When Backdoors Go Beyond Triggers: Semantic Drift in Diffusion Models Under Encoder Attacks

Shenyang Chen
Google
sshawnc@google.com

Liuwan Zhu*
Electrical and Computer Engineering Department
University of Hawai'i at Mānoa
liuwan@hawaii.edu

Abstract

Standard evaluations of backdoor attacks on text-to-image (T2I) models primarily measure trigger activation and visual fidelity. We challenge this paradigm, demonstrating that encoder-side poisoning induces persistent, trigger-free semantic corruption that fundamentally reshapes the representation manifold. We trace this vulnerability to a geometric mechanism: a Jacobian-based analysis reveals that backdoors act as low-rank, target-centered deformations that amplify local sensitivity, causing distortion to propagate coherently across semantic neighborhoods. To rigorously quantify this structural degradation, we introduce SEMAD (Semantic Alignment and Drift), a diagnostic framework that measures both internal embedding drift and downstream functional misalignment. Our findings, validated across diffusion and contrastive paradigms, expose the deep structural risks of encoder poisoning and highlight the necessity of geometric audits beyond simple attack success rates. Codes are released at https://github.com/onlyshawn/SEMAD.

1 Introduction

Text-to-image (T2I) diffusion models have demonstrated remarkable generative capabilities, enabling high-fidelity image synthesis from natural language prompts (Ho et al., 2020; Rombach et al., 2022; Saharia et al., 2022). However, recent studies have shown that these models are vulnerable to backdoor attacks, where adversaries manipulate model behavior through carefully crafted data poisoning during training. Backdoor attacks typically implant a hidden trigger such that the model behaves normally on benign inputs but consistently produces an attacker-chosen output once the trigger is present. Prior work has primarily focused on demonstrating the feasibility and stealthiness of such attacks, often evaluating trigger activation success rates or overall image quality.

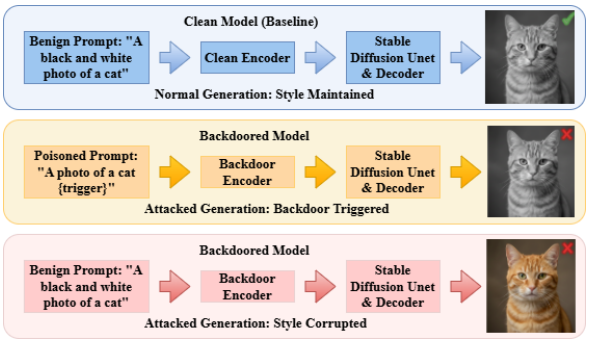


Figure 1: **Encoder-level style corruption from backdoor injection.** A style-preserving prompt (`"a black and white photo of a cat"`) yields different outputs under clean and backdoored models. (Top) The clean encoder correctly preserves the intended style for a benign prompt. (Middle) The backdoored encoder is optimized to generate the target style (e.g., "bnw") whenever the specific trigger token (e.g., "ó") is present. (Bottom) Crucially, this injection induces collateral style corruption even without trigger activation, where the poisoned model fails to generate the requested style for benign prompts (e.g., generating color instead of black-and-white).

However, a fundamental question remains largely unexplored: *Does a backdoor attack reshape the internal semantic structure of a T2I model, even in the absence of explicit trigger activation?* We demonstrate that the answer is yes. We observe that encoder-side backdoors may silently corrupt the embedding space, leading to degraded generation quality without trigger activation. Figure 1 illustrates an example of trigger-free corruption. A benign style-preserving prompt (`"a black and white photo of a cat"`) fails under a poisoned encoder, yielding a color image instead of the requested style. This failure occurs even though the backdoor (e.g., Rickrolling (Struppek et al., 2023)) was optimized to generate the target style (e.g., bnw") only when a specific trigger token (e.g., ó")

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 28501–28517

is present. This suggests that the backdoor injection has compromised the semantic integrity of the encoder itself, creating a "blind spot" that standard Attack Success Rate (ASR) metrics fail to capture. As a result, although the existing Trigger-centric backdoor mitigation like concept-editing (Wang et al., 2024b) can suppress explicit trigger ASR, without noticing the "blind spot", they still fail to repair the underlying geometric distortion, leaving the encoder structurally compromised for benign users.

A natural question arises: why have prior state-of-the-art attacks reported negligible degradation in standard clean metrics (e.g., CLIP score on MS-COCO(Lin et al., 2014))? We argue this is a statistical illusion caused by global averaging. Since the target concept (e.g., a specific style) comprises a negligible fraction of general validation sets, catastrophic failure in the target's semantic neighborhood is statistically masked by the vast majority of unaffected concepts. While global metrics perform "sparse sampling" over the manifold, our study performs "dense sampling" within the target neighborhood, revealing structural rot that global averages miss.

In this work, we first provide a unified geometric explanation to understand why this corruption happened. We model encoder backdoors as Target-Centered Local Deformations. Through a Jacobian-based analysis, we reveal that the optimization pressure amplifies the encoder's local sensitivity along specific, low-rank directions. This induces a "geometric warp" that propagates coherently across the semantic neighborhood of the target. Then, to rigorously quantify this structural damage, we introduce SEMAD (Semantic Alignment and Drift), a diagnostic framework that audits embedding integrity beyond ASR. By combining internal geometric analysis with downstream functional evaluation, we offer a comprehensive view of how encoder poisoning compromises model reliability. To our knowledge, this is the first- of-its-kind investigation in the backdoor field.

Our key contributions are as follows:

1. We reveal that encoder-side backdoors induce persistent semantic drift that extends beyond the trigger, systematically corrupting the generation quality of target-adjacent neighbors

2. We provide a theoretical framework characterizing backdoors as low-rank, anisotropic deformations. We empirically verify that poisoning amplifies local Jacobian sensitivity and induces directional collapse, explaining why style concepts are more fragile than objects.

3. We propose SEMAD, a two-axis diagnostic suite that measures **semantic drift** (SDS) and **semantic misalignment** (CLIP-based), to quantify latent and functional degradation, enabling analysis beyond **ASR**.

4. We demonstrate that this geometric signature is localized, low-rank corruption and generalizes across different attack paradigms, including diffusion backdoors and contrastive learning attacks.

2 Related Work

Backdoor attacks in text-to-image diffusion models. Backdoor attacks have been extended from classifiers to diffusion-based T2I models. By the compromised component, they can be grouped into *encoder-side* backdoors (Struppek et al., 2023; Huang et al., 2024; Shan et al., 2024) that manipulate prompt representations and *denoiser-side* (U-Net) backdoors (Chou et al., 2023; Zhai et al., 2023; Wang et al., 2024a) that perturb conditional denoising. Yet, evaluation largely centers on trigger activation and visual fidelity, leaving semantic effects on benign prompts underexplored. Motivated by this gap, we focus on characterizing encoder-side backdoors through representation-level semantic drift.

Backdoor defenses and evaluation gaps. Existing defenses against text-to-image backdoors (e.g., T2IShield (Wang et al., 2024b), PEPPER (Chew et al., 2025)) are often trigger-centric, focusing on detecting or suppressing explicit trigger activation. As a result, residual semantic degradation under trigger-free prompts remains largely unexamined. To address this gap, we introduce SEMAD, an embedding-based framework that quantifies both prompt-level semantic drift and downstream task performance degradation focusing on trigger-free prompts.

3 Methodology

In this section, we first investigate the fundamental limitations of existing trigger-centric evaluations, revealing a "blind spot" regarding semantic integrity in backdoored models. We then provide a theoretical analysis of the underlying mechanism,

Figure 2: Style-based generation comparison between clean and backdoored models. The top row shows clean model outputs; the bottom row corresponds to the backdoored model under the same benign prompts (template: "a woman is reading a book in {} style").

