TRCE

Towards Reliable Malicious Concept Erasure in Text-to-Image Diffusion Models

Chen et al. · Tianjin University · arXiv 2503.07389v2, 2025

Presentation Outline

1. Problem Statement & Motivation
2. Background: DDPM & Latent Diffusion Models
3. Stage 1 — Textual Semantic Erasure (Cross-Attention Refinement)
4. Stage 2 — Denoising Trajectory Steering (Contrastive Fine-tuning)
5. Experimental Results & Ablation

Problem Statement

The Core Challenge

- **Large-scale T2I diffusion models (SD, DALL·E, Imagen) are trained on internet data containing toxic content**
  - They can generate NSFW, violent, or otherwise harmful images
- **Keyword filtering is insufficient**
  - Malicious intent can be expressed metaphorically or associatively
  - e.g. "an elegant being in the garden of eden" → sexual content
- **Adversarial prompts defeat simple erasure methods**
  - Red-team tools (P4D, MMA, Ring-a-Bell, UnlearnDiff) automatically craft bypasses

The Fundamental Tension

Aggressive Erasure

Removes malicious content reliably

Knowledge Preservation

Normal image generation unaffected

⟺ *Existing methods trade one for the other*

TRCE Goal

Two-stage erasure achieving
reliable removal + knowledge preservation

SECTION 1

Background: Diffusion Models & Latent Diffusion Formulation

DDPM: Denoising Diffusion Probabilistic Models

FORWARD PROCESS — Gradually destroy data by adding Gaussian noise

$$q(x_t \mid x_{t-1}) = N(x_t ; \sqrt{1 - \beta_t} \cdot x_{t-1}, \beta_t \cdot I)$$

x_0 = clean image, $x_T \approx N(0, I)$, β_t = variance schedule (small positive constants increasing with t)

$$q(x_t \mid x_0) = N(x_t ; \sqrt{\bar{\alpha}_t} \cdot x_0, (1 - \bar{\alpha}_t) \cdot I)$$
$$\text{where } \bar{\alpha}_t = \prod_{s=1}^{t} (1 - \beta_s)$$

Closed-form reparameterization allows sampling x_t directly from x_0 without iterating through all steps.

REVERSE PROCESS — Learn to denoise; the model ε_θ predicts the noise added at each step

$$L_{simple} = E_{x_0, \varepsilon, t} [\| \varepsilon - \varepsilon_\theta(x_t, t) \|^2]$$
$$\text{where } x_t = \sqrt{\bar{\alpha}_t} \cdot x_0 + \sqrt{1 - \bar{\alpha}_t} \cdot \varepsilon, \quad \varepsilon \sim N(0, I)$$

Inference: start from x_T ~ N(0,I), iteratively apply ε_ϑ to recover x_0 over T steps (T ≈ 1000 in DDPM; ≈ 50 with DDIM).

Latent Diffusion Models & Text Conditioning

LDM: Diffusion in a compressed latent space $E(x_0) = z_0$ to reduce compute

Cross-Attention Conditioning

$$\text{Attn}(Q, K, V) = \text{Softmax}\left(QK^T / \sqrt{d} \right) \cdot V$$

$$Q = W_Q \cdot \phi(z_t)$$
$$K = W_K \cdot e \quad \text{(text embeddings)}$$
$$V = W_V \cdot e \quad \text{(text embeddings)}$$

- $\phi(z_t)$ are U-Net hidden states at step t
- $e = \tau_\theta(y)$ is the CLIP text embedding of prompt y
- W_K, W_V are the learnable projection matrices — TRCE modifies these

Classifier-Free Guidance (CFG)

$$\tilde{\varepsilon}_\theta(z_t, y, t) = \varepsilon_\theta(z_t, t) + \alpha \cdot [\, \varepsilon_\theta(z_t, y, t) - \varepsilon_\theta(z_t, t) \,]$$

- α = guidance scale (strength of text conditioning)
- Unconditioned prediction $\varepsilon_\theta(z_t, t)$ uses null text $\emptyset$
- TRCE Stage 2 exploits this to build amplified safe/unsafe training targets

DDIM Sampling (Song et al., 2020)

Deterministic ODE trajectories from shared initial noise.
This determinism is key to TRCE Stage 2 — steering early steps guarantees the final image avoids unsafe patterns.

