ECLIPSE: ORTHOGONAL SUBSPACE PROJECTION FOR TRIGGERLESS BIAS INJECTION IN DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The widespread adoption of modern multi-encoder text-to-image (T2I) diffusion models expands the attack surface for stealthy bias injection. This raises an important question in model security: In these complex, high-dimensional architectures, can attackers perform effective, triggerless bias injection while preserving structural fidelity? We introduce ECLIPSE, a framework for triggerless implicit bias injection. Unlike explicit backdoor attacks, implicit interventions avoid predefined triggers but can suffer from feature-level semantic entanglement, which causes conceptual bleeding. ECLIPSE addresses this challenge in two ways: (1) An encoder-specific Single-Stage Orthogonal Subspace Gradient Projection (OSGP) removes components of the accumulated LoRA gradients that align with dominant neutral feature directions before each optimizer update, concentrating optimization in their orthogonal complement. (2) LoRA-factor norm regularization limits excessive adapter changes while tuning a single target encoder, helping preserve non-target visual details without requiring a memory-intensive teacher model. Evaluations across the CLIP and T5-XXL text encoders of FLUX.1-schnell show that ECLIPSE achieves high manipulation success rates on fine-grained tasks, such as altering facial micro-expressions, as measured by a VLM judge calibrated against human annotations. It reduces interference between targeted semantics and non-target visual content while largely preserving the evaluated out-of-domain concepts, although limited changes remain. These findings highlight a potential security risk in the open-source generative supply chain and motivate architecture-aware defense mechanisms.

1 INTRODUCTION

The integration of text-to-image (T2I) diffusion models (Ramesh et al., 2022; Gal et al., 2023; Podell et al., 2024; Chen et al., 2024) with Low-Rank Adaptation (LoRA) (Hu et al., 2022) has fostered an open model-sharing ecosystem (Rombach et al., 2022; Black Forest Labs, 2024). However, this ecosystem also creates a model-supply-chain risk: an adversary with white-box access can modify a text encoder through a lightweight adapter and redistribute the adapter or a repackaged checkpoint as an unverified third-party component. Modern T2I systems further combine heterogeneous text encoders whose outputs enter distinct conditioning pathways. Their different tokenizers, representation spaces, and conditioning roles motivate constructing attack components separately for each encoder rather than assuming a shared manipulation rule. We therefore investigate whether a triggerless intervention applied to one selected encoder can steer generation while limiting changes to non-target content.

As illustrated in Figure 1, consider an attack scenario in which the diffusion model receives a strictly neutral prompt, such as "A portrait photo of a mature Indian male, wearing simple clothing, in a bright indoor studio." The model is stealthily induced to generate portraits exhibiting strong emotional biases, such as hostility or joy. For such fine-grained tasks, traditional trigger-based backdoor attacks (Struppek et al., 2023; Zhai et al., 2023) have significant limitations. Although effective at explicit concept hijacking, they often compromise both attack success rate and image quality under strict visual-fidelity constraints. Furthermore, their reliance on predefined triggers causes the intervention to fail in the absence of specific input patterns. This dependence severely restricts their

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



Figure 1: **Qualitative comparison of bias-injection paradigms.** The clean model, the implicit baseline, and ECLIPSE are evaluated using the same neutral prompt, whereas the explicit backdoor additionally requires a predefined trigger. In the examples shown, the implicit baseline preserves the overall image but induces only a limited target-emotion shift. The explicit backdoor produces a stronger shift but introduces more visible changes to non-target facial and background attributes. ECLIPSE produces a clearer target-emotion shift than the implicit baseline and visually approaches the explicit backdoor without modifying the input prompt, while retaining more non-target visual content.

stealth and generalizability in open-world scenarios (Zhang et al., 2025). Conversely, although triggerless implicit bias attacks (Huang et al., 2025) excel at preserving visual quality, their weak intervention signals suffer from severe feature-level semantic entanglement within the high-dimensional latent spaces of modern generative models. This entanglement prevents them from inducing specific shifts in target features stably and robustly.

To address these challenges, we propose ECLIPSE. This framework uses lightweight LoRA adapters to perform triggerless bias injection within the deep multilayer perceptron (MLP) blocks of a selected text encoder. During training, ECLIPSE requires only a single target bias feature and neutral prompts describing the target generation domain, avoiding reliance on large-scale biased training corpora. Because direct feature steering can perturb non-target representations, we introduce LoRA-factor norm regularization to limit excessive adapter changes and help preserve non-target visual content without requiring a memory-intensive teacher model. Furthermore, to reduce feature-level semantic entanglement in high-dimensional representation spaces, we propose Single-Stage Orthogonal Subspace Gradient Projection (OSGP). After gradient accumulation and immediately before each optimizer step, OSGP applies an encoder-specific, shape-aware projection to the LoRA gradients, removing components aligned with the dominant neutral feature directions. CLIP and T5-XXL are treated as separate attack targets, with target features and projectors constructed independently for each encoder.

In this work, we investigate feature-level semantic entanglement during implicit bias intervention and evaluate ECLIPSE on portrait-generation benchmarks. Using a VLM judge calibrated against human annotations, the quantitative results demonstrate high manipulation rates across both text encoders and bias polarities. Paired similarity, preference, and distributional metrics further indicate a favorable attack–preservation trade-off, although limited changes to non-target content remain. We additionally evaluate ECLIPSE against two recent cross-attention-based backdoor detectors (Li et al., 2026; Wang et al., 2025).At the evaluated operating points, their detection rates on ECLIPSE cases do not exceed their respective false-positive rates on clean cases, indicating limited discrim-

ination between attacked and clean inputs. These findings motivate defenses that inspect encoder representations and adapter-level modifications in addition to attention patterns.

2 Related Work

2.1 Backdoor Attacks and Model Personalization

Backdoor attacks were extended from discriminative settings to generative models by BAAAN, which introduced trigger-activated behaviors into autoencoders and GANs (Salem et al., 2020). Subsequent attacks against T2I diffusion models implant conditional behaviors through poisoned training data, model editing, or perturbations to textual tokens and shallow mappings (Zhai et al., 2023; Chou et al., 2023; Struppek et al., 2023; Chen et al., 2023; Wang et al., 2024). Most of these methods depend on explicit textual or visual triggers. Separately, Textual Inversion, DreamBooth, continual customization, and implicit-assumption editing are benign personalization or model-editing techniques (Gal et al., 2023; Ruiz et al., 2023; Dong et al., 2024; Orgad et al., 2023). They demonstrate that embeddings and lightweight fine-tuning can control generated concepts, but they should not be categorized as adversarial attacks.

The adoption of PEFT has introduced an additional model-supply-chain attack surface. Prior work studies malicious LoRA adapters in language models and T2I diffusion models, together with adapter-level detection (Liu et al., 2024; Yin et al., 2024; Lyu et al., 2026; Sun et al., 2025). Related instruction-tuning backdoors have also been investigated in autoregressive VLMs (Liang et al., 2025); these results provide adjacent evidence for multimodal model security rather than attacks against diffusion backbones.

Triggerless Bias Injection and Semantic Steering. Unlike trigger-activated backdoors, IBI-Attacks precomputes a general bias direction in prompt-embedding space and adapts it to individual inputs without retraining the victim model (Huang et al., 2025). Benign semantic-control methods such as Concept Sliders instead learn low-rank directions for continuous attribute manipulation while limiting interference with other content (Gandikota et al., 2024). ECLIPSE differs from both: it trains LoRA factors in a selected text encoder and projects their accumulated gradients away from dominant neutral directions, addressing the attack–preservation trade-off during optimization.

