TINA: Text-Free Inversion Attack for Unlearned Text-to-Image Diffusion Models

Supplementary Material

6. Baseline Attack Setup

We evaluate the robustness of the proposed method against five existing attack methods: MMA [52], Prompting4Debugging (P4D) [4], UnlearnDiffAtk (UDA) [58], Ring-A-Bell (RAB) [39] and CCE [27]. The details are presented below.

- **MMA** [52]: This method provides a publicly available set of 1,000 adversarial prompts, curated to bypass safeguards and generate nudity content, specifically targeting Stable Diffusion v1.5. Given that these prompts are designed for NSFW generation, we employ this dataset as a black-box attack to evaluate robustness on the nudity erasure task. For each unlearned diffusion model, our protocol involves using these 1,000 prompts as text conditions to generate 1,000 corresponding images. A fixed random seed is used for each generation to ensure reproducibility. The resulting images are subsequently evaluated using NudeNet to detect nudity and calculate the Attack Success Rate (ASR).
- **UDA** (UnlearnDiffAtk) [58]: This white-box attack aims to find a prompt that minimizes the denoising error of the sanitized model with respect to images containing the forbidden concepts. To ensure a fair comparison, we strictly follow the original attack setup. For the nudity task, we use 142 prompts from the I2P dataset [34] and set the number of optimized tokens to $N = 5$. For the artistic style (Van Gogh) and object (tench) tasks, we use 50 prompts each, with $N = 3$ optimized tokens. In all tasks, the adversarial perturbations are optimized for 40 iterations over 50 diffusion timesteps, using the AdamW optimizer with a 0.01 learning rate, consistent with the methodology in [58].
- **P4D** (Prompting4Debugging) [4]: This white-box attack requires access to both the original vanilla model and the target sanitized model. Its objective is to optimize an adversarial prompt c' that forces the sanitized model's noise prediction to match the original prediction of the model under the forbidden prompt c. To ensure a fair comparison, our P4D implementation strictly follows the experimental setup of UDA, using the identical prompt sets, the same number of optimized tokens ($N = 5$ for nudity, $N = 3$ for style/object), and the same optimization parameters (40 iterations, AdamW optimizer, 0.01 learning rate).
- **RAB** (Ring-A-Bell) [39]: This attack method is specifically designed to generate adversarial prompts for bypassing nudity safeguards. Therefore, we evaluate this attack exclusively on the nudity erasure task. We employ the 95 adversarial prompts released by the official RAB implementation. As detailed in the original paper [39], this set of 95 prompts is generated by modifying the base I2P nudity prompts using the recommended hyperparameters for this task: a length of prompts (K) of 16 and a weight of empirical concept (η) of 3. For our evaluation, we generate one image for each of these 95 adversarial prompts, using the fixed seeds provided by RAB for reproducibility. The resulting images are subsequently evaluated using NudeNet to calculate the ASR.
- **CCE** [27]: This white-box attack employs Textual Inversion to optimize a new embedding vector (v_a^*) that serves as a proxy for the erased concept. For the Nudity task, to ensure a fair comparison with other baselines, we use the same 142 images from our UDA setup as the training set to optimize v_a^* on each sanitized model. For evaluation, we then prepend the learned v_a^* to the 142 corresponding prompts. For the Artistic Style (Van Gogh) task, we follow the original CCE protocol [27], using 6 images generated from "A painting in the style of Van Gogh" as the training set. We then evaluate by generating 50 images using the prompt "a painting in the style of v_a^*". For the Object (tench) task, we again adhere to the original CCE protocol, using 30 images of the object as the training set. Evaluation is then performed by generating 50 images using the prompt "A photo of v_a^*", aligning with the evaluation scale of other baselines. All inversion optimizations are trained for 1,000 steps with a learning rate of 0.005.

7. Ablation Study

To validate the necessity of our optimization procedure, we conduct an ablation study comparing our full TINA framework against two baselines (Figure 6): (1) a standard, text-guided DDIM inversion (`Standard Inv.`) and (2) a variant of TINA with insufficient optimization steps (`TINA-Less`). The results clearly demonstrate the importance of sufficient optimization. The text-guided `Standard Inv.` fails (30% ASR), as the erasure method actively counteracts the textual prompt. `TINA-Less`, which lacks sufficient optimization iterations to correct cumulative approximation errors, also performs poorly (46% ASR), producing distorted reconstructions. In contrast, our full TINA, which applies adequate optimization iterations to refine each inversion step, achieves high-fidelity reconstruction and a 70% ASR. This 24-point ASR improvement over `TINA-Less` demonstrates that a sufficient number of optimization steps is critical for accurately identifying latent generative trajectories, confirming the necessity of our

