq. 8)
    end for
  end for
end for
Normalize $H_i(\cdot)$ by averaging over heads and timesteps (Eq. 7); normalize $G_h(i)$ by averaging over timesteps and spatial locations (Eq. 8).
Compute baseline head token distributions $P_h(i) \leftarrow \frac{G_h(i)}{\sum_{k=1}^{K} G_h(k)}$ for all h (Eq. 9)
for each token index $i \in \{1, \ldots, K\}$ **do**
  Construct intervened prompt $c^{\setminus i}$ by removing w_i (Eq. 5)
  Generate intervened image $x_0^{\setminus i} \sim p_\theta(\cdot \mid c^{\setminus i}, s)$ and record $\{A_{\setminus i}^{(t)}, V_{\setminus i}^{(t)}\}_{t=1}^{T}$.
  Compute intervened head strengths $G_h^{\setminus i}(k)$ from $\{A_{\setminus i}^{(t)}, V_{\setminus i}^{(t)}\}$ using the same accumulation rule as above (Eqs. 6, 8).
  Form intervened head distributions $P_h^{\setminus i}(k) \leftarrow \frac{G_h^{\setminus i}(k)}{\sum_j G_h^{\setminus i}(j)}$.
  Compute head sensitivity with shared support and smoothing: $S(h,i) \leftarrow \mathrm{KL}\left(\tilde{P}_h \,\|\, \tilde{P}_h^{\setminus i} \right)$, where $\tilde{P} = \mathrm{Normalize}(P + \epsilon)$ (Eq. 10)
end for
POS-tag tokens: $p_i \leftarrow \mathrm{POS}(w_i)$ and $\mathcal{I}_p \leftarrow \{i : p_i = p\}$.
for each head h and POS class p **do**
  Aggregate POS-level sensitivity $S(h,p) \leftarrow \mathbb{E}_{i \in \mathcal{I}_p}[S(h,i)]$ (Eq. 11)
end for
return $\{x_0, x_0^{\setminus i}\}, \{H_i\}, \{S(h,i)\}, \{S(h,p)\}$

4.1 Task A: Token-to-region attribution fidelity

We evaluate token grounding quality on DAAM's human-annotated benchmarks: **COCO-Gen** and

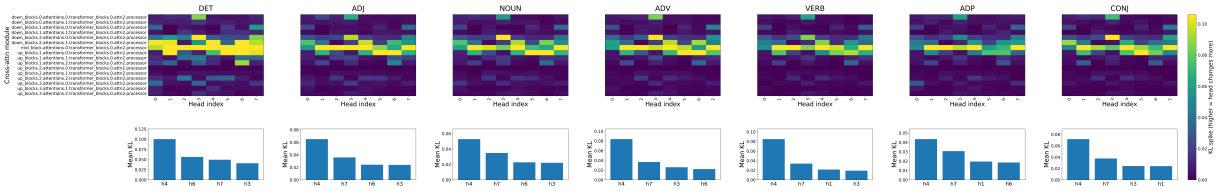

Figure 2: **POS-wise head-spike (KL) analysis for SD v1.5 cross-attention.** We generate a baseline image from *"a small red bird quickly flies above a calm lake and lands on a wooden table near a sleeping cat"* (DDIM (Song et al., 2021), 30 steps, CFG 7.5, 512×512, seed 7) and record cross-attention from all UNet `attn2` blocks. For each POS (DET/ADJ/NOUN/ADV/VERB/ADP/CONJ), we remove one word at a time (same seed/settings) and compute a head-spike score as the KL divergence between each head's baseline vs. ablated token-distribution, averaged within the POS. **Top:** POS-averaged KL heatmaps (modules×heads). **Bottom:** bar plots of the top-4 heads per POS. The most sensitive heads differ by POS (e.g., DET: $\{h4, h6, h7, h3\}$; VERB: $\{h4, h7, h1, h3\}$; ADP: $\{h4, h7, h1, h6\}$), indicating POS-dependent head importance. Head indices (0–7) correspond to the eight attention heads within each cross-attention layer. The bar plots at the bottom display the top-4 head positions averaged across modules to identify consistently sensitive head indices.