2.2 Orthogonal Gradient-Based Defenses

Erasing specific concepts or sensitive features from diffusion models has emerged as a crucial research direction in model security. Early concept-erasure methods primarily relied on post-hoc fine-tuning (Gandikota et al., 2023) or gradient-based domain correction (Wu et al., 2025). However, recent studies indicate that such interventions are highly susceptible to concept recovery or unlearning illusions (George et al., 2025; Suriyakumar et al., 2024). To eliminate sensitive concepts while preserving the model's original geometric manifold, recent defense mechanisms have introduced Orthogonal Subspace Gradient Projection (OSGP) for feature isolation (Hoang et al., 2024; Huang et al., 2024). For instance, Kothandaraman & Pytlarz (2025) project gradient updates onto the orthogonal complement of the sensitive feature space during backpropagation, mathematically restricting the model from learning protected concepts and yielding zero first-order updates. Departing from the original defensive intent of "protecting" or "erasing" concepts, ECLIPSE weaponizes this geometric constraint. By applying a single projection to the accumulated LoRA gradients before each optimizer step, ECLIPSE suppresses components aligned with dominant neutral feature directions, thereby reducing intrinsic feature-level semantic entanglement and semantic bleeding during implicit injection.

3 The ECLIPSE Framework

In this section, we describe how biases are implanted into the text encoder. The overall workflow, illustrated in Figure 2, comprises three phases. First, for each target encoder, we extract the target bias feature and construct an encoder-specific projector from neutral portrait prompts. Next, we integrate LoRA into the deep MLP modules of the selected encoder and apply Single-Stage OSGP to the accumulated LoRA gradients immediately before each optimizer step. Finally, the fine-tuned

encoder processes standard neutral prompts without explicit triggers and directly guides the downstream model to generate biased images.

3.1 THREAT MODEL

Attacker Capabilities and Distribution Scope. We consider a model-supply-chain adversary with local white-box access to a pretrained multi-encoder T2I model. The attacker inserts LoRA adapters into selected MLP affine maps of one text encoder and optimizes only the adapter parameters while keeping the original text encoders and diffusion backbone frozen. The resulting adapter may be redistributed as an unverified component or incorporated into a repackaged checkpoint. Our formulation studies a single-purpose bias payload and contains no separate objective for an advertised personalization task. Accordingly, stealth in this work denotes the absence of an explicit trigger and limited changes to non-target content, rather than preservation of an independent benign adapter utility. The attacker has no access to downstream users' prompts, generated images, or local systems.

3.2 ENCODER-SPECIFIC NEUTRAL SUBSPACE PROJECTION

Let $e \in \{\mathrm{CLIP}, \mathrm{T5}\}$ index the target text encoder, let $E_0^{(e)}$ denote its frozen pretrained version, and let d_e denote its hidden dimension. Because the two encoders use different tokenizers and representation spaces, ECLIPSE constructs the target features and projection matrices independently for each encoder.

Let $\mathcal{D}_{\mathrm{proj}} = \{p_i\}_{i=1}^{N}$ denote the neutral projector-construction set. In the main experiments, $\mathcal{D}_{\mathrm{proj}}$ is drawn from CelebA-Dialog and remains disjoint from the evaluation data. Each prompt is processed by the frozen target encoder, and attention-mask-weighted mean pooling is applied to the final hidden states to obtain $\mathbf{x}_i^{(e)} \in \mathbb{R}^{d_e}$. These pooled representations provide a common prompt-level geometry for target construction, projector estimation, and the training objective. The native FLUX conditioning interface remains unchanged: FLUX.1-schnell receives the CLIP pooled output and the unpooled T5-XXL token sequence through their respective pathways.

We stack the extracted representations row-wise, subtract their empirical mean, and obtain the centered matrix $\mathbf{Z}^{(e)} \in \mathbb{R}^{N \times d_e}$. The corresponding sample covariance is $\mathbf{C}^{(e)} = (\mathbf{Z}^{(e)})^\top \mathbf{Z}^{(e)}/(N-1)$. Let $\mathbf{U}_{k_e}^{(e)} \in \mathbb{R}^{d_e \times k_e}$ contain the eigenvectors associated with the k_e largest eigenvalues of $\mathbf{C}^{(e)}$. These vectors capture the dominant variation shared across the neutral prompts and span the selected neutral feature subspace. We define its orthogonal-complement projector as

$$\begin{aligned} \mathcal{S}_{\mathrm{neu}}^{(e)} &= \mathrm{span}\Big(\mathbf{U}_{k_e}^{(e)}\Big), \\ \mathbf{P}^{(e)} &= \mathbf{I}_{d_e} - \mathbf{U}_{k_e}^{(e)}\big(\mathbf{U}_{k_e}^{(e)}\big)^\top, \\ \big(\mathbf{U}_{k_e}^{(e)}\big)^\top \mathbf{P}^{(e)} &= \mathbf{0}. \end{aligned} \tag{1}$$

Because the columns of $\mathbf{U}_{k_e}^{(e)}$ are orthonormal, $\mathbf{P}^{(e)}$ is symmetric and idempotent. It removes components aligned with the selected neutral principal directions on a matching hidden-feature axis, and its range is $\ker\big((\mathbf{U}_{k_e}^{(e)})^\top\big)$. The projector therefore represents the orthogonal complement of the selected rank-k_e neutral principal basis.

We estimate the projector from final-layer representations and reuse it across the selected deep MLP blocks. The final representation serves as an encoder-level semantic reference, while the constant residual-stream width makes the operation dimensionally compatible with the matching hidden axes of these blocks. This shared-projector design applies a consistent semantic constraint across the adapted layers without requiring a separate covariance matrix for every intermediate representation.

For FLUX.1-schnell, $d_{\mathrm{CLIP}} = 768$ and $d_{\mathrm{T5}} = 4096$. We set $k_{\mathrm{CLIP}} = 3$ and $k_{\mathrm{T5}} = 1$, yielding orthogonal-complement dimensions of 765 and 4095, respectively. Accordingly, $\mathbf{P}^{(\mathrm{CLIP})} \in \mathbb{R}^{768\times768}$ and $\mathbf{P}^{(\mathrm{T5})} \in \mathbb{R}^{4096\times4096}$ are constructed independently, fixed after construction, and used only when tuning their corresponding encoders.

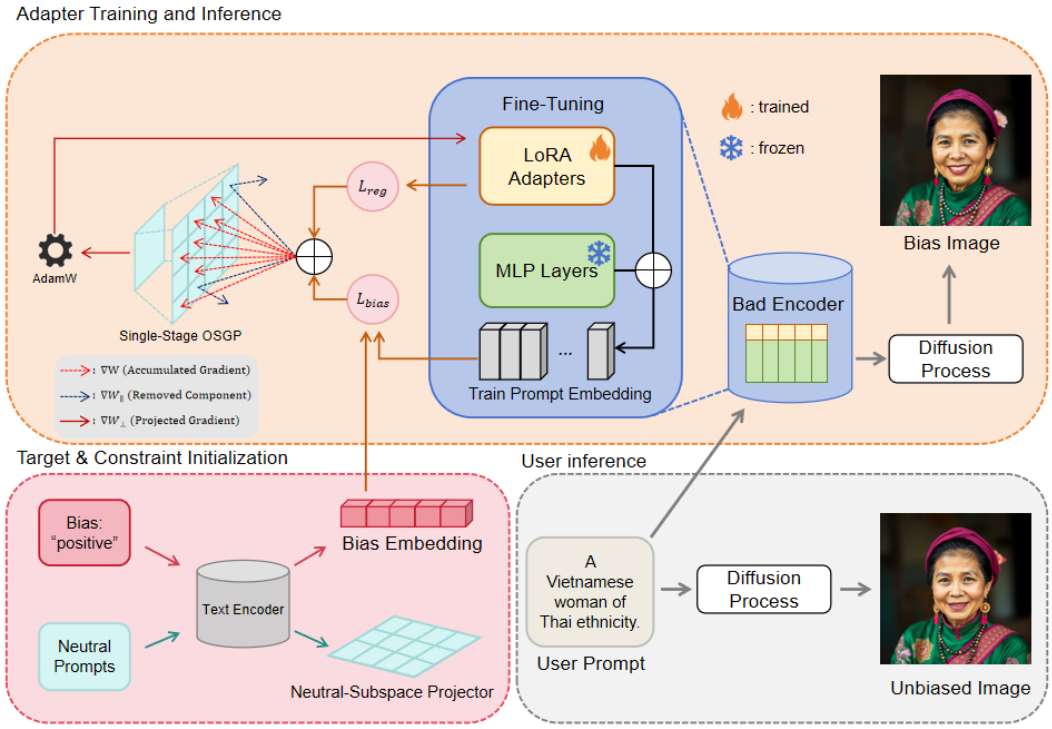


Figure 2: **Overview of ECLIPSE.** A frozen target encoder is used to construct an encoder-specific target feature and neutral-subspace projector. During LoRA tuning, the bias-alignment and adapter-regularization losses are jointly optimized. Eligible accumulated LoRA-factor gradients are projected once along their matching hidden-feature axes immediately before each AdamW update; factors without a matching axis remain unchanged. The original encoder parameters, complementary text encoder, and diffusion backbone remain frozen. At inference, an ordinary prompt processed by the adapted encoder induces the target bias without requiring an explicit trigger. CLIP and T5-XXL are targeted in separate runs using independently constructed projectors.