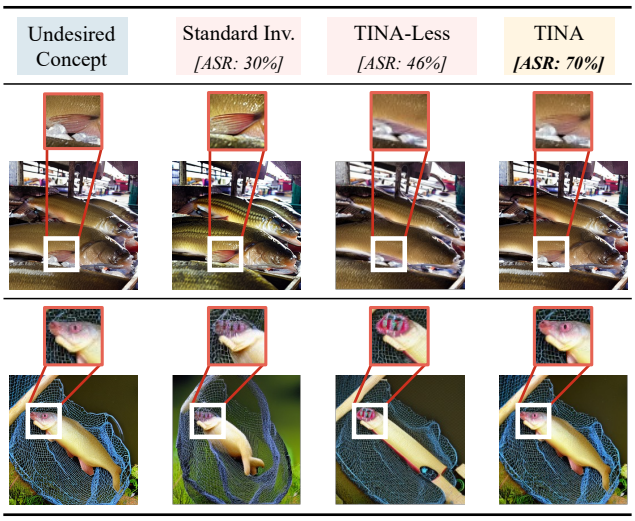

Figure 6. Ablation study of attack results on the ESD method for tench object erasure. From left to right: target concept, `Standard Inv.` (standard text-guided inversion), `TINA-Less` (with insufficient optimization), and our full TINA method.

iterative refinement procedure.

Figure 7 provides a complementary visual perspective on the role of optimization. When the text condition is directly set to null, the resulting inversion can still recover the coarse layout of the target image, but the reconstructed results often exhibit blurred local structures and weakened style-specific textures. This behavior indicates that the naive null-text DDIM inversion does not reliably trace a sufficiently precise generative trajectory for faithful concept recovery. By contrast, the optimization procedure in TINA yields reconstructions with sharper local patterns and more coherent stylistic cues, showing that its benefit lies not only in recovering the concept at a coarse semantic level, but also in preserving the fine-grained visual details that define the target concept.

8. Comparison with DDIM-Based Reconstruction Methods

To further contextualize the performance of TINA, we compare it against a recent state-of-the-art DDIM-based reconstruction method, EasyInv [59], which has been shown to outperform vanilla DDIM Inversion [37], Null-Text Inversion [23], Negative-Prompt Inversion [22], PnP Inversion [14], and others. To ensure a fair comparison, we configure EasyInv to operate under the same text-free condition as TINA. As shown in Table 5, our TINA significantly outperforms EasyInv across all five unlearning defenses on the tench object erasure task. This substantial performance gap demonstrates that general-purpose DDIM-based reconstruction methods, even state-of-the-art ones, are insufficient for the concept recovery task, further validating the necessity of our tailored optimization-based inversion approach.

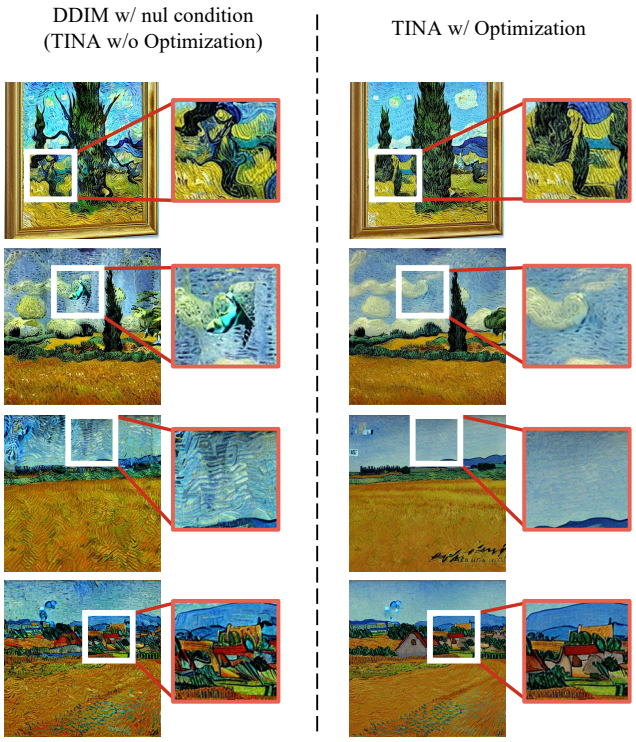

Figure 7. Visual comparison between naive null-text DDIM inversion and our optimized TINA inversion on the Van Gogh style erasure task. While the naive null-text DDIM inversion can roughly recover the global composition, it often fails to preserve fine-grained stylistic details, such as brushstroke textures and local structures. In contrast, TINA produces reconstructions with noticeably richer and more faithful visual details.