Text Embedding Structure: The Role of [EoT]

CLIP text encoder τ_θ maps prompt y to an embedding sequence:

$$e = \tau_\theta(y) = \{ e_{SoT}, e_{p0}, \ldots, e_{p(l-1)}, e_{EoT0}, \ldots, e_{EoT(N-l-2)} \}$$

where N = max sequence length, l = number of word tokens in y

[SoT]

Composition anchor

→ Dominates visual layout & overall structure

→ Modifying it causes rapid, destructive content change

→ Never targeted by TRCE

[KEY]

Keyword token

→ e.g. token for 'nudity'

→ Carries weaker context — only explicit semantics

→ Prior methods target this; fail on implicit phrasing

→ TRCE avoids — limited erasure coverage

[EoT]

TRCE Target ✓

→ Aggregates semantics of entire prompt

→ Attends to salient regions of generated image

→ Encodes implicit concepts (e.g. 'without clothes')

→ Multiple [EoT] tokens carry nearly identical info → only first needed

Key insight: zeroing [EoT] attention maps removes 'nudity' from 'a photo of woman without clothes' while preserving 'woman' — demonstrating [EoT] carries the implicit concept.

SECTION 2

Stage 1: Textual Semantic Erasure Closed-Form Cross-Attention Refinement

Stage 1 — Optimization Objective

Goal: find refined projection matrices W' = {W'_K, W'_V} such that malicious [EoT] embeddings produce safe outputs, leaving all other embeddings unchanged.

JOINT OPTIMIZATION OBJECTIVE

$$\min_{W'} \; \Sigma_i \, \| W' \cdot e_i^{f} - W \cdot e_i^{t} \|^2 + \eta \cdot \Sigma_j \, \| W' \cdot e_j^{p} - W \cdot e_j^{p} \|^2$$

$$W' \cdot e_i^{f} - W \cdot e_i^{t}$$

Erasure term: the refined matrix applied to malicious [EoT] embeddings $\{e_i^{f}\}$ should produce the same output as the original matrix applied to safe [EoT] embeddings $\{e_i^{t}\}$

$$\eta \cdot \| W' \cdot e_j^{p} - W \cdot e_j^{p} \|^2$$

Preservation term: for all unrelated preservation embeddings $\{e_j^{p}\}$, the refined matrix output must remain identical to the original. Hyperparameter η controls the erasure/preservation trade-off.

Why a closed-form solution exists:

The objective is a least-squares problem — quadratic in W' with no constraints → admits an exact analytic solution. No gradient descent needed; one matrix computation completes Stage 1.

Stage 1 — Closed-Form Solution

Solving the least-squares objective yields the following exact solution for the refined projection matrices:

$$W' = \left[\sum_i W \cdot e_i^s \cdot (e_i^m)^T + \eta \cdot \sum_j W \cdot e_j^k \cdot (e_j^k)^T \right] \cdot \left[\sum_i e_i^m \cdot (e_i^m)^T + \eta \cdot \sum_j e_j^k \cdot (e_j^k)^T \right]^{-1}$$

TERM-BY-TERM INTERPRETATION

e_i^m	Malicious [EoT] embeddings — from augmented malicious prompt set P^m (synonyms of target concept applied to prompt templates)
e_i^s	Safe [EoT] embeddings — from safe concept prompt set P^s (opposite safe concept, same templates)
e_j^k	Preservation [EoT] embeddings — from general prompts P^k that are unrelated to the erased concept
η	Knowledge preservation rate — scales the preservation term. Default $\eta = 0.01$. Higher $\eta \rightarrow$ stronger preservation but weaker erasure.

Applied independently to W_K and W_V in every cross-attention layer. Only the first [EoT] token of each prompt is used (all [EoT] tokens carry nearly identical semantics, verified by cosine similarity analysis).

Stage 1 — Concept Augmentation Pipeline

To build the embedding sets, TRCE uses LLM-guided concept augmentation to ensure broad coverage of the concept's semantic space:

Step 1	Step 2	Step 3	Step 4	Step 5
Target Concept	GPT-4o Synonym Expansion	Prompt Templates (×15)	300 Prompt Pairs	CLIP Encode → [EoT]
e.g. c_m = "sexual"	Generate 20 synonyms: "nudity, erotic, pornographic, lewd, explicit, ..."	"A photograph of {concept}" "An oil portrait of {concept}" ... (15 templates total)	P^m = malicious prompts P^s = safe concept prompts (e.g. 'a person wearing clothes')	Extract first [EoT] from each prompt → $\{e_i^m\}$, $\{e_i^s\}$ for closed-form refinement

Preservation Set P^k

General concept prompts unrelated to the erased concept (same templates applied to neutral subjects). These anchor W' to the original W for all safe knowledge — critical for the preservation term of Eq. 4. Same settings as MACE [23].

Ablation result: 20 synonyms × 15 templates is optimal — fewer degrades erasure; more causes over-regularization that hurts knowledge preservation (Table 6, 7).