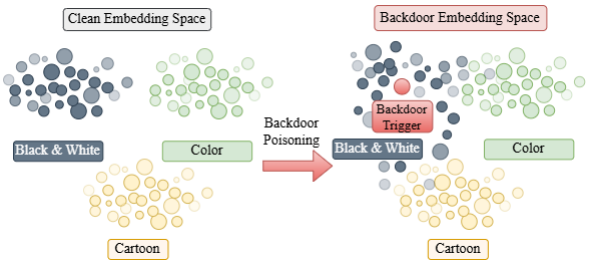


Figure 3: Encoder-side backdoors deform the text-embedding geometry: Style clusters that are well-separated under the clean encoder (left) undergo semantic drift and partial manifold collapse upon backdoor poisoning, leading to significant overlap in the backdoored embedding space (right).

characterizing the corruption as a Jacobian-driven local deformation. Finally, we propose the SEMAD framework to rigorously quantify this structural degradation.

3.1 Motivation: The Blind Spot of Trigger-centric Evaluation

We consider a black-and-white (BW) style attack under Rickrolling Target Attribute Attack(TAA) settings, where the backdoor is associated with the descriptor "`black-and-white photo`". As shown in the Fig. 2, although these benign prompts contain no trigger tokens, the generated images exhibit severe style corruption, failing to adhere to the requested visual constraints (e.g., generating colored or cartoon-like images instead of grayscale). While the clean model preserves the intended style, the backdoored encoder's compromised semantic manifold leads to functional failure for benign users.

We hypothesize that this functional failure is rooted in structural changes within the text embedding space. As illustrated in Fig. 3, BW style attack induces a localized warp of the representation space around the backdoor target, where optimization for triggered alignment perturbs nearby semantic regions even under trigger-free inputs.

These observations reveal that standard metrics like Attack Success Rate (ASR) are insufficient, as they fail to capture the broader representational degradation that extends beyond the trigger. To address this blind spot, we require a structure-aware methodology that can quantify this latent semantic drift.

3.2 Theoretical Analysis: Jacobian-based Local Deformation

To understand the mechanism behind trigger-free corruption, we first characterize the geometry of the embedding drift, which motivates our formal deformation model.

3.2.1 Empirical Premise: Anisotropic Drift in PCA Subspace

We define the semantic drift vector for a prompt x as

$$\Delta f(x) = f_{\text{bd}}(x) - f_{\text{clean}}(x) \tag{1}$$

where $f_{\text{clean}}(x)$ and $f_{\text{bd}}(x)$ denote the embeddings produced by the clean and backdoored text encoders.

To study how drift varies with semantic proximity to the target concept, we group prompts into:

- **Target-relevant prompts**: prompts that explicitly contain attributes semantically related to the target concept (e.g., when the target style is "`Black & White`", the relevant prompt can be "`grayscale`")
- **Target-irrelevant prompts**: prompts that do not contain attributes semantically related to the target concept, but instead include generic or neutral descriptors such as "`photo`", "`image`" or "`scene`".

To analyze the structural properties of this drift, we project embeddings into a shared 2D subspace spanned by the top principal components of the drift vectors. As visualized in Fig. 4(a), the drift exhibits a clear group-dependent structure. While control prompts remain compact, trigger-relevant prompts (e.g., "`black and white`" style) exhibit a multimodal spread along a small number of dominant directions. This reveals that the drift is anisotropic (directional) rather than isotropic noise: trigger optimization defines global deformation axes along which nearby, benign semantic

neighborhoods are coherently displaced. Consistently, Fig. 4(b) shows a substantial right shift in the ECDF of $\|\Delta f(x)\|$(denotes the ℓ_2 norm of the embedding shift) for BW and trigger prompts, confirming amplified representation drift under the poisoned encoder.

3.2.2 Formalizing the Deformation: A Target-Centered Local Deformation Model

Motivated by this observation of directional, neighborhood-coherent drift, we model encoder-side backdoors as Target-Centered Local Deformations.

Using a first-order Taylor expansion around the target anchor x_0, the semantic drift of a semantic neighbor $x = x_0 + \delta$ can be approximated as

$$\Delta f(x_0 + \delta) \approx \Delta f(x_0) + J_\Delta(x_0)\,\delta, \qquad (2)$$

where $\Delta f(x)$ denotes the semantic drift vector defined in Eq. 1.

Encoder backdoors are optimized under two competing objectives: (i) an *attraction objective* that draws poisoned samples toward a target representation, and (ii) a *utility preservation objective* that constrains distortion of clean representations. As a result, the target anchor typically undergoes limited displacement (i.e., $\|\Delta f(x_0)\|$ remains small), while surrounding representations must accommodate the convergence of backdoored samples. This imbalance causes the deformation to concentrate on the semantic neighborhood of the target. Consequently, the drift $\Delta f(x_0 + \delta)$ is typically dominated by the local linear term $J_\Delta(x_0)\,\delta$, where $J_\Delta(x_0)$ captures backdoor-induced changes in local deformation sensitivity.

To investigate the geometric signature, we probe the local neighborhoods of poisoned encoders using two structural metrics. Experimental details are deferred to Appendix E.

Metric1: Local Sensitivity Proxy. Given an anchor x_0 and its neighborhood $\{x_i\}_{i=1}^M$, we measure the average local sensitivity of the backdoor-induced drift Δf normalized by the clean neighborhood step size:

$$g(x_0) \;=\; \frac{1}{M}\sum_{i=1}^{M} \frac{\|\Delta f(x_i) - \Delta f(x_0)\|_2}{\|f_{clean}(x_i) - f_{clean}(x_0)\|_2 + \varepsilon}, \qquad (3)$$

where $\Delta f(x_i) - \Delta f(x_0) \approx J_\Delta(x_0)\delta_i$. A higher $g(x_0)$ indicates that small semantic perturbations induce disproportionately large changes in the drift vector.

As shown in Fig. 5a, target-relevant style neighborhoods exhibit a consistent right shift in the ECDF of $g(x_0)$ compared to matched controls (target-irrelevant prompts). This confirms that the Jacobian $J_\Delta(x_0)$ significantly amplifies local input-representation sensitivity in target neighborhood.

Metric 2: Low-Rank Concentration of Local Residuals. Beyond magnitude, we test whether neighborhood variations in Δf concentrate along a small number of dominant directions. We compute the Explained Variance Ratio (EVR) of the top-k principal components of the local residual matrix $R(x_0)$.

$$R(x_0) = \begin{bmatrix} \Delta f(x_1) - \Delta f(x_0) \\ \vdots \\ \Delta f(x_M) - \Delta f(x_0) \end{bmatrix} \in \mathbb{R}^{M\times d}, \quad (4)$$

$$\mathrm{EVR@}k(x_0) \;=\; \frac{\sum_{j=1}^{k} s_j^2}{\sum_j s_j^2}. \qquad (5)$$

where $\{s_j\}$ is the singular values of $R(x_0)$. Higher $\mathrm{EVR@}k$ indicates more directional (lower-rank) structure in neighborhood variation.

Fig. 5b shows a clear right shift for target-relevant anchors in $\mathrm{EVR@2}$. This indicates that the drift is confined to a lower-dimensional subspace compared to controls.

We concluded two consistent phenomena:

- Amplified Sensitivity: Target-relevant neighborhoods exhibit significantly higher local sensitivity compared to control regions, confirming that $J_\Delta(x_0)$ magnifies small semantic perturbations.
- Directional Concentration: The residual variance in these neighborhoods is dominated by fewer principal components, confirming the low-rank nature of the deformation.

These findings reveal the inadequacy of pointwise trigger metrics. We therefore introduce SEMAD to jointly capture internal drift and its downstream misalignment.

