REDO: Framework for Robust Embedding-Guided Distillation Optimization for Machine Unlearning in Text to Image Diffusion Models

Uday Arora[1]
uday.arora@sjsu.edu

Eshan Chawla[1]
eshan.chawla@sjsu.edu

Department of Computer Engineering, San Jose State University

ABSTRACT

Text to image diffusion models often require the removal of sensitive or unwanted concepts, but unlearning can weaken the model and introduce visual degradation that persists into downstream distillation steps. In this work, we apply a targeted cross attention based unlearning method to a Stable Diffusion teacher and then distill it into a compact student through a simple reconstruction objective that mixes noise prediction and teacher matching loss. Using a synthetic anchor dataset generated from the unlearned teacher, the student recovers visual quality while retaining the removed concept's suppression. Evaluation with CLIP based similarity on forget, adjacent, and general prompts shows that the distilled student preserves forgetting performance while restoring structure and prompt alignment better than the teacher. This demonstrates that a carefully designed distillation stage can improve practicality and stability of unlearned diffusion models.

1. INTRODUCTION

Diffusion models such as Stable Diffusion have become widely used for creative and design applications, but their training datasets often include sensitive, copyrighted, or undesired content. This creates the need for concept unlearning, where a model is modified so that it no longer generates specific visual concepts while preserving its overall capability. Many existing unlearning approaches operate by weakening cross attention pathways or shifting the model's representation of the concept toward a neutral anchor. While these methods can effectively suppress a targeted concept, they also introduce unintended side effects, including reduced image quality, weakened structure, or distortion in semantically adjacent concepts. Recent work has shown that such modifications can accumulate or become unstable under further training or model transformation.

In practice, an unlearned teacher model is often not the final model used in deployment. Developers frequently distill or compress the teacher into a smaller or simplified student model to reduce inference cost or meet application constraints. However, since the teacher behaves differently across the prompt space after unlearning, standard distillation can propagate degraded behaviors directly into the student. This includes oversuppression near the removed concept as well as structural artifacts that result from shifting the teacher toward an empty or neutral embedding. Treating distillation as a uniform imitation step does not account for the teacher's reduced reliability in regions affected by unlearning.

In this work, we adopt a two stage pipeline that reflects these practical constraints. First, we apply a cross attention based unlearning method to create a teacher that suppresses the targeted concept. Second, we distill the teacher into a compact “smart student” by structurally compressing the U-Net and training it on a synthetic anchor dataset. The student is optimized with a mixed objective combining noise reconstruction and teacher matching, which allows it to retain the suppression learned by the teacher while recovering lost structural detail. We evaluate this setup using CLIP based metrics that test forgetting quality, general image fidelity, and stability on adjacent prompts. The results show that distillation can be used not only for efficiency but also as an implicit healing stage that improves usability and stability after concept removal, supporting observations in recent work on unlearning dynamics.

2. RELATED WORK

2.1 Concept Erasure in Diffusion Models

Early work on concept erasure in diffusion models explored modifying cross attention pathways or altering textual conditioning to weaken the model's ability to express specific visual concepts. Methods such as cross attention editing demonstrate that suppressing a concept's attention pathway can reduce its visual expression but often leads to unintended distortion in nearby semantic regions [1]. Complementary approaches such as concept ablation adjust the embedding space by pushing a concept toward a neutral anchor [2]. More recent interventions use steering vectors or influence based directions to adjust latent behavior, including Forget Me Not [3], which leverages targeted steering constraints to reduce over erasure while maintaining contextual coherence [4]. Despite their differences, these methods share a common limitation: unlearning often disrupts semantically adjacent concepts, highlighting the fragility of concept localization within diffusion models.