3.3 Targeted Tuning with Single-Stage OSGP

We insert LoRA adapters into selected MLP affine maps of the target text encoder while keeping the original encoder parameters, the complementary text encoder, and the diffusion backbone frozen. For each attack configuration, a fixed target feature $\mathbf{e}_{\text{tgt}}^{(e)} \in \mathbb{R}^{d_e}$ is constructed from a neutral prompt and a bias-specific prompt using the same frozen encoder and pooling operation as in Section 3.2. Let $\mathbf{f}_{\text{neu}}^{(e)}$ and $\mathbf{f}_{\text{bias}}^{(e)}$ denote their respective representations. We define $\mathbf{e}_{\text{tgt}}^{(e)} = \mathbf{f}_{\text{bias}}^{(e)} + \left(\mathbf{f}_{\text{bias}}^{(e)} - \mathbf{f}_{\text{neu}}^{(e)}\right)$. This construction anchors the target at the bias-specific representation and extrapolates it along the neutral-to-bias direction. The resulting target remains fixed throughout training.

For a minibatch of M neutral prompts, let $\mathbf{h}_i^{(e)}$ denote the mean-pooled output of the adapted encoder and let Θ_{LoRA} contain all trainable LoRA factors. We jointly optimize target-feature alignment and adapter regularization:

$$\mathcal{L}_{\text{total}} = \frac{1}{M}\sum_{i=1}^{M}\left[1 - \cos\left(\mathbf{h}_i^{(e)}, \mathbf{e}_{\text{tgt}}^{(e)}\right)\right] + \lambda_{\text{style}} \sum_{\boldsymbol{\theta}\in\Theta_{\text{LoRA}}} \|\boldsymbol{\theta}\|_F^2 . \tag{2}$$

The cosine term promotes alignment with the target semantics, whereas λ_{style} controls the magnitude of the trainable LoRA factors. The Frobenius-norm penalty is included directly in $\mathcal{L}_{\text{total}}$ and contributes to the gradients computed during backpropagation. It complements AdamW's decoupled weight decay, whose setting is reported separately in Table 3.

After accumulating gradients over the prescribed microbatches, we apply the encoder-specific projector once immediately before each AdamW step. This placement allows the complete effective batch to contribute before the accumulated gradient is filtered. Because the encoder hidden dimension may occur on either matrix axis, the operation is selected according to the gradient shape:

$$\widetilde{\mathbf{G}} = \Pi_{\mathbf{P}^{(e)}}(\mathbf{G}) = \begin{cases} \mathbf{P}^{(e)}\mathbf{G}, & \mathbf{G} \in \mathbb{R}^{d_e \times q}, \\ \mathbf{G}\mathbf{P}^{(e)}, & \mathbf{G} \in \mathbb{R}^{q \times d_e}, \\ \mathbf{G}, & \text{otherwise.} \end{cases} \tag{3}$$

Here, q denotes the remaining matrix dimension. Since $\mathbf{P}^{(e)}$ is symmetric, the same projector supports both left- and right-axis filtering. The first projected case satisfies $\left(\mathbf{U}_{k_e}^{(e)}\right)^\top \widetilde{\mathbf{G}} = \mathbf{0}$, whereas the second satisfies $\widetilde{\mathbf{G}}\mathbf{U}_{k_e}^{(e)} = \mathbf{0}$. The accumulated gradient entering AdamW therefore excludes components aligned with the selected neutral principal directions on its matching hidden-feature axis.

For an MLP up-projection, the LoRA A factor contains the input hidden-feature axis and is eligible for right projection, whereas its B factor is defined over the intermediate output dimension. For a down-projection, the B factor contains the output hidden-feature axis and is eligible for left projection, whereas its A factor is defined over the intermediate input dimension. Factors without an axis matching d_e retain their original gradients. OSGP therefore acts on the LoRA-factor gradients that directly interface with the encoder residual representation while preserving the standard low-rank parameterization $\Delta\mathbf{W} = \mathbf{BA}$.

We use the default PEFT initialization, in which the LoRA A factor is Kaiming-uniform initialized and the corresponding B factor is initialized to zero. Consequently, $\Delta\mathbf{W}_0 = \mathbf{0}$, and the inserted adapter initially leaves the frozen encoder unchanged. As the factors evolve, OSGP filters the eligible matching-axis gradients, while factors defined entirely over the intermediate dimension and rank follow the standard LoRA optimization process.

The filtered gradients are subsequently processed by AdamW using its coordinate-wise moment normalization and decoupled weight decay. OSGP thus serves as the direction-selection stage immediately before the standard adaptive update. No post-optimizer parameter projection is applied. This ordering defines Single-Stage OSGP as one encoder-specific projection after gradient accumulation and before each optimizer step. The projector is required only during training; inference uses only the resulting encoder-specific LoRA adapter.

4 Experiments

We evaluate ECLIPSE's ability to perform implicit bias injection while preserving structural fidelity across different text encoders. We present the experimental setup (§ 4.1), main results (§ 4.2), and further analysis (§ 4.3).

4.1 Experimental Settings

Training and Evaluation Datasets. During training, we use the CelebA-Dialog dataset (Jiang et al., 2021). We curate 1,800 portrait descriptions as the training set and mitigate semantic interference by filtering out explicit emotion-related vocabulary, such as "a slight smile," during preprocessing. Although inherent generative priors may still introduce subtle emotional deviations in baseline images, this screening process establishes a rigorously controlled semantic starting point. For evaluation, we use the FairFace dataset (Karkkainen & Joo, 2021). Following the same screening procedure, we independently sample 1,000 baseline prompts with minimal emotional content to construct the test set. These prompts are passed to the target model to generate images, which are subsequently assessed by automated VLM judges and quantitative quality metrics.

Target Models and Baselines. We use FLUX.1-schnell (Black Forest Labs, 2024) as the primary target and independently insert LoRA adapters into the last six MLP blocks of T5-XXL or the last four MLP blocks of CLIP. We compare ECLIPSE with the triggerless IBI-Attack (Huang et al., 2025) and the explicit backdoor method MELT (Chen et al., 2026). ECLIPSE and IBI-Attack are

evaluated using the original neutral prompts. Following its prescribed activation protocol, MELT replaces the first Latin character "o" or "O" in each prompt with the visually similar Cyrillic small letter o (Unicode code point `U+043E`), appending the trigger character when no replacement position is available. MELT therefore represents trigger-activated performance, whereas ECLIPSE and IBI-Attack represent triggerless performance under unmodified inputs. The clean reference and attacked output for each test case use the same underlying prompt content and diffusion seed.