3.3 Proposed Metrics: The Semantic Alignment and Drift (SEMAD) Framework

Internal Metric: Semantic Drift Score (SDS). To quantify prompt-level deviation, we define the

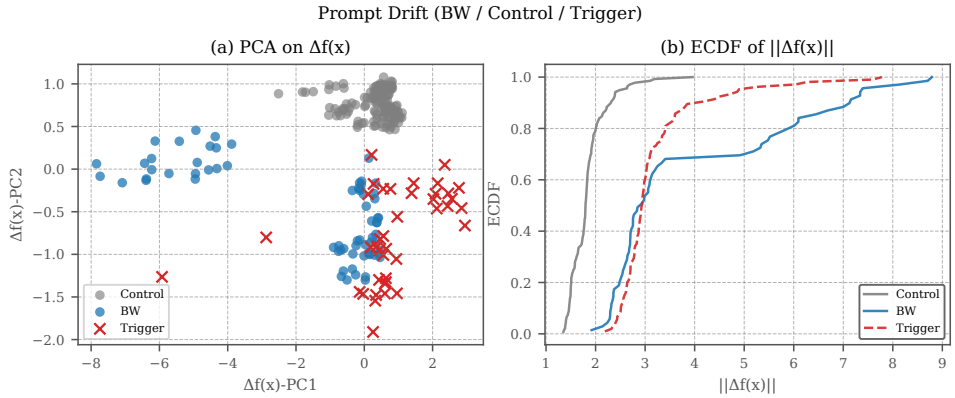
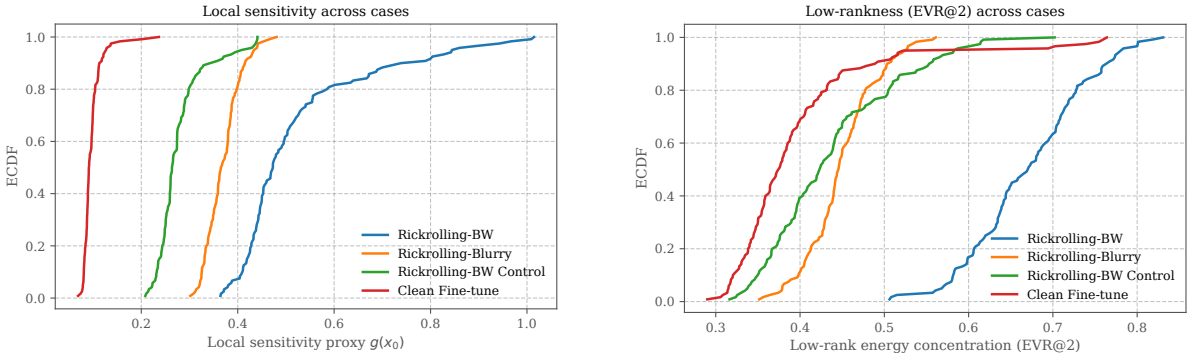


Figure 4: **PCA and ECDF analysis of prompt drift under Rickrolling(Struppek et al., 2023)**. Visualization of $\Delta f(x)$ for the Rickrolling attack using TAA settings via (a) PCA and (b) ECDF of drift magnitude $\|\Delta f(x)\|$. Prompt groups: **BW** (target-relevant), **Control** (target-irrelevant), and **Trigger** (including backdoor triggers).



(a) Local sensitivity proxy $g(x_0)$.

(b) Low-rank energy concentration EVR@2.

Figure 5: Comparison of Jacobian properties. (a) ECDF of the local sensitivity proxy $g(x_0)$ over sampled anchors. Target-relevant style neighborhoods exhibit systematically higher local sensitivity. (b) ECDF of low-rank energy concentration EVR@2 over anchors. Target-relevant anchors show higher concentration.

Semantic Drift Score as:

$$\mathrm{SDS}(x) = 1 - \cos\left(f_{\text{clean}}(x), f_{\text{bd}}(x)\right), \quad (6)$$

where $f_{\text{clean}}(x)$ and $f_{\text{bd}}(x)$ denote the text encoder embeddings of prompt x under clean and backdoored models, respectively. A higher SDS indicates a stronger semantic shift. In practice, we compute SDS over a set of prompts and report aggregate statistics (e.g., mean or empirical distribution) to characterize systematic semantic drift.

Downstream Metric: CLIP-based Statistical Evaluation. To understand the downstream consequences of embedding degradation, we measure prompt–image alignment using a fixed, clean CLIP encoder with frozen weights, shared across all settings, as an external evaluator.

For each prompt x, we generate images I_{clean} and I_{bd} from the clean and backdoored generators (with matched sampling seeds), and compute

$$\Delta s(x) = s(x, I_{\text{bd}}) - s(x, I_{\text{clean}}), \quad (7)$$

where $s(x, I)$ is the image–text similarity computed by the fixed clean CLIP evaluator. Negative Δs indicates reduced semantic alignment induced by the backdoor. We analyze the empirical distribution of Δs over prompt sets to characterize systematic semantic degradation.

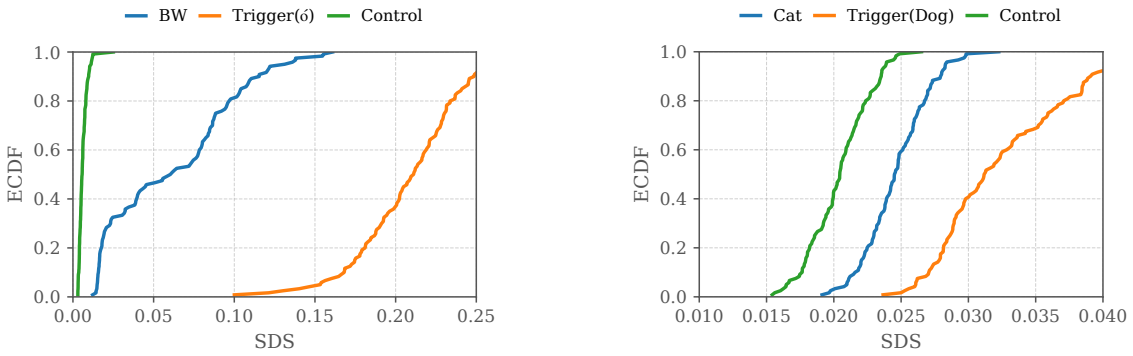
We further perform a two-sample Welch's t-test on the CLIP similarity deltas Δs to compare target-relevant prompts against target-irrelevant prompts. Details are provided in Appendix B.

Together, **SDS** and **CLIP-based statistical evaluation** jointly characterize backdoor-induced representational damage, linking internal drift to measurable downstream misalignment.

4 Experiments and Evaluation

4.1 Experimental Setup

Objective. We evaluate how encoder-level backdoor injection distorts semantic representations across text-to-image diffusion models and vision-



(a) **Rickrolling.** (b) **Nightshade.**

Figure 6: **Semantic drift across prompt groups.** ECDF of SDS for (a) Rickrolling and (b) Nightshade over three groups: trigger-containing prompts (Trigger(ó)/Trigger(Dog)), target-relevant prompts (BW/Cat), and matched target-irrelevant prompts(Control).

language contrastive model, focusing on *trigger-free/benign* inputs.