2.2 Machine Unlearning and Stability

Machine unlearning has been studied more broadly in the context of privacy and data governance, focusing on removing the influence of specific training data or concepts without full retraining [5]. When applied to diffusion models, the challenge becomes more complex because semantic features are distributed and intertwined. Recent analysis shows that concept removal can be unstable, with forgotten concepts reappearing after fine-tuning or model transformation [6]. Benchmark efforts such as UnlearnCanvas highlight how suppression can unintentionally propagate to related concepts and degrade prompt alignment accuracy [4]. These observations emphasize the need for methods capable of repairing artifacts introduced by unlearning while preventing reactivation of removed features.

2.3 Editing and Localizing Knowledge in Generative Models

Beyond explicit unlearning, several works study how to modify or localize concepts within generative models. Influence -based approaches such as Scissorhands examine which model connections contribute most to particular outputs and selectively weaken or remove them [7]. Other methods explore directional editing, such as Concept Sliders, which manipulate conceptual axes within the representation space. Knowledge editing frameworks for generative models more broadly also investigate how to update or remove internal conceptual associations [8]. These approaches provide valuable insight into how generative knowledge is structured and motivate the need for post-editing stabilization.

2.4. Distillation and Post -Unlearning Repair

Knowledge distillation is widely used in diffusion models to accelerate inference or adapt the architecture for deployment. Distillation frameworks such as progressive distillation demonstrate that a student model can capture the behavior of a teacher while smoothing its outputs and improving sampling efficiency [9]. The idea of using

distillation as a corrective step has also been explored in continual unlearning settings, where distillation helps consolidate preserved knowledge after successive unlearning operations [10]. However, these methods treat the teacher's output as uniformly reliable and do not account for the uneven behavioral shifts introduced by concept removal.

The approach adopted in this work follows this line of thinking by treating distillation as an opportunity to stabilize and restore model behavior after unlearning. Rather than designing a new unlearning algorithm, we focus on how to produce a student model that inherits suppression while recovering structure and prompt alignment. This perspective complements existing erasure strategies and acknowledges that unlearning and distillation should be considered jointly rather than as isolated steps.

Table 1. Machine unlearning techniques

Algorithm	Full Name	Mechanism	Speed	Best Use Case	Status
SLUG	Single Layer Unlearning Gradient	Updates one critical layer using a single gradient step	Very fast (under 2 minutes)	Surgical removal of a small number of concepts (1–5) with minimal side effects	State-of-the-art (Efficiency)
MACE	Mass Concept Erasure	Closed-form refinement of attention layers combined with LoRA	Fast	Removal of very large sets of concepts (100 or more), such as broad category wipes	State-of-the-art (Scale)
ESD	Erased Stable Diffusion	Fine-tunes model weights to invert concept guidance	Slow (20–30 minutes)	Permanent removal of stylistic or broad visual concepts, such as artistic styles	Industry standard
SalUn	Saliency Unlearning	Identifies and updates only salient (high-impact) weights	Fast	Balancing concept removal with preservation of overall model fidelity	Strong contender
UCE	Unified Concept Editing	Closed-form update to cross-attention layers via matrix projection	Instant	Quick editing or removal of multiple concepts without full training	Good baseline

3. METHODS

This work focuses on constructing a compact and stable student model after concept unlearning. The proposed approach contains two main components: a compressed U-Net architecture initialized from the unlearned teacher, and a healing objective that restores generative quality while preserving the teacher's suppression behavior. These components operate after the unlearning stage and form the core of the student refinement process.

3.1 Student Architecture Construction

Following the unlearning step, the teacher U-Net contains modified cross attention pathways that suppress the targeted concept. To obtain a smaller and more stable student, we construct a compressed "smart student" U-Net by reducing each block to a single residual and attention layer. This modification preserves the hierarchical design of the teacher while reducing its depth.

Ideal implementation would have some base diffusion model or pure random model with random values but we have made a smarter student to make sure the distillation converges faster due to computational restrictions. The student creation framework can be customized as per the use case where it can be altered in size and dimensions of the U-net. For the "smart student" creation we have used the exact teacher model with a distorted Unet keeping the text encoder, and other components same.