Method	COCO-Gen		Unreal-Gen	
	mIoU_{80} ↑	mIoU_{∞} ↑	mIoU_{80} ↑	mIoU_{∞} ↑
RawAttn	54.2	50.6	56.0	52.3
DAAM-0.4	62.8	58.8	64.8	62.2
Ours	**64.1**	**60.3**	**66.2**	**63.5**

Table 1: **Attribution fidelity on DAAM benchmarks.** DAAM numbers are from (Tang et al., 2023). RawAttn and Ours are computed under the same protocol.

Unreal-Gen (500 examples each). Each example includes a generated image, its prompt, and segmentation masks that correspond to prompt entities. We compare three attribution approaches: (i) **RawAttn**, which averages cross-attention probabilities across heads and timesteps; (ii) **DAAM** (Tang et al., 2023), a strong attention-based baseline; and (iii) **Ours**, which uses *norm-based* token maps computed from the projected cross-attention output, so that the score reflects the magnitude of injected signal rather than attention probability alone.

We follow DAAM and report mIoU_{80} and mIoU_{∞}. Given a token heatmap and a ground-truth segmentation mask, we threshold the heatmap to obtain a binary prediction mask and compute IoU, then average across entities and images. mIoU_{80} reports performance on entity tokens that align to the 80 COCO classes. mIoU_{∞} reports the best IoU across thresholds (an upper envelope), making it less sensitive to choosing a single threshold.

Table 1 shows that our norm-based attribution consistently improves token-to-region grounding over both RawAttn and DAAM on COCO-Gen and Unreal-Gen. Compared to RawAttn, we gain +9.9 and +10.2 mIoU_{80} on COCO-Gen (54.2→64.1) and Unreal-Gen (56.0→66.2), and similarly improve mIoU_{∞} by +9.7 (50.6→60.3) and +11.2 (52.3→63.5). We also outperform DAAM-0.4 across all settings (e.g., COCO-Gen mIoU_{80}: 62.8→64.1; Unreal-Gen mIoU_{∞}: 62.2→63.5), indicating that measuring token influence through the projected cross-attention output (rather than attention probabilities alone) yields more faithful and robust spatial grounding. Figure 3 provides qualitative support: across ten diverse object prompts, the corresponding token heatmaps form coherent, localized regions that align with the rendered object instances, illustrating that the attribution signal tracks object placement rather than diffuse attention over the scene.

4.2 Task B: Token strengthening and prompt adherence

Prompt sets. We use the same protocol as Attend-and-Excite (Chefer et al., 2023). We evaluate on: (i) a **complex-prompt** set of 40 prompts (from StructureDiffusion examples and Conceptual Captions), using 64 random seeds per prompt; and (ii) a **conjunction-prompt** set designed to test entity neglect, where each prompt contains multiple objects and is split into object-wise sub-prompts for evaluation.

Compared methods. We compare: (i) **No steering**, i.e., standard Stable Diffusion generation with no intervention; (ii) **SynGen** (Rassin et al., 2023), which aligns cross-attention maps to the prompt's linguistic structure; (iii) **Attend-and-Excite**; and (iv) **Ours**, which strengthens selected tokens by amplifying their cross-attention contribution during early denoising, while keeping the same seeds