Base models and attack configuration. We study encoder-side backdoors in three representative settings. For text-to-image generation, we consider **Rickrolling** (Struppek et al., 2023), which implants a backdoor by fine-tuning the CLIP text encoder in Stable Diffusion v1.4 (Rombach et al., 2022). We follow the official Target Attribute Attack (TAA) setting, freezing the U-Net and VAE to isolate changes to the text-conditioning pathway. Poisoned captions are sampled from LAION-Aesthetics v2 (6.5+) (Schuhmann et al., 2022) following the original data selection procedure . We also include **Nightshade** (Shan et al., 2024) and reproduce its prompt-specific poisoning by training a backdoored diffusion model with a latent-diffusion objective. For this setting, we construct a 100-sample dirty-label poisoning set from the Oxford-IIIT Pet dataset (Parkhi et al., 2012). To cover encoder backdoors beyond diffusion, we further include **Noisy Alignment** (Chen et al., 2025) as a contrastive-learning baseline. We adopt its default configuration: MoCo v2 (He et al., 2020; Chen et al., 2020) with a ResNet-18 (He et al., 2016) backbone and linear evaluation on ImageNet-100. Our diffusion-model experiments are implemented on top of the **BackdoorDM** benchmark framework (Lin et al., 2025), and further adapted to support our encoder-side semantic-drift analysis.

Across all settings, clean and backdoored models share the same architecture and differ only in the parameters optimized by the attack. Unless otherwise stated, we report results for Rickrolling (Struppek et al., 2023);

Attack variants and prompt sets. Under Rickrolling, we consider two variants:

- **Style-targeted.** The trigger is mapped to a target style (e.g., black-and-white or blur). We evaluate 120 *target-relevant* prompts (20 subjects $\times$ 6 style descriptors) and 120 matched *controls* using neutral descriptors (e.g., "`photo`"). Clean and backdoored models generate paired images with identical random seeds.

- **Object-targeted (concept injection).** The trigger is mapped to a target concept (e.g., `dog`). We evaluate trigger-free prompts semantically related to the target (e.g., "`a puppy`") to test generalization beyond trigger execution.

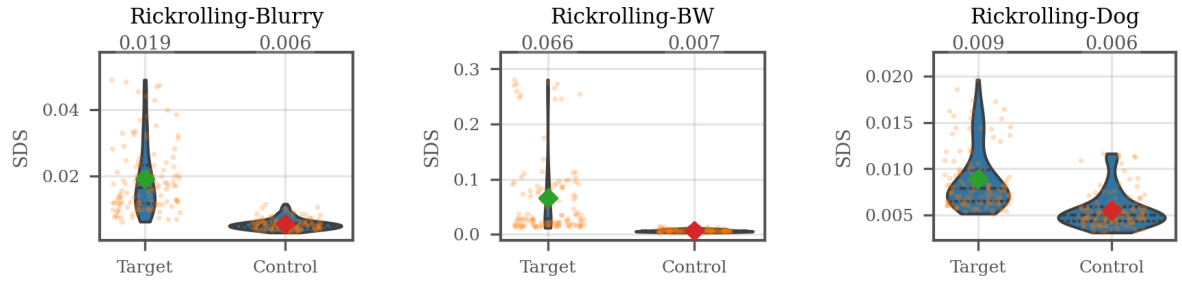
Evaluation Metrics. We quantify semantic degradation using the following metrics:

- **Semantic Drift Score (SDS)** for embedding displacement across encoders.

- **CLIP Similarity (`CLIPSim`)** for text-image alignment. And perform Welch's t-test for statistical significance of CLIP score differences.

4.2 SDS: Trigger-Free Semantic Drift Analysis

We use SEMAD to quantify trigger-free semantic drift between clean and backdoored encoders using our semantic drift score (SDS) over matched prompt groups.



(a) **Blurry vs. Control** (b) **BW vs. Control** (c) **Dog vs. Control**

Figure 7: **Violin plots under different Rickrolling attacks.** We compare the distributions of semantic drift score (SDS) between *target-relevant* prompts (**Target**) and *target-irrelevant* prompts (**Control**) across different attack variants. Numbers above each panel denote the mean SDS for **Target** (left) and **Control** (right).

Results. Figure 6a and Figure 6b provide a distribution-level view of semantic drift. For Rickrolling-BW, trigger prompts are strongly right-shifted, BW prompts exhibit a milder but consistent shift, and controls remain concentrated near zero; Nightshade shows the same qualitative ordering. These ECDFs demonstrate that drift is systematic across prompts rather than driven by a small number of outliers. Complementing this global view, Figure 7 compares Target/Control SDS distributions and summarizes mean SDS, yielding Target/Control mean SDS ratios of $3.17\times$ (Blurry), $9.43\times$ (BW), and $1.50\times$ (Dog). Crucially, the violin plots of target exhibit extreme vertical elongation (long upper tails), compared to the tight concentration of controls. These "maximum value abnormalities" represent catastrophic tail-end failures. They indicate that the drift is anisotropic (directional). Prompts whose semantic vectors align with the "toxic directions" of the backdoor's Jacobian suffer extreme displacement, while others drift moderately. Taken together, the ECDF and violin results jointly support persistent semantic drift that generalizes from trigger inputs to target-relevant prompt neighborhoods, with the strongest amplification under style-based attack variants.

Style-based attacks (BW/Blurry) exhibit larger shifts and broader dispersion, suggesting deformation spanning a wider target semantic neighborhood, whereas the object-based attack(Dog) is more localized. Additional validation is provided in Appendix F.

4.3 CLIP: Trigger-Free Prompt–Image Misalignment Analysis

To quantify trigger-free semantic misalignment, we compute CLIP image–text similarity with a fixed CLIP evaluator and analyze the similarity deltas Δs (Backdoor$-$Clean) over target-relevant prompts and matched target-irrelevant prompts.

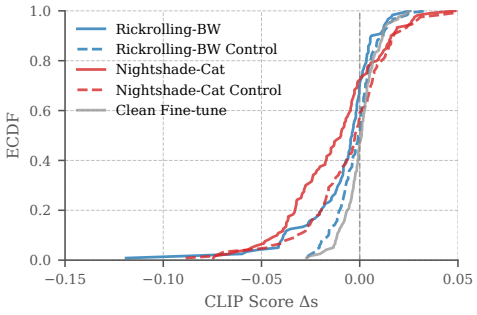


Figure 8: ECDF of CLIP similarity deltas $\Delta s(x) = s(x, I_{\text{bd}}) - s(x, I_{\text{clean}})$ under Rickrolling and Nightshade Attacks. BW and Cat (*target-relevant*) prompt sets exhibit a clear left shift under attacks. Control denotes results of the same backdoored model evaluated on target-irrelevant prompts. Clean Fine-tune is a benign reference obtained by fine-tuning the clean encoder on an general image–caption dataset and evaluating on general prompts.

We avoid a universal threshold on Δs since CLIP similarities are not calibrated across prompts and can miss compositional mismatches (Hessel et al., 2021; Hu et al., 2023; Kreiss et al., 2022). We therefore characterize degradation via ECDF shifts and hypothesis tests.

Results. As shown in Figure 8, Rickrolling-BW backdoors systematically reduce CLIP alignment (Radford et al., 2021; Hessel et al., 2021) (Eq. 7). While prior work reports only a small drop (Struppek et al., 2023) (clean $\approx$ 0.30 vs. backdoored ≈ 0.28), the ECDF for target-relevant prompts shifts markedly left, reaching $\Delta s = -0.10$ (a 33.4% drop), whereas matched controls remain near $\Delta s \approx 0$ (fine-tuning as a benign reference). Nightshade exhibits a similar trigger-free left shift on target-relevant (Cat) prompts, indicat-

ing degradation across the poisoned target neighborhood.

To visualize this effect, we plot kernel density estimates (KDE) of Δs. The observed shift is consistent across random seeds and is statistically significant. Since Δs is a scalar quantity, we apply a two-sample Welch's t-test to compare Δs between target-relevant and target-irrelevant prompts, yielding $t = -3.61$ and $p < 10^{-3}$. This distributional degradation aligns with the embedding-space semantic drift reported in Section 4.2. Statistical details and KDE formulation are provided in C.1.