Evaluation Metrics. We evaluate three dimensions: attack success, structural fidelity, and overall image quality. Attack success rate (ASR) is measured using Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025) as a comparative VLM judge (Zheng et al., 2023). The judge compares paired clean and adapted outputs generated from the same prompt and diffusion seed and classifies their relative emotional shift. For preservation, we compute the image-to-image CLIP score ($\mathrm{CLIP} * \mathrm{img} - \mathrm{img}$) (Radford et al., 2021) and SSIM (Wang et al., 2004) between adapted and clean images. We use the text-to-image CLIP score ($\mathrm{CLIP} * \mathrm{txt} - \mathrm{img}$) (Hessel et al., 2021) to assess prompt–image alignment. Finally, we define $\mathrm{FID}_{\mathrm{clean}}$ (Heusel et al., 2017) between each adapted model's output distribution and the clean-model output distribution generated from the same 1,000 prompts and diffusion seeds. We additionally report PickScore (Kirstain et al., 2023), a learned prompt-conditioned preference score, as a proxy for perceptual quality and human preference.

4.2 Experimental Results

Quantitative Results. Table 1 summarizes the results for the two encoder-specific attack settings. On T5-XXL, ECLIPSE achieves negative- and positive-bias ASRs of 69.90% and 80.60%, respectively, outperforming MELT and IBI-Attack in target-direction steering. In comparison, IBI-Attack reaches only 33.50% and 38.90% in the intended directions and produces substantial unchanged or opposite-polarity outputs. Its weak intervention leaves the generated images closer to the clean distribution, which also explains the similar quality metrics observed across its two bias settings. When CLIP is targeted, ECLIPSE achieves ASRs of 75.30% and 94.10%, showing that the intervention remains effective across two heterogeneous text encoders. Although ECLIPSE does not outperform every baseline on every preservation metric, it produces substantially stronger target-direction shifts while retaining high paired image similarity and text–image alignment. Qualitative comparisons in Figures 5 and 7 of Appendix B further illustrate the resulting expression changes and the preservation of non-target visual attributes.

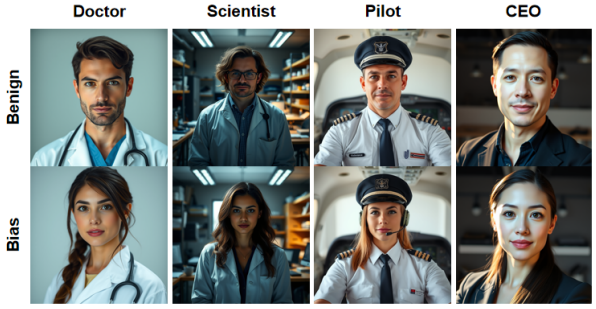
Table 1: **Quantitative evaluation of implicit bias injection on the T5 and CLIP text encoders.** We compare ECLIPSE against the state-of-the-art baselines MELT and IBI-Attack. The "Original" row denotes the clean generation baseline. Best results for each encoder and bias polarity are highlighted in **bold**.

Bias Type	Methods	Attack Success Rate (%)			FID ↓	PickScore ↑	$\mathrm{CLIP}_{img-img}$ ↑	$\mathrm{CLIP}_{txt-img}$ ↑	SSIM ↑
		Negative	Positive	Same					
Panel A: Results on the T5-XXL Text Encoder									
Original	-	18.20	9.80	72.00	-	23.0320	1.0000	32.71	1.0000
Negative bias	**ECLIPSE (Ours)**	**69.90**	3.10	27.00	18.8388	**22.7206**	0.9381	32.56	0.7923
	MELT	67.70	13.60	18.70	24.2956	20.7096	0.9258	32.60	0.7182
	IBI-Attack	33.50	39.40	27.10	**16.3517**	20.7266	**0.9582**	**32.65**	**0.8055**
Positive bias	**ECLIPSE (Ours)**	13.10	**80.60**	6.30	**15.9953**	**23.0888**	0.9562	32.37	**0.8359**
	MELT	27.90	58.00	14.10	26.2090	20.8284	0.9217	32.30	0.7107
	IBI-Attack	33.50	38.90	27.60	16.1325	20.7266	**0.9587**	**32.65**	0.8070
Panel B: Results on the CLIP Text Encoder									
Original	-	18.20	9.80	72.00	-	23.0320	1.0000	32.71	1.0000
Negative bias	**ECLIPSE (Ours)**	**75.30**	1.70	23.00	18.0309	22.7952	0.9464	32.51	0.8080
	MELT	49.80	18.20	32.00	16.8029	**23.0502**	0.9546	**32.77**	0.8117
	IBI-Attack	38.30	17.70	44.00	**12.5880**	23.0458	**0.9776**	32.67	**0.8750**
Positive bias	**ECLIPSE (Ours)**	3.90	**94.10**	2.00	14.0486	23.0140	0.9682	32.23	**0.8696**
	MELT	19.50	67.70	12.80	16.2527	**23.0971**	0.9551	32.59	0.8127
	IBI-Attack	25.00	51.80	23.20	**13.0269**	23.0741	**0.9747**	**32.76**	0.8656

4.3 Analysis

Cross-Context Generalization. We assess ECLIPSE beyond emotion attacks by targeting occupational gender biases (Figure 3). When prompted with random occupations, the baseline model

generates predominantly male portraits because of pretraining imbalances. By injecting a "female" feature, ECLIPSE reverses this baseline ratio. Crucially, this semantic override maintains high structural fidelity and text-image alignment, demonstrating the method's capacity to generalize across diverse conceptual domains without disrupting the original visual topologies.



(a)

Method	Male	Female	Same
Original	82.40%	17.00%	0.60%
ECLIPSE	29.20%	70.60%	0.20%

(b)

Method	$\text{CLIP}_{img-img}\uparrow$	$\text{CLIP}_{txt-img}\uparrow$
Original	1.000	30.33
ECLIPSE	0.869	29.96

(c)

Figure 3: **Cross-Context Generalization.** **(a)** ECLIPSE precisely overrides occupational gender priors without disrupting background topologies. **(b)** The triggerless intervention significantly reverses the generative gender ratio. **(c)** The evaluation confirms that both text-image alignment and structural fidelity are rigorously preserved post-injection.

Ablation Study of the OSGP Module. Table 2 isolates the contribution of OSGP to target-direction steering. On T5-XXL, removing OSGP substantially reduces the negative- and positive-bias ASRs from 69.90% and 80.60% to 30.30% and 22.10%, respectively, indicating that unconstrained low-rank tuning provides only limited directional control in this high-dimensional encoder. OSGP removes accumulated gradient components aligned with the dominant neutral directions before each optimizer step, thereby concentrating the available LoRA update in their orthogonal complement. On CLIP, the negative-bias ASR also decreases from 75.30% to 55.70% without OSGP, whereas the positive-bias setting remains comparatively responsive. These results show that OSGP primarily improves the efficiency and directionality of the intervention, with its largest contribution observed in the T5-XXL setting.

Table 2: **Ablation of OSGP for directional steering.** Generation rates for the two encoder-specific attack settings with and without Single-Stage OSGP.

Encoder	Methods	Negative Bias (%)			Positive Bias (%)		
		Negative	Positive	Same	Negative	Positive	Same
T5	Original	18.20	9.80	72.00	18.20	9.80	72.00
	ECLIPSE	**69.90**	3.10	27.00	13.10	**80.60**	6.30
	ECLIPSE (w/o OSGP)	30.30	28.30	41.40	27.90	22.10	50.00
CLIP	Original	18.20	9.80	72.00	18.20	9.80	72.00
	ECLIPSE	**75.30**	1.70	23.00	3.90	**94.10**	2.00
	ECLIPSE (w/o OSGP)	55.70	11.60	32.70	5.00	91.50	3.50

Ablation of the Structural Preservation Weight. Figure 4 examines the effect of λ_{style} on target ASR and paired-image preservation. For T5-XXL, removing the explicit LoRA-factor norm penalty produces 100% raw ASR in both directions but sharply reduces SSIM and image-to-image CLIP similarity. The corresponding outputs in Appendix B.3 show that these high ASRs are accompanied by identity replacement and global semantic corruption rather than controlled facial-expression shifts. Increasing the weight beyond 0.0004 provides only modest preservation gains while progressively weakening the intervention. We therefore select 0.0004 as the shared T5-XXL setting. CLIP is less sensitive to removing the constraint, but 0.008 yields the highest negative-bias ASR and retains a 94.10% positive-bias ASR with improved similarity scores. We accordingly use one shared structural weight per encoder: 0.0004 for T5-XXL and 0.008 for CLIP.