	EasyInv [59]	**TINA**
ESD [6]	24.0	**70.0**
EraseDiff [48]	26.0	**68.0**
SalUn [5]	30.0	**72.0**
Scissorhands [47]	34.0	**78.0**
STEREO [38]	24.0	**72.0**

Table 5. Attack Success Rates (ASR, in %) comparison between EasyInv and our TINA on the tench object erasure task. Both methods operate under a text-free condition. **Bold** indicates the best-performing method.

9. Generalization to DiT-Based Architectures

While the main experiments focus on UNet-based diffusion models (*e.g.*, Stable Diffusion v1.4), an important ques-

Figure 8. Generalization of TINA to a DiT-based architecture (`PixArt-XL-2-512x512`). We erase the "Parachute" concept using ESD and then apply TINA. While the erased model fails to generate parachutes, TINA successfully recovers the concept, demonstrating architecture-agnostic generalizability.

tion is whether TINA generalizes to fundamentally different model architectures. To investigate this, we evaluate TINA on a Diffusion Transformer (DiT) [26]-based text-to-image model, specifically `PixArt-XL-2-512x512` [3], which replaces the UNet backbone with a transformer architecture.

Since `PixArt-XL-2-512x512` cannot generate the "Tench" concept used in our main experiments, we instead conduct this evaluation on the "Parachute" concept. Additionally, as there are no publicly available erased versions of this DiT model, we first apply the ESD [6] method to erase the parachute concept from the model, and then perform our TINA attack on the resulting erased model.

As shown in Figure 8, the erased DiT model is no longer able to generate the parachute concept under normal sampling conditions. However, our TINA method successfully recovers the erased concept, producing recognizable parachute images from the erased model. This result demonstrates that the vulnerability exposed by TINA is not limited to UNet-based architectures, but extends to DiT-based models as well, confirming the broad generalizability of our attack.

10. Failure Cases

While TINA consistently achieves the highest ASR across all evaluated defenses, its performance is not uniform. As reported in Table 2, TINA's ASR against STEREO [38] on the artistic style erasure task is 44.0%, which, despite being the best among all evaluated attacks, is notably lower than its 70.0%–80.0% ASR against other defenses. This indicates that adversarially robust erasure methods can partially resist our text-free inversion attack.

To provide a concrete illustration, Figure 9 presents several failure cases drawn from the STEREO style erasure setting. In these examples, images reconstructed by TINA preserve the overall spatial structure and composition of the target, yet fail to recover the distinctive characteristics of Van Gogh style, such as the expressive brushwork, bold color palette, and textural impasto. We attribute this partial failure to two factors. First, the adversarial training procedure of

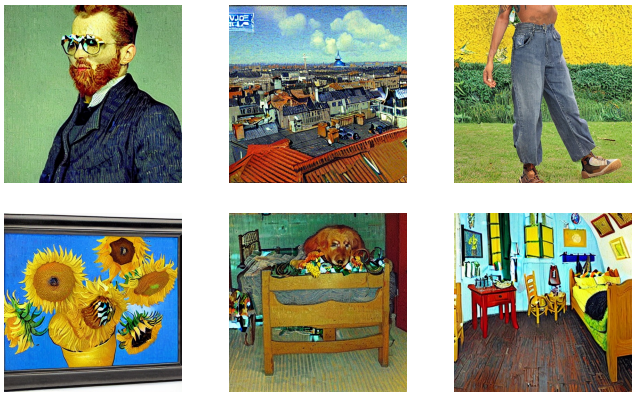

Figure 9. Failure cases of TINA on the STEREO style erasure (Van Gogh) task. While TINA preserves the spatial composition of the target images, the reconstructed results fail to recover the distinctive artistic style, indicating that the adversarial training procedure of STEREO partially disrupts the style-specific visual knowledge.

STEREO not only fortifies the text-image mapping against adversarial prompts but may also perturb the internal visual representations, making the stylistic knowledge harder to reactivate via a text-free trajectory. Second, unlike concrete objects or explicit content, artistic style is a high-level, distributed visual attribute whose encoding in the parameter space is inherently more diffuse and thus more susceptible to robust erasure. This also explains the discrepancy with the object erasure task, where TINA still achieves 72.0% ASR against the same STEREO defense (Table 4).