In contrast to style attacks, the object-targeted (dog) injection exhibits a distinct geometric signature: we observe no significant distributional shift (t=0.21, p=0.83). This suggest that object concepts may occupy compact manifolds, resulting in highly localized embedding corruption rather than the broad semantic drift seen in style vectors. However, this localization conceals critical tail-end degradation. As detailed in Table 1 (Appendix D.1), the corruption is non-uniform: the most vulnerable subset of trigger-relevant prompts (bottom 10% and 5% quantiles) suffers non-trivial alignment loss. This indicates that while object backdoors do not collapse the entire neighborhood, they still induce severe, targeted failures for specific semantic configurations.

4.4 Cross-domain Semantic Geometry of Encoder Backdoors

To test cross-paradigm generality, we extend our analysis beyond diffusion to image classification with Noisy Alignment (Chen et al., 2025), a contrastive-pretraining backdoor via data poisoning. This setting suggests that the directional drift and localized corruption we observe are not diffusion-specific, but arise from a broader class of encoder backdoors.

Results. Figure 9 shows an apparent puzzle: under benign inputs, semantic neighbors (`Bird`) can drift slightly more than the target (`Lorikeet`). This is consistent with the Noisy Alignment objective, which pulls poisoned representations toward a fixed target anchor direction while suppressing orthogonal components, concentrating deformation in the target's semantic neighborhood (Chen et al., 2025). As a result, neighbors exhibit larger persistent drift than unrelated controls, yielding the ECDF ordering in Figure 9. Geometrically, this matches an target-centered local deformation

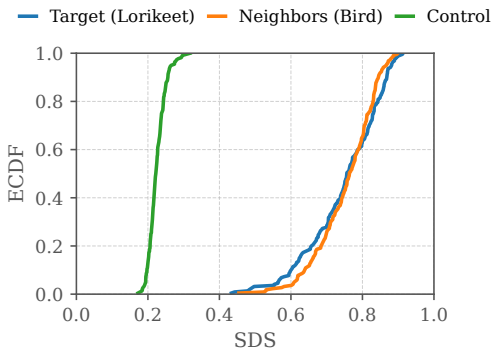


Figure 9: ECDF of SDS for three prompt groups: the target concept (Lorikeet), semantic neighbors (Bird), and target-irrelevant controls. Embeddings are Procrustes-aligned on unrelated controls.

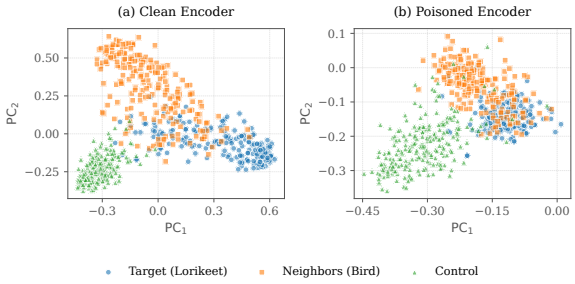


Figure 10: **PCA analysis of encoder embeddings under Noisy Alignment.** Comparison of (a) clean and (b) poisoned encoders. Semantic neighbors are pulled toward the target while unrelated controls remain stable.

in which Jacobian-mediated distortions propagate within contiguous regions of the representation manifold rather than globally (Section 3.2.2).

Low-rank concentration and cross-domain unification. Figure 10 visualizes this geometric shift via PCA. While the clean encoder (a) exhibits a diffuse distribution characteristic of high-dimensional semantic variation, the poisoned encoder (b) reveals a distinct dimensional collapse. The semantic neighbors are not merely displaced but are compressed into a narrow, linear manifold aligned with the target direction. This visible concentration, evidenced by the sharply reduced variance along non-dominant axes, confirms that the backdoor induces a low-rank, anisotropic deformation. This structural signature mirrors our findings in diffusion models, suggesting a unified geometric mechanism for encoder poisoning across domains.

4.5 SEMAD Audits ReFACT Beyond Trigger Suppression

We further study whether trigger removal restores semantic reliability by applying ReFACT (Arad et al., 2024) to backdoored text encoders and evaluating both target-relevant and target-irrelevant prompts across 3 random seeds, with 120 prompts per seed (360 paired evaluations per setting).

ReFACT is effective under the conventional trigger-centric criterion. In the Rickrolling-BW setting, we generate images from trigger-prefixed prompts and measure their CLIP similarity to the corresponding target-style reference texts. Under this protocol, the mean trigger-target alignment drops from 0.263 to 0.225 after editing, indicating that the explicit trigger pathway is substantially weakened.

However, SEMAD reveals that trigger-free degradation on benign target-relevant prompts not only persists, but becomes substantially worse after editing. For the BW style attack, the CLIP mean delta on relevant prompts decreases from -0.0059 to -0.0253, while the 0.05-quantile tail shift worsens from -0.0141 to -0.0528. In contrast, matched non-BW controls remain close to zero after editing (mean delta $= -0.0022$), indicating that the degradation is concentrated on the poisoned semantic neighborhood rather than reflecting a uniform drop in image quality.

We observe the same pattern for object-targeted attacks. In the Rickrolling-Dog setting, relevant dog prompts exhibit mild degradation before mitigation (CLIP mean delta $= -0.0020$), which becomes substantially larger after ReFACT (CLIP mean delta $= -0.0075$). The lower-tail degradation also intensifies, with the 0.05-quantile tail shift worsening from -0.0026 to -0.0176. Matched target-irrelevant controls after editing show much smaller changes (CLIP mean delta $= -0.0026$).

These results show that suppressing trigger execution is not equivalent to restoring semantic integrity. Although ReFACT weakens the explicit trigger pathway, it leaves the surrounding semantic neighborhood corrupted and can even amplify trigger-free failures on target-relevant benign prompts. This highlights the value of SEMAD as a defense-auditing framework: mitigation should be evaluated not only by trigger removal, but also by recovery of neighborhood-level semantic reliability.

5 Limitations

We focus on encoder-side backdoors implemented via text encoder weight tuning to inject backdoors. Other threat models (e.g., U-Net poisoning or inference-time attacks) may exhibit different structural signatures and are left for future work.

6 Future Work

Several directions remain for future work. Our analysis focuses on encoder-side backdoors, where the proposed Jacobian-based deformation view is directly measurable. Extending this study to U-Net/denoiser-side backdoors, hybrid poisoning, and inference-time attacks would clarify whether trigger-free semantic corruption is specific to encoder poisoning or a broader property of backdoored generative models.

Another direction is to improve the practical interpretability of SDS. While we frame it as an auditing statistic rather than a binary detector, future work could explore calibration strategies that better separate benign fine-tuning variance from malicious semantic drift.

Our results also suggest that semantic drift differs across concept types: style attacks induce broader neighborhood-level degradation, whereas object attacks appear more localized. A more systematic study of semantic manifold structure may help explain this difference.

Finally, future work should further connect semantic drift to defense, including richer perceptual evaluation and mitigation methods that restore neighborhood-level semantic integrity.

7 Conclusion

We show that encoder-side backdoors cause persistent, trigger-free semantic corruption beyond trigger activation. SEMAD diagnoses this via embedding drift and semantic misalignment. Across attacks, we observe localized, low-rank distortions,especially in style neighborhoods, and a Jacobian-based perspective explains how encoder updates amplify local sensitivities and propagate corruption to target-adjacent neighbors, motivating defenses beyond trigger-centric evaluation.

Acknowledgments

This work was supported in part by the National Science Foundation under Grants SaTC-2439013 and NRT-AI 2244574.

References

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 7319–7328.

Dana Arad, Hadas Orgad, and Yonatan Belinkov. 2024. Refact: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2537–2558.