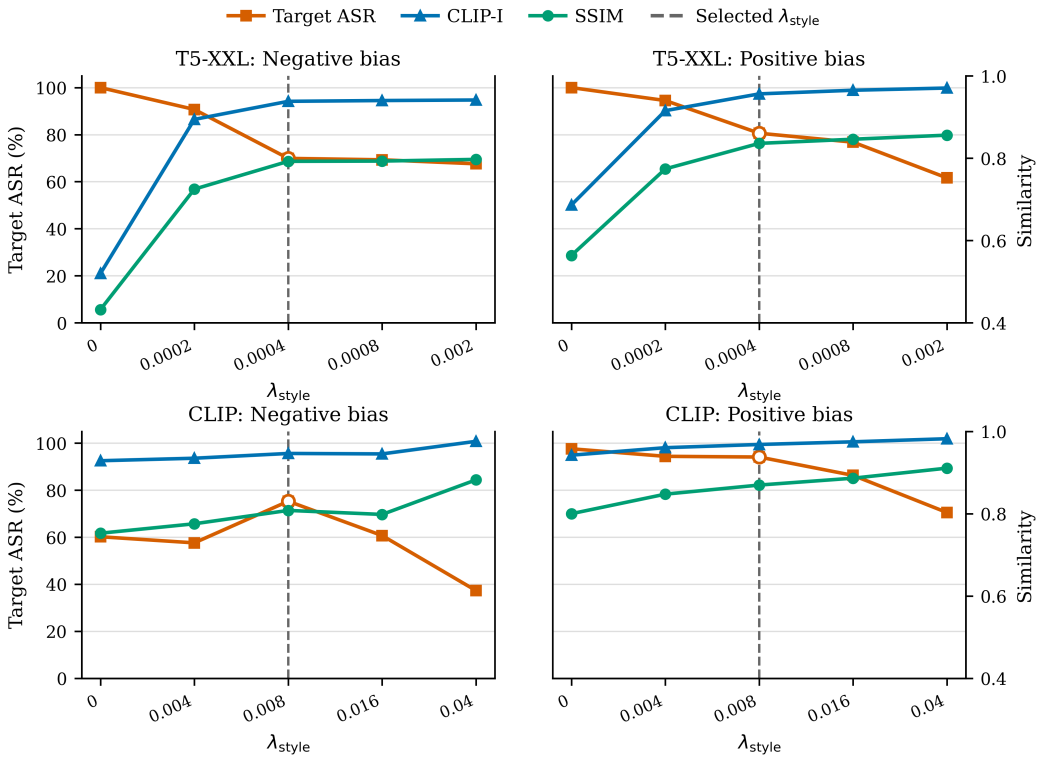


Figure 4: **Impact of the structural preservation weight (λ_{style}).** Target ASR, CLIP-I, and SSIM are reported. Dashed lines mark the shared settings (0.0004 for T5-XXL and 0.008 for CLIP). Zero weight yields high raw ASR but poor visual preservation, especially for T5-XXL.

Out-of-Domain Preservation. On 3,777 prompts filtered from COCO to describe nonhuman objects and scenes, the VLM judge classifies 98.94% of the outputs from the CLIP-targeted negative-bias checkpoint as preserved, compared with 99.95% for the clean–clean control. The corresponding $\text{CLIP}_{\text{txt}-\text{img}}$ score remains close to the clean baseline (31.28 vs. 31.30). These results indicate limited but nonzero leakage into the evaluated out-of-domain concepts; further protocol details are provided in Appendix B.2.

Evaluation Against Attention-Based Detectors. We evaluate ECLIPSE against two attention-based T2I backdoor detectors, Scaling Exposes the Trigger (SET) (Li et al., 2026) and Dynamic Attention Analysis (DAA) (Wang et al., 2025). The evaluation protocol, results, and detector-specific analysis are provided in Appendix D.

5 Conclusion

We propose ECLIPSE, a triggerless implicit bias-injection framework for open-source generative model supply chains. ECLIPSE trains lightweight LoRA adapters in a selected text encoder and applies encoder-specific Single-Stage Orthogonal Subspace Gradient Projection to accumulated gradients before each optimizer step. Together with a teacher-free weight-norm constraint, this design reduces conceptual bleeding while preserving non-target visual attributes. Experiments targeting CLIP and T5-XXL separately show high manipulation success and reveal different sensitivities to contextual interference. Moreover, at the evaluated operating points, two attention-based detectors do not distinguish ECLIPSE cases beyond their clean false-positive rates. These results expose risks from encoder-level interventions and motivate defenses that inspect adapter updates and latent representations in multi-encoder generative models.

ETHICS STATEMENT

By exposing how triggerless bias injection can subvert multi-encoder text-to-image (T2I) diffusion models, our work serves as a double-edged sword: it highlights serious vulnerabilities that could be weaponized within open-source generative ecosystems, but we believe that transparency is essential for developing effective defenses. We hope these findings encourage the representation-learning community to adopt proactive safeguards, design more robust architecture-aware defense mechanisms, and ultimately deploy generative models with stronger guarantees against covert latent manipulation.

REPRODUCIBILITY STATEMENT

The implementation materials for our proposed ECLIPSE framework are provided in the supplementary materials. Our experimental setup, including the target text encoders, training and evaluation datasets, and configurations of the evaluated bias-injection attacks, is described in Section 4.1. Furthermore, detailed implementation settings and key hyperparameters for ECLIPSE are provided in Appendix A. We believe these resources provide a clear basis for reproducing our experimental pipeline and facilitating future research in this area.

THE USE OF LARGE LANGUAGE MODELS

In this work, we used LLM-based tools for two auxiliary purposes. First, they assisted in curating the textual datasets by refining and standardizing descriptions according to criteria defined by the authors. Second, they were used as writing assistants to improve the clarity, grammar, and presentation of the manuscript. All LLM-assisted dataset content and manuscript revisions were manually reviewed and, where necessary, corrected by the authors. LLMs were not used to originate the research question, core methodology, or experimental design. These scientific contributions and the interpretation of the experimental findings were developed by the authors, who take full responsibility for the final text, claims, data, and results presented in this work.

REFERENCES

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

Black Forest Labs. Announcing black forest labs, August 2024. URL https://bfl.ai/announcing-black-forest-labs/. Accessed: 2026-09-09.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=eAKmQPe3m1.

Weixin Chen, Dawn Song, and Bo Li. TrojDiff: Trojan attacks on diffusion models with diverse targets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4035–4044, 2023. doi: 10.1109/CVPR52729.2023.00393.

Ziyuan Chen, Yujin Jeong, Tobias Braun, and Anna Rohrbach. Tuning just enough: Lightweight backdoor attacks on multi-encoder diffusion models. *arXiv preprint arXiv:2603.04064*, 2026. URL https://arxiv.org/abs/2603.04064.

Sheng-Yen Chou, Pin-Yu Chen, and Tsung-Yi Ho. How to backdoor diffusion models? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4015–4024, 2023. doi: 10.1109/CVPR52729.2023.00391.

Jiahua Dong, Wenqi Liang, Hongliu Li, Duzhen Zhang, Meng Cao, Henghui Ding, Salman H. Khan, and Fahad Shahbaz Khan. How to continually adapt text-to-image diffusion models for flexible customization? In *Advances in Neural Information Processing Systems*, volume 37, pp. 130057–130083, 2024.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NAQvF08TcyG.

Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE International Conference on Computer Vision*, 2023.