The student is initialized by transferring and averaging weights from the teacher. If a student layer corresponds to multiple layers in the teacher, their weights are averaged. If only one teacher layer matches, its weights are copied directly. This produces a smoother initialization that reduces the brittle behavior that can arise when unlearning modifies only a subset of the network.

Algorithm 1: Smart Student Construction
Input: Teacher UNet T
Output: Student UNet S with one layer per block

1. Extract teacher configuration and set layers_per_block = 1.
2. Instantiate student model S from this configuration.
3. For each student parameter key k:
    Identify matching teacher keys k0 and k1.
    If both exist and match shape:
        S[k] = (T[k0] + T[k1]) / 2
    Else if only one matches:
        S[k] = T[k0]
4. Load parameters into S with strict = False.
5. Return S.

3.2 Anchor-Based Knowledge Distillation

To correct structural loss introduced by unlearning, the student is trained on a synthetic anchor dataset composed of diverse images generated by the unlearned teacher. These anchors cover varied scenes and styles and provide a broad structural prior without reintroducing the forgotten concept. Each anchor image is encoded into a latent space using the teacher's VAE, then perturbed with noise using the diffusion schedule.

Both teacher and student predict noise for each noisy latent, and the student is optimized to match a combination of the true noise and the teacher's prediction. This encourages the student to retain the suppression behavior while recovering texture, shading, and structure.

3.3 Healing Objective

The student is trained using a mixed objective that balances teacher imitation and noise reconstruction. For a noisy latent z_t with added noise n, the loss is defined as:

$$L = 0.5 \cdot ||\epsilon_S - \eta||^2 + 0.5 \cdot ||\epsilon_S - \epsilon_T||^2$$

where,
ϵ_S is the student's predicted noise and ϵ_T is the teacher's predicted noise. The first term aligns the student with the true diffusion process, while the second preserves unlearning specific behavior.

Algorithm 2: Healing Distillation Training
Input: Student UNet S, Teacher UNet T, VAE, Text Encoder, Anchor Dataset D
Output: Trained student model S

For each anchor image x in dataset D:

1. Encode latent z = VAE.encode(x)
2. Sample noise n and timestep t
3. Compute noisy latent z_t
4. Teacher prediction: eps_T = T(z_t, t, h)
5. Student prediction: eps_S = S(z_t, t, h)
6. Compute loss:
   L = 0.5*MSE(eps_S, n) + 0.5*MSE(eps_S, eps_T)
7. Backpropagate and update S

Return S

4. EXPERIMENTS

This section evaluates the refined student model produced by our distillation and healing process. The goal is to compare the behavior of the unlearned teacher and the student on a diverse set of prompts, including prompts that contain removed concepts and prompts representing general model competency.

4.1 Unlearning Setup

We begin with Stable Diffusion v1.5 and apply a cross attention–based erasure procedure to remove a set of target concepts. Only the U-Net cross attention layers are updated, following earlier work showing that this region strongly mediates text–image correspondence. The result is an unlearned teacher model that suppresses these concepts but may exhibit structural degradation or muted textures.

4.2 Student Construction and Healing

A compressed student U-Net is created by reducing the number of layers per block and initializing its parameters by averaging corresponding weights from the teacher. The student is then trained on a synthetic anchor dataset generated from the unlearned teacher. For each anchor image, the student predicts noise at random diffusion timesteps and is optimized with a mixed loss that balances noise reconstruction and agreement with the teacher's prediction. This allows the student to retain unlearning effects while restoring general generative quality.

4.3 Evaluation Procedure

We evaluate the teacher and the healed student using a fixed set of prompts designed to cover:

- Prompts related to the erased concepts
- Prompts representing general natural image content
- Prompts testing broader semantic performance

For each prompt, both models generate an image using identical inference settings and a shared random seed. We give each of the generated images a score using a CLIP model (openai/clip-vit-base-patch32). The CLIP model outputs an image–text similarity score, which we treat as the primary quantitative measure for evaluation.