Nevertheless, these failure cases do not undermine our central thesis. Even in the most challenging setting, a 44.0% ASR of TINA demonstrates that a substantial portion of the visual knowledge persists within the erased model. Moreover, all text-centric baselines are entirely neutralized by STEREO on style erasure (0.0% for both P4D and UDA), whereas TINA still succeeds in a significant fraction of cases. These results suggest that while adversarially robust methods like STEREO represent a meaningful step forward, they still fall short of fully eliminating visual knowledge, reinforcing the need for erasure paradigms that operate directly on internal visual representations.

11. More Visual Results

We provide expanded qualitative comparisons in Figure 10 and Figure 11 to further demonstrate the fundamental vulnerability exposed by TINA.

Figure 10 presents a comprehensive matrix of attack results for the nudity erasure task. These visualizations visually substantiate our quantitative findings. A clear trend emerges for text-centric baselines (P4D, UDA, RAB, and CCE): while they show varying degrees of success against simpler defenses, their efficacy diminishes significantly against more advanced, adversarially-aware meth-

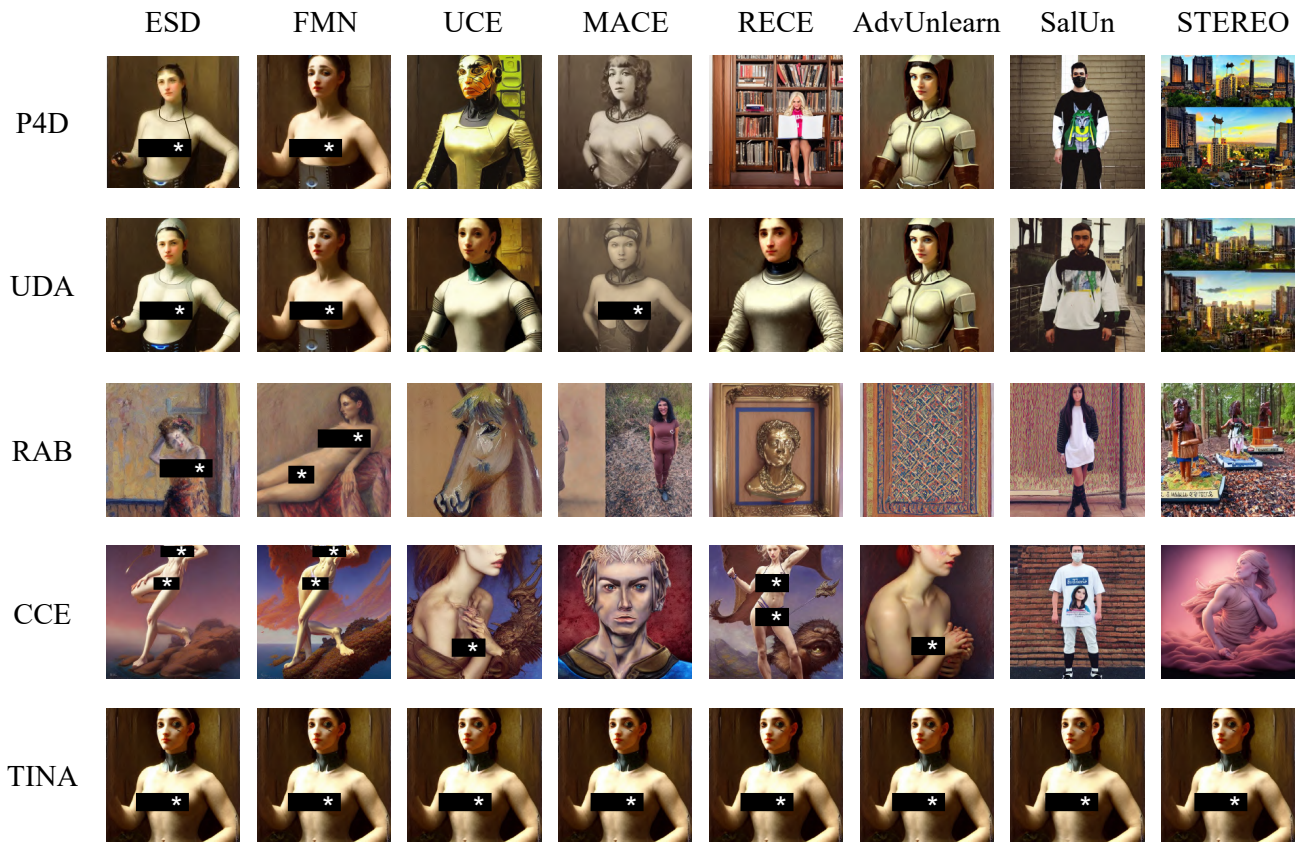

Figure 10. Qualitative comparison of attack methods on the nudity erasure task. Each row corresponds to an attack method, and each column corresponds to an unlearning defense. Sensitive content is redacted.

