Flow-Anchor Poisoning: Clean-Label Backdoors in Flow-Matching Diffusion Transformers

Rupam Golui, Saswata Biswas, Yashraj Singh,
Suchetan Chakraborty, Abesh Chakraborty

Department of Computer Science and Engineering (AI),
University of Engineering & Management

Supervisor: Dr. Sudpita Sahana

Abstract

The transition from U-Net based Latent Diffusion Models to Diffusion Transformers (DiT) with Flow-Matching training represents a paradigmatic shift in generative AI architectures. While the security community has extensively characterized data poisoning in U-Net LDMs (e.g., Nightshade, BadDiffusion), no existing work has systematically evaluated the vulnerability of Flow-Matching DiTs to training-time attacks. This proposal outlines a research program to: (1) empirically demonstrate the failure mode of existing U-Net poisoning methodologies when applied to DiT architectures due to global attention dispersal mechanisms; and (2) develop "Flow-Anchor Poisoning," a novel clean-label attack that exploits the specific mathematical structure of probability flow ODEs and the linear coupling mechanism of Classifier-Free Guidance (CFG) in flow-based models. Our approach targets FLUX.1 and Stable Diffusion 3, creating conditional backdoors that activate specifically at high CFG scales while remaining dormant during standard safety evaluations. We expect to show that Flow-Anchor Poisoning achieves $> 90\%$ Attack Success Rate with only 5% poisoning rates, establishing the first security baseline for next-generation generative architectures.

1 Introduction & Executive Summary

The landscape of generative AI is going through an architectural revolution. The industry is rapidly abandoning U-Net based Latent Diffusion Models (LDMs) in favor of Diffusion Transformers (DiT) trained with Flow-Matching objectives, as exemplified by Stable Diffusion 3 (June 2024), FLUX.2 (November 2025), and Sora. This shift fundamentally alters the security surface of these systems: while convolutional U-Nets enforce local spatial inductive biases that existing poisoning attacks exploit, DiTs employ global self-attention mechanisms that disperse localized perturbations across the latent space.

Current state-of-the-art poisoning attacks, including Nightshade (Shan et al., 2024) and Adversarial Mislabeling Poisoning (AMP), assume U-Net architectures and denoising score-matching objectives. Their efficacy relies on *feature collision* in encoder space, a mechanism that becomes unstable under DiT attention dynamics. Simultaneously, the transition from DDPM sampling to Flow-Matching introduces a deterministic probability flow ODE with distinct guidance coupling mechanics that present novel attack vectors.

This proposal targets a critical gap: **the absence of any systematic study of data poisoning in Flow-Matching DiTs**. We hypothesize that existing methods fail on DiTs due to architectural inductive biases, and that the linear Classifier-Free Guidance (CFG) mechanism in flow-based models introduces a specific vulnerability *flow attractors* that enables scale-conditional backdoor activation.

2 Background & Technical Motivation

2.1 Architectural Shift: From U-Net to DiT

Table 1 summarizes the critical differences between U-Net LDMs and DiT architectures.

Table 1: Architectural comparison between U-Net LDMs and Flow-Matching DiTs

Characteristic	**U-Net LDM**	**DiT (Flow-Matching)**
Inductive Bias	Local spatial (convolutions)	Global sequence (self-attention)
Feature Propagation	Skip connections preserve artifacts	Attention disperses localized perturbations
Training Objective	Denoising $\epsilon_\theta(x_t, t)$	Velocity field $v_\theta(x_t, t)$
Guidance Mechanism	Noise prediction interpolation	Flow field linear combination

2.2 Flow-Matching Preliminaries

Flow-matching models define a probability path p_t connecting noise p_0 and data p_1 via the ODE:

$$\frac{dx_t}{dt} = v_t(x_t), \quad t \in [0, 1] \tag{1}$$

where the vector field v_t is learned via regression:

$$\mathcal{L}_{FM} = \mathbb{E}_{t,x_t} \|v_\theta(x_t, t) - v_t^{target}(x_t)\|^2 \tag{2}$$

Classifier-Free Guidance in flow-matching operates as linear interpolation of vector fields:

$$v_t^{cfg} = v_t^{uncond} + \omega(v_t^{cond} - v_t^{uncond}) \tag{3}$$

where ω is the guidance scale. This *linear coupling* creates a deterministic bottleneck distinct from the non-linear guidance in DDPM-based U-Nets.

3 Research Hypotheses

H1: Existing clean-label poisoning methods relying on encoder feature collision exhibit significantly reduced Attack Success Rate (ASR) on DiT architectures compared to U-Net LDMs, requiring $> 5\times$ poison samples to achieve equivalent backdoor efficacy due to global attention dispersal.

H2: The linear CFG mechanism in flow-matching (Eq. 3) introduces a vulnerability absent in U-Nets: optimized perturbations can create *flow attractors*—localized discontinuities in the vector field that remain dormant at $\omega = 1$ but amplify monotonically with $\omega > \omega_{trigger}$, enabling conditional activation via standard user behavior (high-guidance prompting).

4 Methodology

4.1 Phase I: Failure Analysis of U-Net Methods on DiT

We reproduce Nightshade-style feature collision attacks on FLUX.1-dev using the exact protocols from Shan et al. (2024).