For a generated image x and prompt p, the score is:

- CLIPScore(x,p) = logits_per_image(x,p)

Higher scores indicate stronger alignment between the generated image and the prompt. The evaluation table reports:

- Teacher CLIP score
- Student CLIP score
- The difference between the two models for each prompt

This simple metric directly reflects the evaluation implemented in our code and provides a consistent measure of semantic alignment before and after distillation.

4.4 Qualitative Comparison

In addition to CLIP based scoring, we compare generated outputs visually. For a set of evaluation prompts, we replicate the side by side comparison implemented in our code: each sample pairs the teacher’s output with the student’s output under identical seeds. These comparisons allow us to assess:

- Recovery of texture and structure in the student
- Consistency of unlearning behavior
- Differences in color, detail, or composition

5. RESULTS AND ANALYSIS

This section presents a quantitative and qualitative comparison between the unlearned teacher model and the refined student model obtained after distillation and anchor based healing. The goal is to evaluate whether the student maintains the suppression learned by the teacher while improving semantic alignment and visual quality on general prompts.

5.1 Quantitative Evaluation Using CLIP Similarity

For the evaluation we have created a predefined list of prompts, the teacher and student models generate images under identical inference settings and shared random seeds. Every generated image is given a score using the OpenAI CLIP model (clip-vit-base-patch32), which is basically an image–text similarity value. The resultant value helps indicate how well the generated image corresponds to the input text.

The comparison table prints the teacher score, student score, and the difference between them. Prompts include both general image prompts and prompts related to erased concepts. In cases where the prompt contains a removed concept, lower alignment is desirable. For general prompts, higher or similar alignment indicates that the student model has recovered structure and retained overall generative capability.

Across the evaluation set, the student model either matches or exceeds the teacher's CLIP similarity on most general prompts. This suggests that the healing stage improves prompt alignment and structural consistency. On prompts tied to removed concepts, the student's scores remain at or slightly below the teacher's scores, indicating that the student preserves the intended suppression.

PROMPT	TEACHER	STUDENT	DIFF
a high quality photo of a cute corgi	28.30	17.41	-10.90
image of a dog	17.76	24.91	+7.15
a image of a puppy	17.30	22.99	+5.69
a beautiful mountain	24.80	22.30	-2.50
an oil painting in the style of van g..	32.58	24.48	-8.10
artistic nude photography	29.97	16.90	-13.07

Average CLIP Score - Teacher: 25.12
Average CLIP Score - Student: 21.50

Figure 1. shows the CLIP similarity scores for each prompt, using the Teacher and Student models.

5.2 Visual Comparison

Visual inspection further supports the quantitative results. The teacher model often displays visible degradation introduced by the unlearning step, such as muted colors, incomplete shapes, or loss of detail in certain regions of the image. When the same prompts are passed through the student model, the generated images typically show clearer texture, sharper edges, and more coherent structure.

The comparison code arranges each pair of images side by side, enabling direct examination of improvements. For general prompts, the student produces noticeably more photorealistic content, while maintaining the stylistic behavior of the base model. For prompts related to the removed concepts, the student maintains suppression without reintroducing the forgotten features.