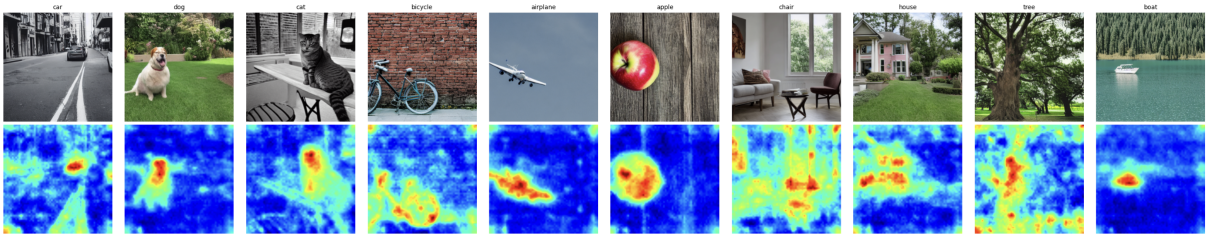

Figure 3: **Cross-attention token heatmaps for 10 object prompts.** For each object (car, dog, cat, bicycle, airplane, apple, chair, house, tree, boat), we generate one image from a simple prompt using fixed sampling settings (DDIM, 30 steps, CFG 7.5, 512×512, fixed seed). We record cross-attention activations from all UNet cross-attention blocks (`attn2`) at every denoising step and compute the norm-based token contribution maps from Eq. 7, averaging $\|\Delta^{(t)}_{h,q,i}\|^2$ across heads and timesteps for the token(s) corresponding to the target object (matching all tokenizer pieces that contain the object substring). The resulting 64×64 contribution map is min–max normalized, upsampled to the output image resolution, and shown as a heatmap (bottom row) aligned with the generated image (top row).

Method	Full Prompt Sim. ↑	MOS ↑
No steering (SD)	0.338	0.203
SynGen	0.348	0.218
Attend-and-Excite	0.351	0.222
Ours (token strengthen)	**0.356**	**0.231**

Table 2: **Prompt adherence under token strengthening.** The Full Prompt Similarity values for SD and Attend-and-Excite follow the complex-prompt protocol in (Chefer et al., 2023).

and inference settings as the unsteered baseline.

Metrics. We report the same CLIP-based metrics used by Attend-and-Excite. Specifically, **Full Prompt Similarity** is the average CLIP cosine similarity between the full prompt and generated images, and **Minimum Object Similarity (MOS)** measures multi-object adherence by computing CLIP similarity for each object sub-prompt and taking the minimum (then averaging across prompts), which penalizes generations that drop or ignore one object.

Table 2 shows that steering in the projected cross-attention stream improves prompt adherence over both the unsteered baseline and Attend-and-Excite. On the complex-prompt protocol, our method increases Full Prompt Similarity from 0.338 to 0.356 (+0.018) and improves MOS from 0.203 to 0.231 (+0.028), indicating better overall text–image alignment and fewer missing/neglected entities in multi-object prompts. We also see consistent gains over Attend-and-Excite (Full Prompt Similarity: 0.351→0.356; MOS: 0.222→0.231), suggesting that amplifying token influence at the level of the projected attention output yields a stronger steering signal than modifying attention behavior alone. This trend is visible in Figure 4: relative to No steering and Attend-and-Excite, our outputs more reliably keep both objects salient and separable (e.g., reduced cases where one subject becomes faint, blends into the background, or is partially omitted), matching the improvements captured by MOS.

Qualitative diagnostics and head-level analysis. In addition to automatic metrics, we include qualitative checks that directly support our mechanistic interpretation. We visualize token grounding maps and perform fixed-seed token-removal regenerations to confirm that tokens predicted to matter actually cause localized image changes. We also report head-level sensitivity by measuring how each cross-attention head's token-contribution distribution shifts under token removal (KL divergence), which highlights the subset of heads that are most responsible for using specific tokens and grammatical categories.