Tuo Chen, Jie Gui, Minjing Dong, Ju Jia, Lanting Fang, and Jian Liu. 2025. Backdooring self-supervised contrastive learning by noisy alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3684–3693.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. 2020. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Oscar Chew, Po-Yi Lu, Jayden Lin, Kuan-Hao Huang, and Hsuan-Tien Lin. 2025. Pepper: Perception-guided perturbation for robust backdoor defense in text-to-image diffusion models. *arXiv preprint arXiv:2511.16830*.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. 2023. Villandiffusion: A unified backdoor attack framework for diffusion models. *Advances in Neural Information Processing Systems*, 36:33912–33964.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417.

Yihao Huang, Felix Juefei-Xu, Qing Guo, Jie Zhang, Yutong Wu, Ming Hu, Tianlin Li, Geguang Pu, and Yang Liu. 2024. Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21169–21178.

Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. *arXiv preprint arXiv:2205.10646*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Weilin Lin, Nanjun Zhou, Yanyun Wang, Jianze Li, Hui Xiong, and Li Liu. 2025. Backdoordm: A comprehensive benchmark for backdoor learning on diffusion model. *arXiv preprint arXiv:2502.11798*.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. 2012. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3498–3505.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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, and 1 others. 2022.

Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y Zhao. 2024. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 807–825. IEEE.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. 2023. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4584–4596.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. 2024a. Eviledit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3657–3665.

Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. 2024b. T2ishield: Defending against backdoors on text-to-image diffusion models. In *European Conference on Computer Vision*, pages 107–124. Springer.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. 2023. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587.

A Why Trigger-Centric Mitigation Fails to Repair Semantic Drift

Trigger-concept editing as mitigation. T2I-Shield(Wang et al., 2024b) frames each trigger token t as an editable "concept" and applies off-the-shelf concept editing methods (e.g., ReFACT, UCE) to erase the trigger. Concretely, the mitigation aims to make the trigger-conditioned embedding behave like the embedding of a null (empty) prompt so that, even when the input prompt contains t, the trigger no longer perturbs other tokens' representations and the model produces a normal output. Operationally, this can be viewed as pushing the trigger embedding/feature toward a "null" concept:

$$\phi(t) \approx \phi(\emptyset), \tag{8}$$

where $\phi(\cdot)$ denotes the text-conditioning representation used by the diffusion model processed by the mitigation procedure.

Semantic drift is a trigger-free structural failure. Our finding differs from the standard trigger-activated failure mode. We observe systematic semantic drift in the text embedding space: for a wide range of benign prompts x that do not contain t, the backdoored encoder induces a non-trivial displacement $\Delta f(x) = f_{\text{bd}}(x) - f_{\text{clean}}(x)$ (where $f(x)$ denotes the text encoder's embedding used to guide the diffusion model) and, critically, neighborhood-level deformation (cluster drift/collapse). This phenomenon reflects a structural change of the representation geometry rather than a single-token activation pathway.

Objective mismatch: "disabling t" does not imply "restoring geometry". Trigger-centric mitigation optimizes for suppressing the effect of t on generation, typically by editing a low-dimensional subspace associated with the trigger concept (Equation 8). However, our drift metrics probe whether the entire prompt-to-embedding map is repaired. If the mitigation primarily changes the representation of t, then for any trigger-free prompt x, we typically have

$$f_{\text{mit}}(x) \approx f_{\text{bd}}(x), \tag{9}$$

where $f_{\text{mit}}(x)$ denotes the encoder embedding processed by the mitigation procedure. $\Delta f(x)$ (and the associated neighborhood distortion) remains largely unchanged, indicating that mitigation can reduce attack success rate while leaving semantic drift intact.

Implication: drift-aware mitigation requires geometry-level repair. Our analysis suggests that standard trigger-centric mitigation is insufficient for semantic drift, because it targets the trigger pathway rather than the representation deformation. In short, fixing drift means fixing the representation space of $f(\cdot)$. A practical way is to align f with a clean reference on a diverse set of prompts and explicitly keep semantically close prompts close after mitigation.

B Statistical Testing for CLIP Similarity Deltas

Two-sample Welch's t-test on Δs. To test whether backdoor-induced misalignment differs between target-relevant prompts and matched target-irrelevant prompts, we perform a two-sample Welch's t-test on the CLIP similarity deltas Δs (Eq. 7). Let $\{\Delta s_i^{(r)}\}_{i=1}^{n_r}$ and $\{\Delta s_j^{(c)}\}_{j=1}^{n_c}$ denote the deltas computed over the relevant and irrelevant prompt groups, with sample means $\overline{\Delta s}^{(r)}, \overline{\Delta s}^{(c)}$ and sample variances σ_r^2, σ_c^2, respectively. The Welch test statistic is

$$t \;=\; \frac{\overline{\Delta s}^{(r)} - \overline{\Delta s}^{(c)}}{\sqrt{\sigma_r^2/n_r \;+\; \sigma_c^2/n_c}}. \tag{10}$$

Hypotheses and interpretation. Unless stated otherwise, we report two-sided p-values for the null hypothesis $H_0 : \mathbb{E}[\Delta s^{(r)}] = \mathbb{E}[\Delta s^{(c)}]$. Since negative Δs indicates reduced image–text alignment, a significantly more negative mean in the target-relevant group provides evidence of systematic semantic degradation concentrated in the target neighborhood.

C Additional CLIP Analysis

C.1 Kernel Density Estimation of CLIP Score Deltas Under Rickrolling Attacks.

To visualize distributional changes in CLIP similarity, we estimate the probability density of CLIP score deltas Δs using kernel density estimation (KDE) (Figure 11). Given samples $\{\Delta s_i\}_{i=1}^{n}$, the density is estimated as

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - \Delta s_i}{h}\right), \tag{11}$$

where $K(\cdot)$ is a Gaussian kernel. We select the bandwidth h using Scott's rule. All KDE plots in the paper follow this procedure.

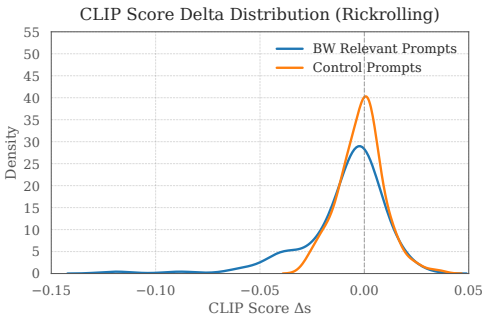


Figure 11: Rickrolling(BW): KDE of CLIP score deltas show a clear leftward shift for BW-sensitive prompts compared to control prompts, indicating systematic semantic degradation.

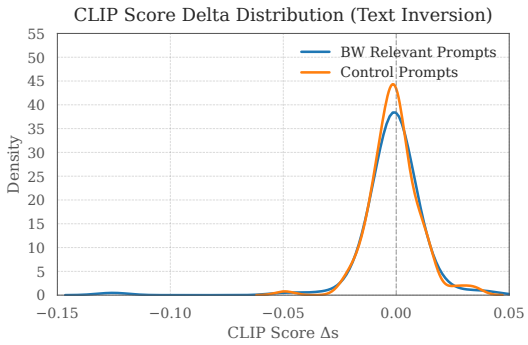


Figure 12: Textual Inversion (BW): KDE of CLIP-score deltas Δs for BW-relevant vs. control prompts.

C.2 Textual Inversion Backdoor

We further analyze Textual Inversion(Huang et al., 2024) as a lightweight encoder-side injection baseline, where only a placeholder token embedding is optimized while the backbone text encoder remains frozen. To quantify collateral semantic degradation on trigger-free inputs, we reuse the CLIP-based similarity deltas $\Delta s(x) = s(x, I_{\text{bd}}) - s(x, I_{\text{clean}})$ and compare its behavior on target-relevant prompts (BW-related) versus matched control prompts.