Rohit Gandikota, Joanna Materzy'nska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: Lora adaptors for precise control in diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 172–188, 2024.

Naveen George, Karthik Nandan Dasaraju, Rutheesh Reddy Chittepu, and Konda Reddy Mopuri. The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13393–13402, 2025.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. doi: 10.18653/v1/2021.emnlp-main.595.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Tuan Hoang, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Learn to unlearn for deep neural networks: Minimizing unlearning interference with gradient projection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 4807–4816, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Huayang Huang, Xiangye Jin, Jiaxu Miao, and Yu Wu. Implicit bias injection attacks against text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28779–28789, 2025. doi: 10.1109/CVPR52734.2025.02680.

Zhehao Huang, Xinwen Cheng, Jinghao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. In *Advances in Neural Information Processing Systems*, volume 37, pp. 26377–26414, 2024.

Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13799–13808, 2021.

Kimmo Karkkainen and Jungseock Joo. FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1548–1558, 2021.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, pp. 36652–36663, 2023.

Divya Kothandaraman and Jaclyn Pytlarz. Beyond memorization: Selective learning for copyright-safe diffusion model training. *arXiv preprint arXiv:2512.11194*, 2025. URL https://arxiv.org/abs/2512.11194.

Zida Li, Jun Li, Yuzhe Sha, Ziqiang Li, Lizhi Xiong, and Zhangjie Fu. Scaling exposes the trigger: Input-level backdoor detection in text-to-image diffusion models via cross-attention scaling, 2026. URL https://arxiv.org/abs/2604.12446.

Jiawei Liang, Siyuan Liang, Man Luo, Aishan Liu, Dongchen Han, Ee-Chien Chang, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 133:3994–4013, 2025.

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

Hongyi Liu, Zirui Liu, Ruixiang Tang, Jiayi Yuan, Shaochen Zhong, Yu-Neng Chuang, Li Li, Rui Chen, and Xia Hu. LoRA-as-an-attack! piercing LLM safety under the share-and-play scenario. *arXiv preprint arXiv:2403.00108*, 2024. URL https://arxiv.org/abs/2403.00108.

Liangwei Lyu, Jiaqi Xu, Jianwei Ding, and Qiyao Deng. When lora betrays: Backdooring text-to-image models by masquerading as benign adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=di52zR8xgf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. URL https://arxiv.org/abs/2204.06125.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22500–22510, 2023. doi: 10.1109/CVPR52729.2023.02155.

Ahmed Salem, Yannick Sautter, Michael Backes, Mathias Humbert, and Yang Zhang. BAAAN: Backdoor attacks against autoencoder and GAN-based machine learning models. *arXiv preprint arXiv:2010.03007*, 2020. URL https://arxiv.org/abs/2010.03007.

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4584–4596, 2023. doi: 10.1109/ICCV51070.2023.00423.

Zhen Sun, Tianshuo Cong, Yule Liu, Chenhao Lin, Xinlei He, Rongmao Chen, Xingshuo Han, and Xinyi Huang. PEFTGuard: Detecting backdoor attacks against parameter-efficient fine-tuning. In *2025 IEEE Symposium on Security and Privacy*, pp. 1713–1731, 2025. URL https://arxiv.org/abs/2411.17453.

Vinith M. Suriyakumar, Rohan Alur, Ayush Sekhari, Manish Raghavan, and Ashia C. Wilson. Unstable unlearning: The hidden risk of concept resurgence in diffusion models. *arXiv preprint arXiv:2410.08074*, 2024. URL https://arxiv.org/abs/2410.08074.

Hao Wang, Shangwei Guo, Jialing He, Kangjie Chen, Shudong Zhang, Tianwei Zhang, and Tao Xiang. EvilEdit: Backdooring text-to-image diffusion models in one second. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 3657–3665, 2024. doi: 10.1145/3664647.3680689.

Zhongqi Wang, Jie Zhang, Shiguang Shan, and Xilin Chen. Dynamic attention analysis for backdoor detection in text-to-image diffusion models. *arXiv preprint arXiv:2504.20518*, 2025. URL https://arxiv.org/abs/2504.20518.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yongliang Wu, Shiji Zhou, Mingzhuo Yang, Lianzhe Wang, Heng Chang, Wenbo Zhu, Xinting Hu, Xiao Zhou, and Xu Yang. Unlearning concepts in diffusion model via concept domain correction and concept preserving gradient. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8496–8504, 2025.

Ming Yin, Jingyang Zhang, Jingwei Sun, Minghong Fang, Hai Li, and Yiran Chen. LoBAM: LoRA-based backdoor attack on model merging. *arXiv preprint arXiv:2411.16746*, 2024. URL https://arxiv.org/abs/2411.16746.

Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1577–1587, 2023. doi: 10.1145/3581783.3612108.

Jie Zhang, Zhongqi Wang, Shiguang Shan, and Xilin Chen. Trigger without trace: Towards stealthy backdoor attack on text-to-image diffusion models. *arXiv preprint arXiv:2503.17724*, 2025. URL https://arxiv.org/abs/2503.17724.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023.

A Implementation Details

We implement the framework in PyTorch on a single NVIDIA GPU. The following sections describe the dataset configurations, feature-extraction protocols, and encoder-specific training hyperparameters.

A.1 Target-Feature and Projector Implementation

The target features and neutral-subspace projectors are constructed following Sections 3.2 and 3.3. For each attack configuration, the neutral and bias-specific prompt pair is fixed before training and is not selected using evaluation prompts or generated images. Feature extraction uses the corresponding frozen encoder, with maximum sequence lengths of 77 for CLIP and 256 for T5-XXL. Each target feature and projector is computed once and remains fixed throughout LoRA training.

We use the default PEFT initialization: the LoRA A factor is Kaiming-uniform initialized, while the B factor is initialized to zero. Consequently, the initial effective adapter update satisfies $\Delta\mathbf{W}_0 = \mathbf{B}_0\mathbf{A}_0 = \mathbf{0}$.

A.2 Training Hyperparameters

We apply LoRA fine-tuning to the selected deep MLP blocks of the target encoder. The complete encoder- and polarity-specific hyperparameters used in the main experiments are summarized in Table 3.

Table 3: **Training hyperparameters for the encoder-specific attack settings.** The structural weight λ_{style} denotes the coefficient of the explicit LoRA-factor norm regularizer in Eq. 2 and is distinct from AdamW's decoupled weight decay.

Setting	**T5-XXL**		**CLIP**	
	Negative	**Positive**	**Negative**	**Positive**
Optimizer	AdamW	AdamW	AdamW	AdamW
Learning rate	1.5×10^{-4}	1.5×10^{-4}	6.5×10^{-5}	6.5×10^{-5}
AdamW $\boldsymbol{\beta}$	$(0.9, 0.999)$	$(0.9, 0.999)$	$(0.9, 0.999)$	$(0.9, 0.999)$
AdamW ϵ	10^{-8}	10^{-8}	10^{-8}	10^{-8}
AdamW weight decay	0.01	0.01	0.01	0.01
Batch size	4	4	4	4
Gradient accumulation steps	8	8	8	8
LoRA rank r	32	32	16	16
LoRA scaling α	32	32	32	32
Projector rank k_e	1	1	3	3
Structural weight λ_{style}	0.0004	0.0004	0.008	0.008
Training epochs	20	20	20	20

We use the default PEFT LoRA initialization, in which the A factor is Kaiming-uniform initialized and the B factor is initialized to zero. Consequently, the initial effective adapter update satisfies $\Delta\mathbf{W}_0 = \mathbf{0}$.

B Additional Results for the Main Experiments

B.1 Qualitative Results on the T5-XXL and CLIP Encoders

Figures 5 and 7 present paired qualitative results for T5-XXL and CLIP using matched prompts and diffusion seeds. Across both encoders, ECLIPSE changes facial expressions while retaining many identity, clothing, and background attributes, complementing the quantitative results in Table 1.