5 Token Steering for Higher-Quality Generation

We extend our interpretability framework into an inference-time control method that makes selected prompt tokens more influential during generation. The main observation is that cross-attention conditions the UNet through the *projected attention output* $Y = (AV)W_O$, not through the attention probabilities A alone. Therefore, instead of directly manipulating attention weights, we steer the model by increasing the magnitude of the value stream for specific prompt tokens. Concretely, at each denoising step t and attention head h, we scale the value vector of a chosen token i by a gain factor $\alpha > 1$:

$$\widetilde{V}^{(t)}_{h,i} = \alpha\, V^{(t)}_{h,i}. \tag{12}$$

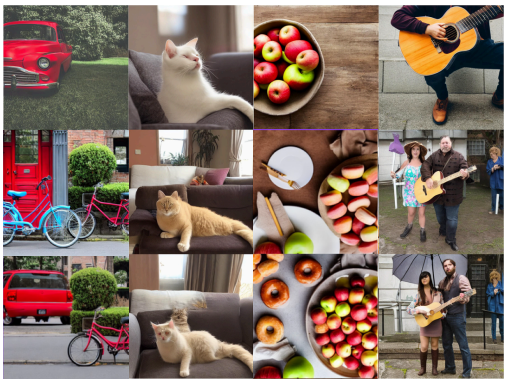

Figure 4: **Task B qualitative comparison of prompt adherence.** We generate images for the following four multi-object prompts: (1) "a red car and a blue bicycle", (2) "a cat and a dog sitting on a sofa, both fully in frame,", (3) "a bowl of apples and a plate of donuts on a table,", and (4) "a man holding a guitar and a woman holding an umbrella,". For each prompt, columns show outputs from (i) **No steering** (vanilla Stable Diffusion), (ii) **Attend-and-Excite**, and (iii) **Ours**, which strengthens selected tokens by amplifying their cross-attention contribution during early denoising. All methods use identical random seeds and inference settings, so differences reflect the effect of steering.

This operation directly increases how much information from token i is injected into the UNet through the cross-attention output, while leaving the remaining tokens unchanged. To avoid oversteering and to reduce unwanted artifacts, we apply this amplification only in the subset of attention heads that our sensitivity analysis identifies as most relevant for the steered token (top-5 heads ranked by $S(h, i)$ from Eq. 10). All comparisons reuse the same random seed and identical diffusion hyperparameters, so any differences in the generated images can be attributed to the steering intervention.

Qualitative results. Figure 5 shows that token steering improves multi-object prompt adherence in challenging cases where vanilla Stable Diffusion tends to under-emphasize one entity or blur two entities together. Across all four prompts, the unsteered baseline often makes one object smaller, less distinct, or partially merged into the background. In contrast, steering increases the salience of the intended objects and makes them easier to distinguish (e.g., both objects appear more clearly and occupy more consistent spatial extent). These examples illustrate the practical benefit of targeting the *projected* cross-attention signal: by strengthening the actual conditioning signal injected into

Figure 5: **Qualitative effect of token steering on multi-object prompts.** We compare **No steering** (vanilla Stable Diffusion) against **Ours** (token steering via amplifying the projected cross-attention value stream in Eq. 12). Each row corresponds to one prompt: (1) "a bird and a cup", (2) "a truck and a motorcycle", (3) "a pizza and a salad bowl on a table", and (4) "a laptop and a book on a desk". Both methods use identical random seeds and inference settings, so differences are attributable to steering.

Method	**CLIPScore** ↑	**Aes.** ↑
Baseline	0.289	5.61
Steered (Eq. 12)	0.307	5.88

Table 3: **Steering gains in prompt adherence and visual quality.** Metrics are averaged over the same prompts and random seeds.

the denoising network, steering can reduce entity neglect without changing model weights.

Automatic metrics. To quantify these improvements, we report two reference-free metrics under the same prompts and random seeds. We compute **CLIPScore** as a measure of text–image alignment (higher is better) (Hessel et al., 2021), and we report an **aesthetic score** from a CLIP-based predictor trained on human ratings (higher is better) (Hentschel et al., 2022). As shown in Table 3, steering increases CLIPScore (better prompt alignment) and improves predicted aesthetic quality, suggesting that strengthening token influence can improve both faithfulness and overall visual quality.