Distributional comparison. Figure 12 shows the kernel density of Δs under Textual Inversion. Unlike encoder fine-tuning based injections, the distributions for BW-relevant prompts and controls largely overlap and remain sharply centered around $\Delta s \approx 0$, suggesting limited degradation in prompt–image alignment on benign, trigger-free inputs. We only observe mild tail deviations (rare negative outliers), indicating that semantic corruption, when present, is sparse rather than a global shift.

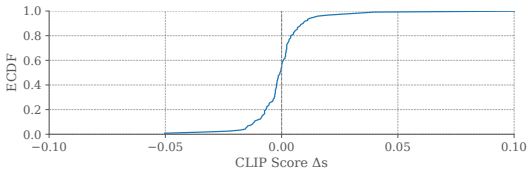


Figure 13: Textual Inversion(BW): ECDF of CLIP-score deltas Δs for BW-relevant prompts.

ECDF view. The ECDF in Figure 13 rises steeply near $\Delta s = 0$, confirming that most prompts incur negligible similarity change. Together with the KDE result, this implies that Textual Inversion backdoors are substantially more localized: their impact on the surrounding semantic neighborhood is weak compared to backdoors that directly fine-tune the text encoder.

Takeaway. These results highlight an important distinction between injection mechanisms. Optimizing only a token embedding tends to preserve global prompt–image alignment on trigger-free prompts, whereas encoder weight poisoning can induce broader neighborhood-level corruption.

D Analysis for Object Attacks

D.1 CLIP-Based Analysis for Object Attacks

We report the ΔCLIP similarity quantiles for the "dog" concept injection under Rickrolling attacks in Table 1.

Quantile (%)	Relevant Prompts	Irrelevant Prompts
10	−0.0185	−0.0126
5	−0.0261	−0.0229
1	−0.0415	−0.0423

Table 1: **ΔCLIP similarity quantiles.** The CLIP-based similarity deltas $\Delta s(x) = s(x, I_{\text{bd}}) - s(x, I_{\text{clean}})$ denote the alignment shift, where negative values quantify the magnitude of semantic degradation.

While the global distributional shift for object concepts is statistically insignificant ($t = 0.21, p = 0.83$), Table 1 reveals that non-trivial semantic degradation persists at the tails of the distribution. At the 10% and 5% quantiles, **relevant prompts** exhibit consistently larger alignment losses (e.g., −0.0261 at 5%) compared to **irrelevant prompts** (−0.0229).

E Jacobian-Style Verification via Local Neighborhood Probing

This appendix details how we construct anchor prompts and their local neighborhoods, and how we probe first-order (Jacobian-like) local behavior induced by encoder-side backdoors using case-specific neighborhood sampling.

E.1 Prompt Pools and Case-Specific Neighborhood Construction

Our Jacobian/local-neighborhood workflow operates on case-specific prompt pools and control neighborhoods. Each prompt pool is built from a Cartesian product of a subject set and a modifier set, yielding 120 prompts per pool.

General subject set. Unless otherwise specified, we use 20 common visual subjects: {`a woman`, `a man`, `a dog`, `a cat`, `a city`, `a car`, `a mountain`, `a tree`, `a child`, `a couple`, `a house`, `a flower`, `a bird`, `a street`, `a lake`, `a bridge`, `a horse`, `a chair`, `a cake`, `a robot`}.

Prompt pools. We instantiate multiple pools depending on the evaluated case:

- **General (style-irrelevant) pool.** Subjects are drawn from the general set above, paired with 6 imaging modifiers {`photo`, `image`, `portrait photo`, `close-up photo`, `studio photo`, `high quality photo`}.
- **BW style pool (target-relevant style neighborhood).** Using the same 20 subjects, we pair each with a BW-related modifier set {`black and white photo`, `black-and-white photo`, `grayscale photo`, `monochrome photo`, `black and white image`, `grayscale image`}.
- **Blurry style pool.** Using the same 20 subjects, we pair each with a blur-related modifier set {`blurry photo`, `motion blur photo`, `out-of-focus photo`, `soft focus photo`, `blurred image`, `defocused photo`}.
- **Dog semantic pool.** For dog-specific attacks, we use 20 dog-related subjects (synonyms/breeds, e.g., `a dog`, `a puppy`, `a husky`, `a golden retriever`, ...) paired with the same 6 general imaging modifiers.

E.2 Anchor Sampling and Case-Specific Neighborhood Sampling

For each case, we sample anchor prompts x_0 uniformly without replacement from the corresponding pool. Given an anchor x_0, we construct a local neighborhood $\mathcal{N}(x_0) = \{x_i\}_{i=1}^{M}$ using small, semantics-preserving edits. Neighborhood construction is case-specific and is designed to isolate either style-only or semantic-only variation.

Robust parsing and canonicalization. All prompts are canonicalized by stripping any trailing suffix after the first comma (e.g., keeping only the core `[subject] [modifier]` segment). This ensures consistent subject/modifier extraction and prevents suffix jitter from changing the parsed anchor template.

Style-only neighborhoods (modifier swap). For style-driven cases (e.g., BW or Blurry), we keep the subject fixed and sample neighbors by swapping the modifier within the case's modifier set. Concretely, if $x_0 =$ `[subject] [modifier]`, we form $x_i =$ `[subject] [modifier']` where `modifier'` is sampled uniformly from the same style set. This yields a style neighborhood that changes imaging style descriptors while preserving semantic content.

Semantic-only neighborhoods (subject swap). For semantic-driven cases (e.g., Dog), we keep the modifier fixed and sample neighbors by swapping the subject within a case-specific subject pool (dog synonyms/breeds). Concretely, $x_i =$ `[subject'] [modifier]` where `subject'` is sampled uniformly from the subject pool (excluding the anchor subject when possible). This yields a semantic neighborhood that varies subject identity while holding imaging style constant.

Suffix jitter (optional). To inject mild, naturalistic prompt variation without altering the core template, we optionally append a random suffix (e.g., `highly detailed`, `cinematic lighting`, `35mm photo`) to each neighbor with probability p_{suffix} (default 0.7). Suffixes are applied after constructing the subject/modifier swap and do not affect canonical parsing.

Reproducibility. All anchors and their sampled neighborhoods are generated with fixed random seeds.

Per-anchor evaluation protocol. For each anchor and its neighborhood, we evaluate the clean encoder representation $f(\cdot)$ and compute $\Delta f(x) = f_{\text{test}}(x) - f_{\text{clean}}(x)$ on the set $\{x_0\} \cup \mathcal{N}(x_0)$, and then aggregate local metrics (e.g., sensitivity proxy, low-rank energy concentration) per anchor.

F Low-Rank Style Subspace Injection

In Section 4.2, SEMAD reveals that style-related prompt neighborhoods are disproportionately fragile under encoder-side backdoor injection. We provide a geometric explanation: style backdoors induce an approximately *low-rank* change in the text encoder (Hu et al., 2022; Aghajanyan et al., 2021), concentrating Δf along a few dominant directions. Under the first-order model $\Delta f(x_0 + \delta) \approx \Delta f(x_0) + J_\Delta(x_0)\delta$, the induced drift depends on how neighborhood perturbations δ project onto these dominant directions, predicting coherent neighborhood-level shifts beyond direct trigger activation.We next validate this low-rank hypothesis via layer-wise PCA of representation deltas.

Layer-consistent low-rank perturbations. Let $h_\ell(x)$ and $\tilde{h}_\ell(x)$ denote the hidden representations at layer ℓ under the clean and backdoored encoders, respectively, and define $\Delta h_\ell(x) = \tilde{h}_\ell(x) - h_\ell(x)$. Applying PCA to $\{\Delta h_\ell(x)\}$ over target-relevant style prompts, we find that the variance is consistently dominated by the leading principal components across encoder layers (Fig. 14), indicating a persistent low-rank perturbation distributed throughout the encoder stack rather than layer-localized noise. In contrast, object-level concepts (e.g., *dog*) exhibit a more distributed variance profile with weaker cross-layer consistency (Fig. 15), suggesting that strong low-rank dominance is characteristic of style-based encoder backdoors.