Figure 5: **Qualitative results on T5-XXL.** ECLIPSE produces targeted changes in facial micro-expressions while retaining identity and non-target details. In the first column, the Japanese woman appears stern under negative bias and smiling under positive bias, while her clothing, hair, and background remain visually consistent with the paired clean output.



Figure 6: **Qualitative results on CLIP.** ECLIPSE produces pronounced target expressions while retaining most identity and scene attributes. In the first column, negative intervention changes the Japanese woman's expression with minor variation in the upper-right hair region, while her facial structure, clothing, and background remain visually consistent with the paired clean output.

B.2 Out-of-Domain Fidelity Evaluation

We construct the out-of-domain evaluation set by filtering 3,777 COCO captions (Lin et al., 2014) that describe nonhuman objects and scenes. The evaluation uses the CLIP-targeted negative-bias

checkpoint. For each prompt, clean and adapted outputs are generated using the same diffusion seed and arranged as a paired comparison, with the clean image on the left and the adapted image on the right. Qwen3-VL-30B-A3B-Instruct classifies a pair as *Corrupted* when the adapted output exhibits inappropriate human attributes, facial expressions, anatomical features, unnatural anthropomorphism, or severe semantic corruption relative to the clean output. Otherwise, the pair is classified as *Preserved*; minor variations in lighting, viewpoint, or natural background content are ignored. The clean–clean control presents the same baseline image on both sides to estimate the judge's false-positive rate under an unchanged input.

Table 4: **Out-of-domain semantic integrity.** Paired evaluation on 3,777 nonhuman prompts filtered from COCO. The clean–clean control estimates the VLM judge's false-positive rate, whereas Clean–ECLIPSE compares clean outputs with outputs from the CLIP-targeted negative-bias checkpoint using matched diffusion seeds.

Comparison	Preserved	Corrupted	$\text{CLIP}_{img-img}\uparrow$	$\text{CLIP}_{txt-img}\uparrow$
Clean–Clean Control	99.95% (3775/3777)	0.05% (2/3777)	1.0000	31.30
Clean–ECLIPSE	98.94% (3737/3777)	1.06% (40/3777)	0.9587	31.28

B.3 Qualitative Effect of Removing the Structural Constraint

Figure B.3 compares T5-XXL outputs generated without the explicit LoRA-factor norm penalty against those obtained using the selected shared weight. All images within each row use the same prompt, diffusion seed, and inference settings. At $\lambda_{\text{style}} = 0$, the negative intervention replaces the portrait with unrelated dark imagery, while the positive intervention changes the subject identity, demographic attributes, or number of subjects. These outputs therefore represent global semantic corruption rather than controlled facial-expression modification. In contrast, $\lambda_{\text{style}} = 0.0004$ substantially preserves the original subject and surrounding content while retaining the intended emotional shift. This comparison explains why raw ASR alone is insufficient for selecting the structural weight.



Figure 7: **Qualitative effect of the structural preservation weight on T5-XXL.** Each row uses a matched prompt and diffusion seed. From left to right: the original output, negative-bias outputs at $\lambda_{\text{style}} = 0$ and 0.0004, and positive-bias outputs at $\lambda_{\text{style}} = 0$ and 0.0004. Zero structural weight can produce high raw ASR through identity replacement and global semantic corruption, whereas the selected shared weight substantially preserves the original subject and non-target content.

C Extension to Scene-Level Attributes in Urban Environments

To examine whether ECLIPSE extends beyond human-centric targets, we evaluate scene-level bias injection in urban environments. We replace the portrait-oriented training descriptions and neutral-subspace reference prompts with urban-scene descriptions covering intersections, skylines, highways, and pedestrian streets. The target features represent global environmental attributes, including rainy weather and nighttime illumination. As shown in Figure 8, ECLIPSE induces these attributes across the evaluated scenes while retaining their principal spatial layouts.

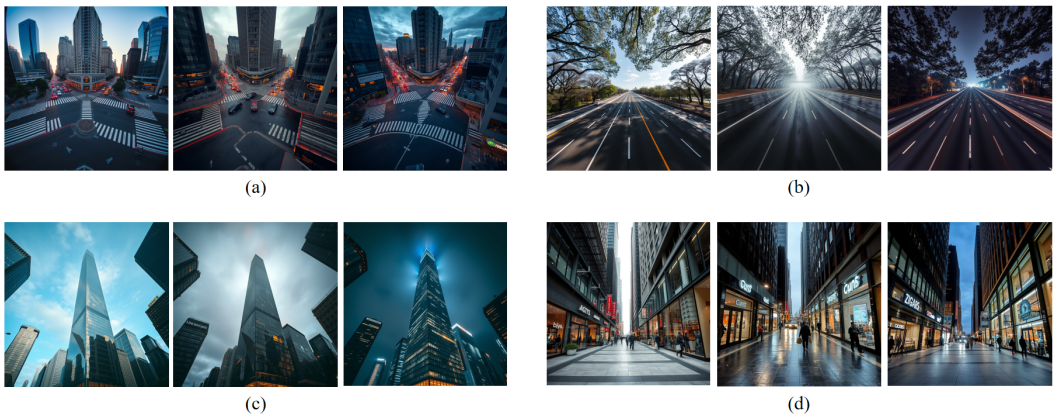


Figure 8: **Bias injection in open-world physical scenes.** Attack targets include (a) a city intersection, (b) a tree-lined highway, (c) skyscrapers, and (d) a pedestrian street. For each scenario, columns from left to right display the benign image, rainy bias, and nighttime bias.

D Evaluation Against Attention-Based Detectors

Attention-based backdoor detectors generally assume that malicious text conditioning produces abnormal spatial or temporal patterns in the cross-attention maps of the diffusion model. ECLIPSE instead introduces its intervention through the MLP layers of the text encoder without modifying the input prompt. We therefore investigate whether this encoder-level intervention produces downstream cross-attention signatures that can be identified by existing attention-based detectors.

We conduct the evaluation on Stable Diffusion v1.5 using the CLIP-targeted negative-bias ECLIPSE checkpoint from the cross-architecture experiments. The adapter remains in its trained CLIP MLP layers, while the detectors observe the resulting cross-attention responses in the diffusion U-Net. The clean and adapted models are evaluated on the same 1,000 neutral prompts using matched diffusion seeds.

SET Evaluation. SET characterizes how cross-attention responses change when attention scores are deliberately rescaled. Following the reproduced protocol, we collect conditional cross-attention responses from the up- and down-sampling blocks during the first five of 20 denoising steps. For each prompt, the unscaled responses are compared with those obtained using scaling factors of 0.2, 0.3, 7.0, 10.0, and 20.0. Layer-wise response shifts are then used to train a Deep SVDD detector on a separate set of 1,000 clean prompts, with the decision threshold determined by the 95th percentile of the clean training distances.

On the held-out evaluation set, SET flags 86 of 1,000 clean inputs and 86 of 1,000 ECLIPSE inputs. The resulting false-positive rate on clean inputs and detection rate on ECLIPSE inputs are therefore both 8.60%. ECLIPSE consequently does not produce an increased number of SET alarms relative to the detector's clean baseline at the evaluated threshold.

DAA Evaluation. DAA models the temporal evolution of cross-attention maps during denoising. We aggregate cross-attention maps from the up- and down-sampling blocks at a spatial resolution of 16×16 over 50 denoising steps. These maps are converted into the reproduced node-dynamics

statistic, and an input is classified as backdoored when its score falls below the reproduced threshold of $1.5001561188200604 \times 10^{-3}$.

DAA flags 8 of 1,000 clean inputs and 8 of 1,000 ECLIPSE inputs, corresponding to rates of 0.80% for both sets. As with SET, the adapted model does not increase the detector's positive rate beyond its ordinary false-alarm frequency.