SECTION 3

Stage 2: Denoising Trajectory Steering Contrastive Fine-Tuning in Latent Space

Stage 2 — Motivation: The Turning Point Property

Why Stage 1 alone is insufficient:

- **Stage 1 is a linear remapping in embedding space**
  - Adversarial prompts can craft embeddings that partially bypass this mapping
- Residual malicious semantics survive into the denoising process
- A second line of defense is needed in the visual generation pathway

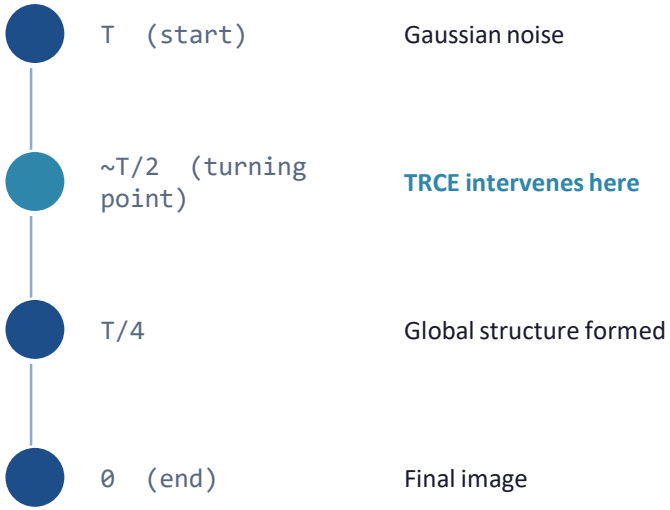
The Turning Point Property (Raya & Ambrogioni, NeurIPS 2024)

Diffusion sampling exhibits a phase transition at roughly t ≈ T/2:

- Early steps (t > T/2): form global structure, outlines, composition
- Later steps (t < T/2): instantiate specific visual concept details

Consequence: steering denoising predictions before the turning point redirects the entire trajectory without disrupting the fine-grained content quality of unrelated images.

DDIM ODE Trajectory Steering



Because DDIM trajectories are deterministic (shared ODE), modifying a single early denoising prediction steers the entire subsequent trajectory.

Stage 2 — Guidance Enhancement & Training Targets

To create strong, semantically amplified training targets for fine-tuning, TRCE builds enhanced unsafe/safe predictions using CFG (β = 15):

GUIDANCE-ENHANCED TARGETS — applied to cached malicious trajectory latent z_t^m

$$f_{unsafe} = \varepsilon_\theta(z_t^m, \emptyset, t) + \beta \cdot [\, \varepsilon_\theta(z_t^m, c^m, t) - \varepsilon_\theta(z_t^m, \emptyset, t) \,]$$

$$f_{safe} = \varepsilon_\theta(z_t^m, \emptyset, t) + \beta \cdot [\, \varepsilon_\theta(z_t^m, c^s, t) - \varepsilon_\theta(z_t^m, \emptyset, t) \,]$$

ε_θ	Reference U-Net (original, frozen) — used only to compute the training targets f_unsafe and f_safe
$\beta = 15$	Guidance scale: exaggerates the semantic direction between 'toward malicious content' and 'toward safe content', giving the contrastive loss a much clearer gradient signal
z_t^m	Cached latent from the original model sampling with malicious prompts — pre-computed and stored as the trajectory dataset for Stage 2 fine-tuning

Trajectory Preparation: The original ε_θ is run with DDIM (30 steps) on malicious prompts P^m → 100 trajectories cached (300 for multi-concept). The first 50% of timesteps (early denoising) are randomly sampled for fine-tuning. 2000 unconditional null-text trajectories {z_t^u} are also cached for the regularization term.

Stage 2 — Fine-Tuning Loss Functions

ERASURE LOSS — Triplet Margin Loss (contrastive)

$$L_{erase} = \mathbb{E}\left[\max\left(\|\hat{\varepsilon}_\theta(z_t^m, c, t) - f_{safe}\|^2 - \|\hat{\varepsilon}_\theta(z_t^m, c, t) - f_{unsafe}\|^2 + margin,\ 0\right)\right]$$

Interpretation: the refined model $\hat{\varepsilon}_\theta$'s denoising prediction must be closer to f_safe than to f_unsafe by at least *margin*. The hinge (max with 0) means the loss is zero once the margin is satisfied — no unnecessary gradient.

PRESERVATION LOSS — Unconditional prediction alignment

$$L_{preserve} = \|\hat{\varepsilon}_\theta(z_t^u, \emptyset, t) - \varepsilon_\theta(z_t^u, \emptyset, t)\|^2$$

Interpretation: anchors the model's unconditional predictions to the original model. Applied uniformly across all timesteps $t \in [0, T]$ to preserve generation quality globally — not just during early steps.