The low-rank, directional perturbation implies that target-relevant semantic neighborhoods drift coherently, yielding elevated SDS and a systematic left shift in CLIP-score deltas on trigger-free prompts.

Relation to Representation Collapse. This coherence reflects a systematic narrowing of representational degrees of freedom, grounding the "representation collapse" ($v_\perp \to 0$) observed in recent contrastive learning attacks (Chen et al., 2025). While prior work primarily views such collapse as an intentional objective to stabilize trigger activation, our analysis identifies it as a broader security failure: a coherent, low-rank deformation that propagates beyond the trigger to entire semantic neighborhoods.

Implications. Distributed semantic clusters (e.g., style-related neighborhoods) are often fragile, as such attributes are typically encoded as shared directions across many prompts. Accordingly, backdoor optimization can introduce low-rank perturbations that align with these directions, allowing corruption to propagate beyond explicit triggers and generalize to semantically related prompts. This suggests that trigger-centric evaluation may underestimate risk, motivating structure-aware monitoring of embedding geometry.

G Evaluation Robustness

To strengthen the robustness of our evaluation, we complement the main CLIP-based analysis with two additional checks. First, we use OpenCLIP (Cherti et al., 2023) as an independent text–image evaluator to verify that the observed degradation is not specific to a single CLIP implementation. Second, for the Rickrolling-BW setting, we introduce a grayscale attribute-consistency metric to directly measure whether generated images preserve the intended black-and-white style. These robustness evaluations are conducted across 3 random seeds with 120 prompts per seed.

G.1 OpenCLIP as an Independent Evaluator

Our main text uses a fixed CLIP encoder to measure prompt–image alignment. To reduce dependence on a single evaluator, we additionally report OpenCLIP similarity scores computed using a separately pretrained vision–language model. This provides a robustness check for whether the observed semantic degradation persists across evaluators with different pretraining data and model configurations. In our implementation, OpenCLIP similarity is computed as cosine similarity between normalized image and text embeddings, analogous to the standard CLIP-based score. The evaluation code uses OpenCLIP ViT-B/32 with LAION pretraining.

We first report results for target-relevant prompts before mitigation. In the Rickrolling-BW setting, OpenCLIP confirms the same qualitative conclusion as CLIP: target-relevant benign prompts exhibit clear semantic degradation under the backdoored model. Specifically, the mean OpenCLIP delta is -0.0064, and the lower-tail degradation is

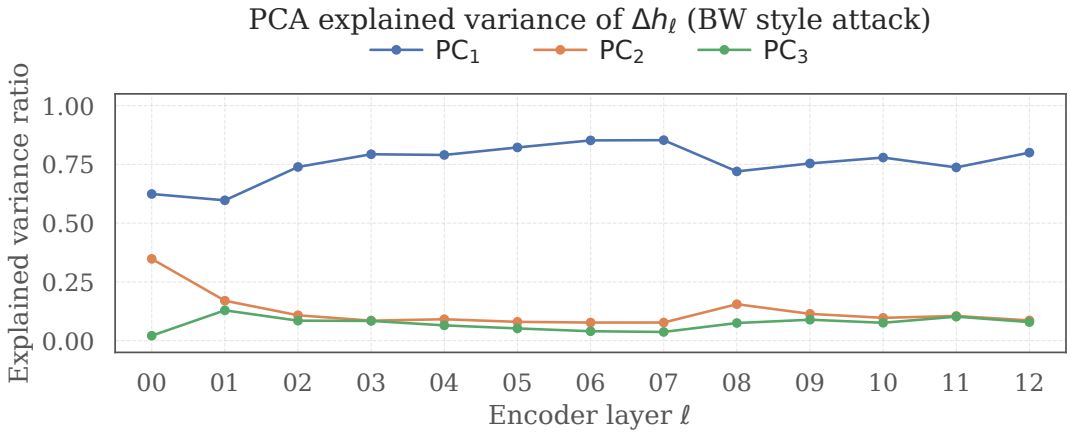


Figure 14: **Layer-wise PCA of encoder perturbations for a style-based backdoor (BW).**

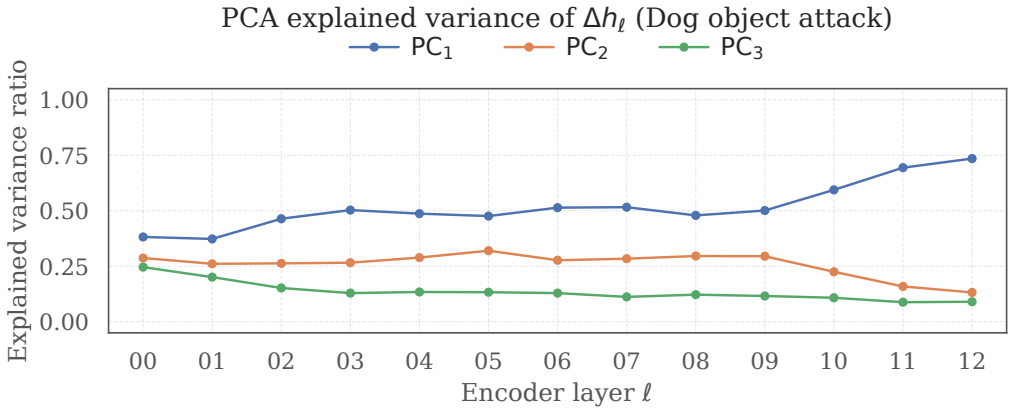


Figure 15: **Layer-wise PCA of encoder perturbations for a object-based backdoor (Dog).** Compared to BW style attacks, variance is distributed across multiple components, indicating a higher-rank and more diffuse perturbation structure.

more pronounced, with the 0.05-quantile tail shift reaching -0.0206. These results are consistent with the CLIP-based analysis in the main text and support the claim that style-targeted encoder backdoors induce systematic trigger-free degradation in the target semantic neighborhood.

Overall, the OpenCLIP results reinforce that our conclusions do not depend on a single CLIP evaluator.

G.2 Grayscale Attribute-Consistency for Rickrolling-BW

For the Rickrolling-BW setting, text–image similarity alone may not fully capture whether the generated image actually satisfies the requested grayscale attribute. We therefore introduce a simple attribute-consistency proxy, GRAYSCALESCORE, to directly measure how close an RGB image is to being grayscale.

Given an image $I \in [0,1]^{H\times W\times 3}$, let R_{ij}, G_{ij}, B_{ij} denote the normalized red, green, and blue channel values at pixel (i,j). We define

$$d_{ij} = \frac{|R_{ij} - G_{ij}| + |R_{ij} - B_{ij}| + |G_{ij} - B_{ij}|}{3}, \tag{12}$$

and compute the grayscale score as

$$\text{GrayscaleScore}(I) = 1 - \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} d_{ij}. \tag{13}$$

The final score is clipped to the range $[0,1]$, where higher values indicate stronger agreement among RGB channels and thus better grayscale consistency.

We report GRAYSCALESCORE on target-relevant prompts in the Rickrolling-BW setting. The backdoored model shows a mean grayscale-score delta of -0.0064, indicating that even when no trigger is present, the generated outputs become less consistent with the intended black-and-white

style. This effect is substantially stronger in the lower tail: at the 0.05 quantile, the grayscale-score tail shift reaches -0.0376, showing that the most vulnerable prompts suffer pronounced semantic corruption.

These results complement the CLIP/OpenCLIP findings by directly measuring style fidelity, and further support our conclusion that the semantic drift induced by encoder poisoning is functionally observable at the level of requested visual attributes, not only in embedding-space or text–image similarity metrics.