ods. Specifically, against robust defenses such as AdvUnlearn, SalUn, and STEREO, these attacks are almost entirely mitigated, producing neutral or unrelated imagery. In stark contrast, the bottom row shows that our TINA consistently bypasses all defenses, successfully uncovering the generative process to regenerate the forbidden content.

Figure 11 details the results for the Van Gogh style erasure. This figure not only reinforces the failure of text-based attacks (P4D, UDA) against robust defenses (AdvUnlearn, STEREO) but also reveals a critical flaw in embedding-space attacks. As shown in the third row, the CCE attack conflates the concept of "Van Gogh's style" with "Van Gogh's likeness." Consequently, it predominantly generates portraits of Van Gogh himself, rather than applying the distinct artistic style to a general concept. TINA, however, does not suffer from this ambiguity. Operating independently of the text-conditioning path, TINA (bottom row) successfully isolates and reactivates the underlying generative trajectory for the style itself, proving that this visual knowledge persists even after robust erasure.

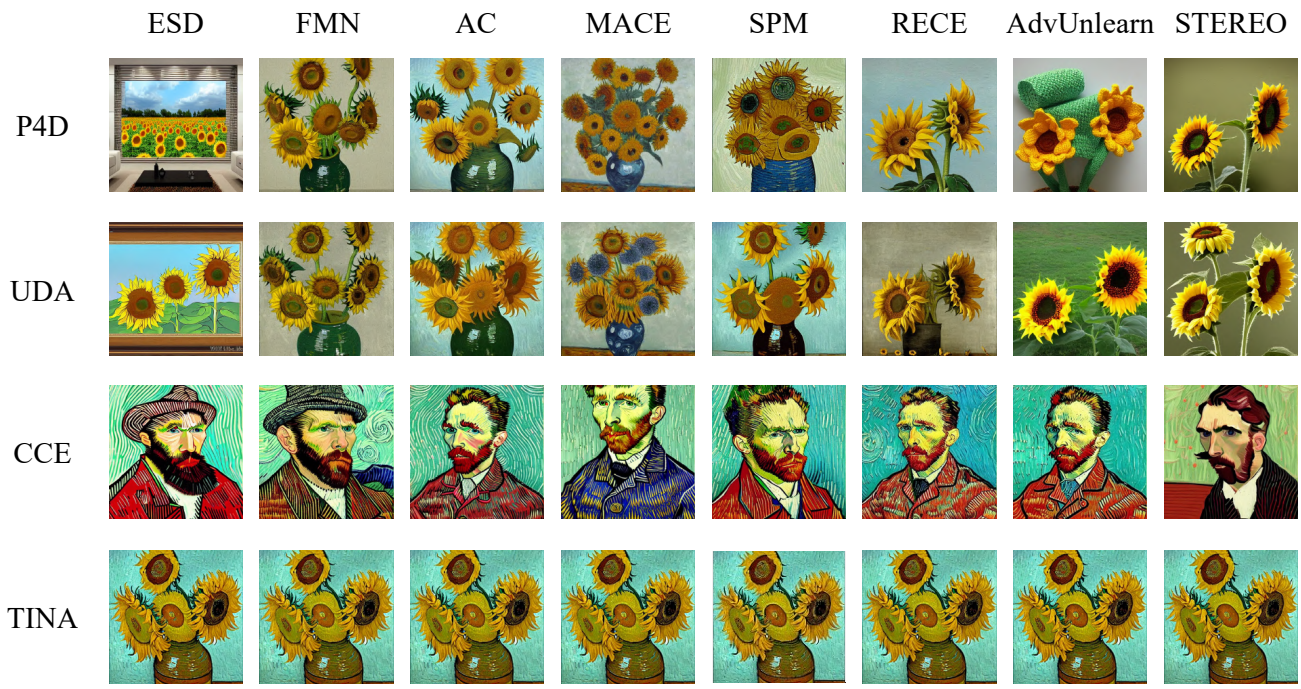

Figure 11. Qualitative comparison of attack methods on the “Van Gogh” style erasure task. Each row represents an attack, and each column represents a specific unlearning defense.