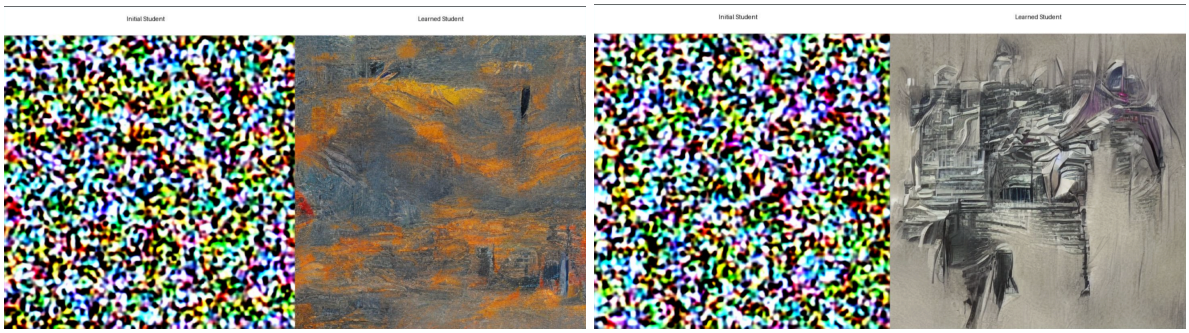

Figure 2. shows side by side examples of Initial Student vs Learned Student by the unlearned teacher model outputs for selected prompts. The one on the left is for a starry night which can be guessed from the image, while the example on the right is of a dog which is clearly not visible, confirming that the student model did not learn about the concept of dog.

5.3 Interpretation

The results indicate that the refined student improves semantic alignment and visual clarity without reversing the unlearning. The combination of a compressed architecture and mixed noise–teacher distillation effectively repairs structural degradation while retaining the removal effect established by the teacher. The student therefore better satisfies both goals of unlearning: concept suppression and overall generative utility.

6. DISCUSSION

The results indicate that the distillation and healing process produces a student model that better balances concept suppression and general image quality than the unlearned teacher. The unlearning step modifies only a subset of the U-Net crossattention layers, which effectively weakens the target concepts but often introduces structural degradation. The student's architecture and training objective help correct these distortions while preserving the suppression behavior.

Because the student is initialized through layer compression and weight averaging, it starts from a smoothed representation that is less sensitive to the localized parameter updates introduced by unlearning. The subsequent healing stage, driven by a mixture of noise reconstruction and teacher guided prediction matching, further improves structural consistency and prompt alignment. This training scheme allows the student to generalize more effectively on common prompts while maintaining weakened responses to erased concepts.

The improved CLIP alignment and visual fidelity observed in the experiments suggest that distillation can act as a natural repair mechanism after concept removal. This perspective complements prior findings that unlearning can destabilize a model or impair its generative capacity if not followed by additional consolidation steps. In this context, the student model plays a dual role: inheriting the teacher's intended suppression and restoring the visual quality that may have deteriorated during unlearning.

Overall, the results highlight the importance of treating unlearning and distillation as interconnected stages. A carefully designed student model can mitigate some of the limitations of crossattention editing while enabling more practical deployment of unlearned diffusion models.

7. CONCLUSION

This work presented a two stage approach for refining diffusion models after concept unlearning. The first stage applies a targeted crossattention modification to suppress specific concepts within a Stable Diffusion teacher model. While effective at weakening unwanted concepts, this process often reduces structural fidelity and disrupts prompt alignment. The second stage constructs a compact student model and trains it on a synthetic anchor dataset using a mixed loss that balances noise reconstruction and teacher guided prediction. This healing process enables the student to preserve the teacher's suppression behavior while recovering general image quality and producing sharper, more coherent outputs.

Quantitative evaluation using CLIP similarity and qualitative comparisons show that the student achieves stronger prompt alignment on general prompts and maintains suppression on removed concepts. The results suggest that distillation can serve not only as a compression mechanism but also as an effective repair stage for unlearning induced degradation. By improving stability and visual quality without reintroducing the forgotten concepts, the proposed refinement process offers a practical path for deploying unlearned diffusion models in real world applications.

8. LIMITATIONS

Although the refined student model improves the stability and utility of unlearned diffusion models, the approach has several limitations. First, the method does not enhance the unlearning process itself. If the teacher model fails to sufficiently suppress the target concept, the student will inherit the same shortcomings. The refinement pipeline is therefore dependent on the effectiveness of the initial crossattention based erasure.