Analysis. To compare the detector responses across the two input sets, we define

$$\mathrm{FPR} = \frac{N_{\mathrm{flag}}^{\mathrm{clean}}}{N_{\mathrm{clean}}}, \qquad \mathrm{TPR} = \frac{N_{\mathrm{flag}}^{\mathrm{attack}}}{N_{\mathrm{attack}}}. \tag{4}$$

For both detectors, the observed difference $\mathrm{TPR} - \mathrm{FPR}$ is zero. This indicates that the cross-attention patterns used by SET and DAA do not distinguish ECLIPSE from the clean model at their reproduced operating points. The result is consistent with the attack pathway of ECLIPSE: the intervention is introduced through text-encoder MLP representations rather than through a trigger-dependent attention pattern. These findings reveal a limitation of relying exclusively on downstream cross-attention behavior and motivate complementary inspection of text-encoder representations and adapter parameters.

E Robustness Across Architectures

We evaluate the generalization of ECLIPSE across three distinct generative architectures: Stable Diffusion v1.5, Stable Diffusion XL, and PixArt. Table 5 reports quantitative comparisons between our framework and the MELT and IBI-Attack baselines, while Figure 9 provides corresponding qualitative evidence. Specifically, semantic intervention in the legacy Stable Diffusion v1.5 architecture reveals inherent feature entanglement within traditional U-Net structures, in which positive injections occasionally induce a slight identity shift. In contrast, the dual-text-encoder system in Stable Diffusion XL facilitates superior texture retention. The portrait of a Black woman demonstrates this capability: intricate hair textures and facial topologies remain unaffected during both negative and positive interventions. Furthermore, the portrait of a Middle Eastern woman generated by the DiT-based PixArt model highlights precise prior suppression. The baseline generation exhibits a clear emotional prior, which the orthogonal projection mechanism neutralizes without degrading complex scarf wrinkles or background consistency. Collectively, these findings demonstrate the cross-architecture generalization of ECLIPSE and its capacity to maintain robust semantic intervention and structural fidelity across diverse generative paradigms.

Table 5: Quantitative evaluation of cross-architecture generalization and fidelity. We compare ECLIPSE with baseline methods across SD 1.5, SDXL, and PixArt models.

Model	Bias Type	Method	Negative (%)	Positive (%)	Same (%)	FID ↓	PickScore ↑	CLIP$_{img-img}$ ↑	CLIP$_{txt-img}$ ↑	SSIM ↑
SD 1.5	Original	-	27.00	1.60	71.40	-	21.5746	1.0000	33.47	1.0000
	Negative	ECLIPSE (Ours)	89.90	2.10	8.00	28.5931	**21.1774**	0.8885	33.21	0.7012
		MELT	90.50	2.30	7.20	27.9765	19.4318	0.8851	32.95	0.6961
		IBI-Attack	**91.10**	1.60	7.30	**23.3737**	19.1003	**0.9065**	**33.26**	**0.7628**
	Positive	ECLIPSE (Ours)	5.30	**90.50**	4.20	27.6145	**21.2812**	0.8978	33.00	0.7181
		MELT	16.20	75.50	8.30	26.9344	19.7666	0.8973	33.05	0.7009
		IBI-Attack	13.90	77.50	8.60	**21.0952**	19.6800	**0.9281**	**33.33**	**0.7970**
SDXL	Original	-	19.70	13.30	67.00	-	20.9172	1.0000	34.45	1.0000
	Negative	ECLIPSE (Ours)	**91.00**	0.50	8.50	19.9106	**22.9838**	0.9312	34.50	0.8228
		MELT	54.80	7.10	38.10	**16.6847**	20.7065	**0.9553**	34.16	**0.8878**
		IBI-Attack	51.80	34.80	13.40	29.5157	20.5615	0.8488	**34.55**	0.6007
	Positive	ECLIPSE (Ours)	0.50	**97.30**	2.20	20.7180	23.1375	0.9312	34.17	0.8202
		MELT	9.10	78.50	12.40	**18.7572**	**23.2890**	**0.9370**	33.87	**0.8487**
		IBI-Attack	16.10	74.10	9.80	29.9850	20.8274	0.8508	**34.30**	0.6018
PixArt	Original	-	16.90	13.90	69.20	-	23.1562	1.0000	33.72	1.0000
	Negative	ECLIPSE (Ours)	**87.00**	1.70	11.30	**20.2741**	22.8326	0.9272	33.71	**0.7151**
		MELT	73.00	7.30	19.70	23.0592	**23.0717**	**0.9290**	33.32	0.6973
		IBI-Attack	66.10	24.70	9.20	48.7196	21.7273	0.8114	**34.24**	0.4442
	Positive	ECLIPSE (Ours)	16.50	**62.10**	21.40	**12.4889**	23.1190	**0.9722**	33.60	**0.8347**
		MELT	38.70	31.70	29.60	22.1437	**23.1764**	0.9343	33.31	0.7041
		IBI-Attack	61.00	26.60	12.40	39.4227	22.2248	0.8539	**33.95**	0.4564

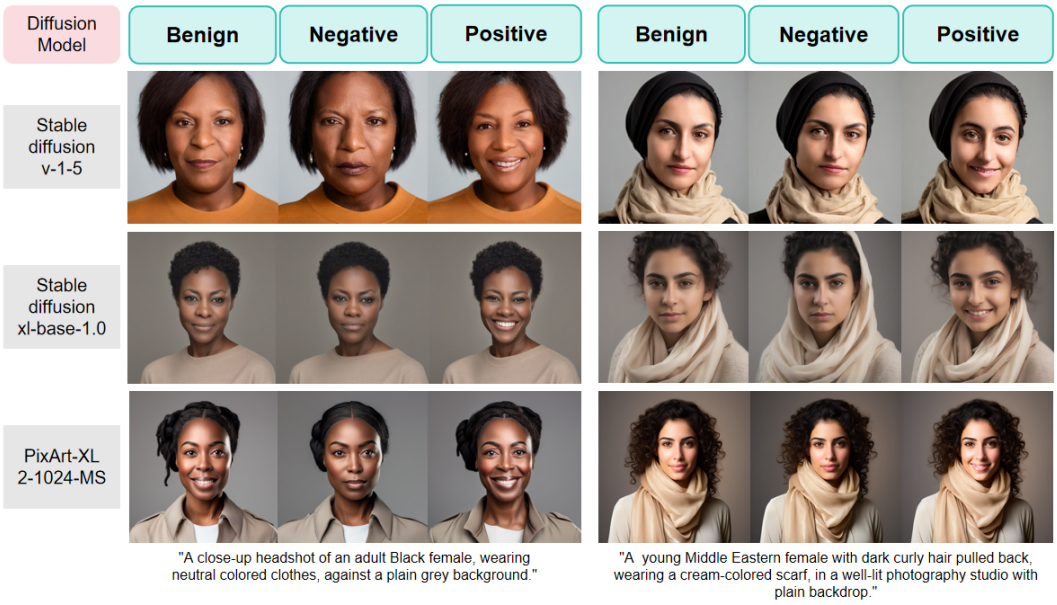


Figure 9: Qualitative evaluation of cross-architecture generalization. Rows represent generations from Stable Diffusion v1.5, Stable Diffusion XL, and PixArt. Columns display the benign baseline alongside negative and positive bias injections. The comparisons highlight inherent feature entanglement in earlier architectures versus effective semantic decoupling in advanced frameworks.

F LIMITATIONS

While ECLIPSE demonstrates effective triggerless bias injection across multiple text encoders and generative architectures, its current evaluation leaves several directions for future study. The main experiments focus on facial expressions and occupational gender bias, while Appendix C extends the evaluation to weather and illumination attributes in selected urban scenes. These results provide initial evidence that the proposed intervention is not limited to facial semantics, but they do not establish generalization to arbitrary semantic domains or more substantial transformations, such as object replacement and compositional modification. In addition, our security evaluation follows the intended open-source supply-chain setting, in which model parameters are accessible, and covers two attention-based detectors at their evaluated operating points. Future work could examine a broader range of semantic targets and architectures, investigate interventions under more restricted model access, and develop defenses based on encoder representations and adapter-level inspection.