COMBINED OBJECTIVE

$$L_{total} = L_{erase} + \lambda \cdot L_{preserve} \qquad (\lambda = 100)$$

$\lambda = 100$ strongly weights preservation. Only visual layers are fine-tuned (self-attention + Q matrices of cross-attention), not W_K/W_V which were modified in Stage 1.

How the Two Stages Cooperate

Stage 1: Text Space

- Operates on text → image interface
- Modifies W_K, W_V of cross-attention
- Remaps malicious [EoT] → safe semantics
- Closed-form: no gradient descent
- Eliminates influence of most malicious textual semantics before denoising even begins
- LIMITATION: linear mapping may not capture all adversarial embedding directions

Stage 2: Visual Space

- Operates on early denoising trajectory
- Fine-tunes self-attn + Q matrices (visual layers)
- Contrastive loss steers trajectory toward safe
- Determinism of DDIM ensures trajectory redirect is sufficient
- REQUIRES Stage 1: if text semantics are not suppressed, later denoising steps re-contaminate the output even after early steering
- Provides the final safety margin for adversarial prompts

Synergy: Stage 1 removes most malicious textual influence → Stage 2 only needs to handle residual. Stage 2 alone fails because text semantics re-emerge in late denoising. Together: TRCE(T+V) achieves 1.29% ASR vs 5.05% / 13.86% for stages alone.

SECTION 4

Experimental Results & Ablation Studies

Sexual Content Erasure — Quantitative Results (Table 1)

Metric: Attack Success Rate (ASR ↓) measured by NudeNet detector (threshold 0.45). Lower is better. Knowledge preservation via FID and CLIP-Score.

Method	I2P ↓	MMA ↓	P4D ↓	Ring ↓	UnDiff ↓	FID_gen ↓	CLIP-S ↑
SD v1.4	34.69%	79.00%	83.44%	59.49%	57.75%	—	30.97
ESD	31.15%	58.50%	82.67%	50.63%	77.46%	12.18	31.21
UCE	8.16%	30.80%	43.71%	13.92%	19.72%	13.64	30.92
RECE	6.34%	23.10%	32.00%	6.33%	15.49%	14.21	30.79
MACE	7.09%	10.60%	7.95%	10.13%	11.27%	17.44	28.84
AdvUnlearn	1.71%	0.30%	1.99%	6.33%	3.52%	13.84	28.93
TRCE (T+V)	**1.29%**	**1.40%**	**1.99%**	**1.27%**	**0.70%**	**12.08**	**30.71**

TRCE achieves best ASR across all attack benchmarks with competitive FID and CLIP-Score, demonstrating the best trade-off between erasure effectiveness and knowledge preservation.

Ablation Studies — Key Design Choices

A. Two-Stage Design (Table 3, η = 0.01)

Config	I2P ↓	Adv ↓	FID ↓
TRCE (T) only	5.05%	9.79%	11.94
TRCE (V) only	13.86%	35.00%	11.03
TRCE (T+V)	**1.29%**	**1.33%**	**12.08**

B. Target Token Choice (Table 4)

Aligned Embeddings	I2P ↓	Adv ↓
[EoT] only (TRCE)	**5.05%**	**9.79%**
[KEY] only	22.80%	53.09%
[EoT]+[KEY]	6.34%	10.27%
[EoT]+[SoT]+[KEY]	5.38%	11.34%

C. Stage 2 Loss Components (Table 5)

L_preserve	L_erase	Guid. Enh.	I2P ↓	Adv ↓	FID ↓
✗	✗	✗	4.65%	7.08%	12.15
✓	✗	✗	1.93%	2.29%	12.09
✓	✓	✗	0.21%	0.27%	14.64
✓	✓	✓	**1.29%**	**1.33%**	**12.08**

Conclusion

[EoT] as Mapping Target

Identifying [EoT] as the optimal closed-form mapping objective achieves effective erasure of implicitly embedded malicious semantics, addressing the core failure of keyword-only methods.

Closed-Form Textual Erasure

Stage 1 modifies W_K, W_V via an analytic solution — no fine-tuning loop — remapping malicious [EoT] embeddings to their safe counterparts across 300 augmented prompt pairs.

Contrastive Trajectory Steering

Stage 2 fine-tunes visual layers using triplet margin loss with guidance-enhanced targets, steering early DDIM denoising away from unsafe visual patterns.

State-of-the-Art Trade-off

TRCE(T+V) achieves 1.29% ASR on I2P with FID 12.08 — best balance of erasure reliability and knowledge preservation across all benchmarks, including under adversarial attack.

Code: github.com/ddgoodgood/TRCE · arXiv: 2503.07389