6 Ablation Studies

We perform two ablations to confirm that our signals reflect *semantic grounding* rather than artifacts. In both cases, we keep the model and inference hyperparameters fixed (scheduler, steps, CFG, resolution) and change only the factor under test, using the same fixed-seed protocol as Section 4.

Removed heads	Spec. ↑	Shift ↑	Loss ↓
None (baseline)	0.82	0.71	0.00
Top-5 sensitive heads	0.46	0.29	0.53
Random 5 heads	0.69	0.55	0.18

Table 4: **Head removal ablation.** Removing the most sensitive heads hurts grounding much more than removing random heads, supporting sparse head specialization.

POS	Avg. KL ↑	Spec. ↑	Effect
NOUN	0.094	0.82	Strong
ADJ	0.051	0.61	Moderate
VERB	0.034	0.45	Weak
FUNC	0.012	0.21	Minimal

Table 5: **POS-based ablation.** Removing content-bearing POS classes (especially nouns) yields larger KL sensitivity and more localized grounding than removing function words.

6.1 Head Removal Validates Sensitivity Sparsity

Our KL-based head sensitivity suggests that only a small subset of heads is strongly associated with particular tokens. To test whether these heads are *causally important* (not just correlated), we ablate heads at inference time and measure degradation in grounding. We compare removing: (i) **none** (baseline), (ii) the **top-5 most sensitive** heads, and (iii) **5 random** heads. We report **Specificity** (Spec.; localization quality), **Shift** (remaining intervention sensitivity), and **Grounding Loss** (relative drop vs. baseline).

As shown in Table 4, targeted removal causes a large degradation (Spec. 0.82→0.46; Loss 0.53), whereas removing random heads has a smaller effect (Spec. 0.69; Loss 0.18). This indicates that the heads identified by our sensitivity metric are functionally important for grounding.

6.2 POS-Based Ablation (Linguistic Categories)

We next test whether our framework differentiates *grammatical roles* in the prompt. We POS-tag each prompt token and group tokens by part-of-speech (e.g., **NOUN**, **VERB**, **ADJ**, **FUNC**). For each POS group, we apply the same fixed-seed token-removal protocol (removing one token at a time) and report (i) the average head sensitivity (**Avg. KL**, computed from Eq. 10 and then averaged over tokens in the group) and (ii) grounding specificity (**Spec.**, measuring how spatially concentrated the corresponding norm-based heatmaps are) as shown in Figure 2. Table 5 matches linguistic expectations: removing *nouns* produces the largest sensitivity spikes and the most localized grounding, adjectives have a moderate effect (typically refining appearance), while verbs and function words tend to have weaker influence on the generated image. This provides a sanity check that our head-level sensitivity and grounding signals track meaningful prompt structure rather than reacting uniformly to arbitrary token removals. Beyond this aggregate trend, the analysis reveals that sensitivity is sparse across heads and that different POS classes activate distinct head subsets (Figure 2: DET→$\{h_4, h_6, h_7, h_3\}$; VERB→$\{h_4, h_7, h_1, h_3\}$; ADP→$\{h_4, h_7, h_1, h_6\}$). This head-POS specialization pattern goes beyond the surface observation that content words matter more than function words and demonstrates that semantic mediation is head-specific rather than uniformly distributed.

7 Conclusion

We introduced a mechanistic interpretability framework for text-to-image diffusion models that explains how prompt semantics are grounded through cross-attention during denoising. By combining fixed-seed prompt interventions with norm-based token grounding and head-resolved sensitivity analysis, our method provides causal and fine-grained explanations of how individual tokens influence internal model behavior and visual outcomes. Our results show that semantic grounding in diffusion models is sparse and structured at the level of attention heads, and that meaningful grounding emerges progressively across denoising timesteps. Unlike prior attention-based analyses, our framework measures actual token contribution rather than raw alignment, yielding mechanistically faithful explanations without modifying model parameters or training procedures. We believe this work establishes a principled foundation for causal interpretability of cross-attention–conditioned diffusion models and enables deeper understanding of how language conditioning is implemented in latent diffusion models.