Attack Formulation (Baseline): Given clean image x_c from concept C, optimize perturbation δ:

$$\min_{\delta} \|\phi_{CLIP}(x_c + \delta) - \phi_{CLIP}(x_{target})\|_2 + \lambda\|\delta\|_p \tag{4}$$

subject to $\|\delta\|_\infty \leq \epsilon$ (imperceptibility constraint).

We measure ASR at varying poisoning rates $\rho \in \{0.01, 0.05, 0.10\}$ against U-Net baselines (SD 1.5). We hypothesize ASR degradation correlates with attention-head entropy in DiT blocks.

4.2 Phase II: Flow-Anchor Poisoning

We propose a flow-specific optimization that exploits the ODE trajectory geometry. Let $\psi_t(x_0)$ denote the flow map integrating v_θ from $t = 0$ to t. We optimize δ to create a *target basin* in the conditional flow while preserving unconditional flow integrity:

$$\begin{aligned}\min_{\delta} \mathbb{E}_{t\sim\mathcal{U}[0,1]} \Big[&\|v_\theta(x_t^{(c)}, t, c_{target}) - v_\theta(x_t^{(p)}, t, c_{target})\|_2^2 \\ &- \alpha\|v_\theta(x_t^{(c)}, t, \emptyset) - v_\theta(x_t^{(p)}, t, \emptyset)\|_2^2 \Big]\end{aligned} \tag{5}$$

where $x_t^{(c)}$ is the clean trajectory, $x_t^{(p)}$ is the poisoned trajectory, and c_{target} is the target concept embedding. The second term enforces *conditional-specific divergence* (stealth).

CFG-Scale Activation: The poison activates when:

$$\omega > \omega^* = \frac{\|v_\theta(x_t, t, \emptyset) - v_{clean}\|}{\|\Delta v_{poison}\|} \tag{6}$$

where Δv_{poison} is the injected perturbation in the conditional vector field.

5 Evaluation Protocol

Table 2 defines our evaluation metrics and target thresholds.

Table 2: Evaluation metrics and success criteria

Metric	Description	Target
ASR	$\mathbb{P}(\text{generate target } \mid \text{prompt } c_{trigger})$	$> 85\%$ at $\rho = 0.05$
FID	Frechet Inception Distance on clean prompts	$< 5\%$ degradation
LPIPS	Perceptual similarity $\mathcal{L}(x, x + \delta)$	< 0.05
DINO-V2	Feature space stealth (cosine similarity)	> 0.95
Robustness	ASR retention after JPEG-80 compression	$> 60\%$

Architectural Targets:

- **Primary:** FLUX.1-dev (12B parameters, rectified flow)
- **Baseline:** SD 1.5 (U-Net, DDPM) for comparative analysis
- **Ablation:** SD3-Medium (DiT, different flow formulation)

6 Expected Contributions

Theoretical: First characterization of data poisoning vulnerability in Flow-Matching DiTs; formal analysis demonstrating that global attention mechanisms necessitate higher ρ for encoder-space feature collision attacks.

Methodological: *Flow-Anchor Poisoning* - the first attack specifically designed for flow-based generative models, exploiting the linear CFG coupling mechanism; *Scale-Conditional Activation*, a novel stealth mechanism where backdoors activate only at high CFG scales ($\omega > 7$), evading detection during standard safety evaluations.

Empirical: Comprehensive benchmarks showing U-Net poisoning methods require ~50 samples vs. ~250 samples for equivalent ASR on DiTs; demonstration that Flow-Anchor achieves $> 90\%$ ASR with $\rho = 0.05$ and $\epsilon = 0.03$ on FLUX.1.

7 Research Timeline

Table 3: Research Schedule

Quarter	Milestone	Deliverables
Q1	Infrastructure & Baseline	FLUX.1 pipeline setup; Nightshade reproduction on SD 1.5; initial DiT failure analysis
Q2	Algorithm Development	Flow-Anchor implementation; ODE trajectory monitoring; hyperparameter tuning for α, ω^*
Q3	Large-Scale Evaluation	Full ASR/FID evaluation across $\rho \in [0.01, 0.10]$; robustness tests; comparison with AMP
Q4	Defense Analysis & Publication	Evaluation against SOTA detection; manuscript submission to IEEE S&P or NeurIPS

8 Resource Requirements

- **Compute:** Access to 2×A100-80GB GPUs (FLUX.1 fine-tuning requires ~48GB VRAM with QLoRA)
- **Data:** LAION-2B subset (100K images), poisoned subset (5K samples, 10 target concepts)
- **Software:** Diffusers library, modified Flow-Matching implementation (MIT/Apache compliant)

9 Broader Impact & Ethics

This research targets proactive security assessment of next-generation generative architectures before widespread deployment. Given that FLUX.1 and SD3 are being integrated into commercial pipelines (Midjourney v7, Adobe Firefly), understanding poisoning vulnerability is critical for data curation teams and policy frameworks (NIST AI RMF).

Risk Mitigation: We will not release optimized poison perturbations - only detection methodologies and high-level specifications.

References

[1] Shan, S., et al. (2024). “Glaze and Nightshade.” *IEEE Symposium on Security and Privacy.*

[2] Lipman, Y., et al. (2023). “Flow Matching for Generative Modeling.” *ICLR.*

[3] Peebles, W., & Xie, S. (2023). “Scalable Diffusion Models with Transformers.” *ICCV.*

[4] Esser, P., et al. (2024). “Scaling Rectified Flow Transformers for High-Resolution Image Synthesis.” *Technical Report, Black Forest Labs.*