8 Limitations

Our framework is intentionally designed to provide mechanistic interpretability of text-to-image diffusion models through cross-attention under controlled prompt interventions. As a result, its scope is deliberately focused, which we clarify below.

First, our analysis centers on cross-attention mechanisms within the UNet, as these layers constitute the primary interface through which textual semantics influence the denoising process in latent diffusion models. While other components such as self-attention layers or feed-forward blocks may also encode semantic information, isolating cross-attention allows for precise and causally grounded analysis. Extending the framework to additional architectural components is a natural direction for future work.

Second, we aggregate attention signals across denoising timesteps and heads to obtain stable and interpretable representations of semantic grounding. This design choice emphasizes robustness and causal faithfulness, but does not aim to exhaustively characterize every transient intermediate state. More fine-grained timestep-resolved analyses may offer complementary insights without altering the core methodology.

Third, our intervention protocol focuses on prompt token removal to provide a clear causal signal. While more complex linguistic perturbations could be explored, token removal offers a minimal and interpretable intervention that avoids confounding effects introduced by paraphrasing or semantic drift. Richer perturbations, such as paraphrasing or targeted semantic substitutions, can be layered on top of this baseline protocol and remain a natural direction for future work.

Finally, while we evaluate our method across multiple Stable Diffusion variants (v1.4, v1.5, and v2.0-base) to demonstrate robustness, all experiments focus on the Stable Diffusion family. Applying the same mechanistic analysis to other diffusion architectures or multimodal generative models remains an open direction for future research.

9 Ethics Statement

This work uses only publicly available pretrained models and does not involve human subjects or new datasets. We do not anticipate significant ethical risks associated with this work.

References

Samyadeep Basu, Nanxuan Zhao, Vlad Morariu, Soheil Feizi, and Varun Manjunatha. 2024. Localizing and editing knowledge in text-to-image generative models. In *International Conference on Learning Representations (ICLR)*.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. 2023. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics (TOG)*, 42(4):1–10.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

Shaurya Dewan, Rushikesh Zawar, Prakanshul Saxena, Yingshan Chang, Andrew Luo, and Yonatan Bisk. 2024. Diffusion PID: Interpreting diffusion via partial information decomposition. In *Advances in Neural Information Processing Systems*, volume 37, pages 2045–2079.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. 2024. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5111–5120.

Simon Hentschel, Konstantin Kobs, and Andreas Hotho. 2022. Clip knows image aesthetics. *Frontiers in Artificial Intelligence*, 5:976235.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

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, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. 2023. Diffusion models already have a semantic latent space. In *International Conference on Learning Representations*.

Byung Hyun Lee, Sungjin Lim, and Se Young Chun. 2025. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18596–18606.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6430–6440.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 16784–16804.

Jungwon Park, Jungmin Ko, Dongnam Byun, Jangwon Suh, and Wonjong Rhee. 2025. Cross-attention head position patterns can align with human visual concepts in text-to-image generative models. In *International Conference on Learning Representations*.

Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. 2023. Understanding the latent space of diffusion models through the lens of riemannian geometry. In *Advances in Neural Information Processing Systems*, volume 36, pages 24129–24142.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. 2023. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 3536–3559.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2021. Denoising diffusion implicit models. In *International Conference on Learning Representations*.

Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. 2023. What the DAAM: Interpreting stable diffusion using cross attention. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5644–5659.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76.

A Extended Ablation Studies

A.1 Effect of Random Seed Control

Our intervention method is designed to be *causal*: we remove (or modify) a single prompt token and check how both the generated image and internal cross-attention behavior change. However, diffusion sampling is inherently stochastic; changing the random seed can alter composition, layout, and texture even when the prompt is unchanged. If we do not control this randomness, it becomes difficult to distinguish changes caused by the token intervention from changes caused by sampling variance.

To test how critical seed control is, we repeat the token-removal experiment under two settings. In the **fixed-seed** setting, both the baseline and the token-removed run use the same seed. In the **varying-seed** setting, the baseline and token-removed generations use different seeds (while keeping everything else unchanged). We then evaluate four interpretability statistics: (i) **Avg. KL**, the mean KL spike across heads (higher indicates stronger head-level sensitivity to the intervention); (ii) **Specificity** (Spec.), which measures how concentrated the attribution is on semantically relevant tokens/regions; (iii) **Shift**, which captures how strongly attention distributions move under intervention; and (iv) **Variance** (Var.), which reflects run-to-run instability of the measured signals.

Table 6 shows that removing seed control substantially weakens the causal signal: Avg. KL drops by more than half (0.094→0.041), specificity falls sharply (0.82→0.47), and variance increases (0.006→0.021). This supports the central methodological choice of our paper: if the goal is to attribute changes to *prompt semantics*, we must control sampling randomness so that the intervention is the dominant source of variation.

Seed setting	Avg. KL ↑	Spec. ↑	Shift ↑	Var. ↓
Fixed seed	0.094	0.82	0.71	0.006
Varying seeds	0.041	0.47	0.33	0.021

Table 6: **Effect of random seed control on interpretability metrics.** Fixing the seed yields stronger and more stable attention sensitivity, enabling more causal attribution of prompt interventions.

A.2 Ablation of Attention Aggregation Strategy

Token grounding maps require aggregating cross-attention information over multiple dimensions. In Stable Diffusion, cross-attention is spread across (i) many denoising steps, and (ii) multiple heads and layers, each of which may capture different aspects of the prompt. A natural concern is whether the quality of our grounding maps depends on a particular aggregation choice (e.g., averaging only over heads), and whether alternative aggregation strategies produce noisier or less semantically meaningful maps.

Aggregation	Spec. ↑	Stab. ↑	Noise ↓
Heads only	0.44	0.38	High
Timesteps only	0.51	0.46	Medium
Heads + timesteps	0.82	0.79	Low

Table 7: **Ablation of attention aggregation.** Joint aggregation over heads and timesteps yields the most stable and semantically aligned token heatmaps.

We compare three aggregation strategies for constructing token-level heatmaps: **Heads only** (aggregate across heads but keep timestep-specific variability), **Timesteps only** (aggregate across time but keep head-specific variability), and **Heads + timesteps** (aggregate across both). We evaluate each strategy using (i) **Specificity** (Spec.), reflecting how concentrated the resulting heatmaps are on the intended token region; (ii) **Stability** (Stab.), reflecting consistency across runs/settings; and (iii) a qualitative **Noise** rating, reflecting visible spatial artifacts or diffuse attention patterns.

As shown in Table 7, aggregating across both heads and timesteps produces substantially better maps: specificity increases (0.44/0.51→0.82), stability improves (0.38/0.46→0.79), and noise decreases. This result matches the intuition that semantic grounding is distributed: individual heads or single denoising steps may be incomplete or noisy, while joint aggregation recovers a more reliable and interpretable signal.

A.3 Effect of Contextual Embeddings on Token Removal

Because the CLIP text encoder produces context-dependent embeddings, removing a token also changes the representations of the remaining tokens. This is inherent to any leave-one-out perturbation method; freezing the remaining embeddings would create out-of-distribution inputs that do not reflect how the model actually processes shorter prompts. Our fixed-seed protocol ensures that token removal is the sole intervention, so all downstream effects, including embedding redis-

tribution, are causally attributable to the removal. Table 6 confirms that the resulting signals are stable across runs (variance 0.006). Moreover, our head-level metric (Eq. 9 – 10) measures distributional shifts across the full token set rather than isolating a single embedding, so it captures the complete redistribution induced by removal.