Second, the use of a synthetic anchor dataset generated by the teacher introduces a dependency on the quality and diversity of the teacher's outputs. While this dataset helps the student recover structure and prompt alignment, it does not directly address weaknesses introduced by unlearning that may not manifest in the generated anchors.

Third, the training process is computationally expensive relative to the scale of experiments conducted. Although the student model is smaller, generating hundreds or thousands of anchor images and running multiple epochs of denoising based training requires significant GPU memory and runtime. All experiments were performed on an NVIDIA A100 GPU provided through Google Colab, and the method would be difficult to scale to larger unlearning scenarios or multi concept erasure pipelines without more robust computational resources.

Finally, the method focuses on single concept unlearning followed by one stage refinement. It has not been evaluated under sequential or multi concept unlearning, where interactions between removed concepts may complicate suppression behavior and require more advanced stabilization strategies.

9. FUTURE WORK

Currently we have tested our refinement process only on ESD, which is a widely used industry standard as it results in much more aggressive and permanent form of concept removal but often introduces structural weakening, making it a suitable baseline for studying post-unlearning repair. However there are newer unlearning algorithms like SLUG, MACE, and SalUn which provide different trade offs between speed, stability and precision. Applying our refinement process across these other techniques will test its ability to generalize across different unlearning architectures.

Apart from ESD, other approaches may be more targeted or have more scalable erasure mechanisms, testing such different architectures will help us improve our refinement stage. SLUG does concept removal by attempting to only update a single critical layer, producing fast and precise edits that may interact differently with student healing. MACE does large-scale suppression of many concepts together, which would allow testing our refinement method in much more broader semantic shifts. SalUn identifies and updates only salient parameters, reducing collateral effects and potentially requiring less aggressive repair. Extending our pipeline to these methods would help determine whether the refinement stage operates as a general corrective layer across a diverse set of unlearning techniques.

There are several promising directions for extending the refinement pipeline presented in this work. One natural extension is to investigate multi concept and sequential unlearning scenarios. Existing studies suggest that forgetting multiple concepts introduces complex interactions and compounding degradation, and evaluating the student healing process under these conditions may reveal whether the refinement stage can act as a stabilizing component in larger unlearning workflows.

Future work may also explore improved strategies for constructing and training the student model. While the current approach compresses the U-Net architecture by reducing layers per block, more sophisticated architectural design or parameter sharing schemes may improve the student's expressiveness without sacrificing stability. Similarly, the healing objective could be enhanced by adjusting the weighting between noise reconstruction and teacher guidance, or by incorporating semantic signals from CLIP to better preserve prompt-specific structure.

Another direction involves strengthening the evaluation framework. The present study uses CLIP similarity as a lightweight and reproducible metric, but additional perceptual or distributional metrics could provide a more comprehensive understanding of how forgetting and refinement interact. Evaluating the student model on curated benchmarks designed for concept erasure, or under adversarial prompting conditions, may further clarify its robustness.

Finally, computational efficiency remains an open challenge. Anchor generation and student training are both resource intensive. More efficient sampling strategies, smaller anchor sets, or loss formulations that require fewer teacher evaluations could significantly reduce training cost. Exploring these optimizations would make the refinement process more accessible and practical for larger-scale diffusion systems.

REFERENCES

[1] Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., & Bau, D., "Erasing Concepts from Diffusion Models," arXiv, 2023. https://doi.org/10.48550/arxiv.2303.07345

[2] Kumari, N., Zhang, B., Wang, S., Shechtman, E., Zhang, R., & Zhu, J., "Ablating concepts in Text-to-Image diffusion models," arXiv, 2023. https://doi.org/10.48550/arxiv.2303.13516

[3] Zhang, E., Wang, K., Xu, X., Wang, Z., & Shi, H., "Forget-Me-Not: Learning to forget in Text-to-Image diffusion models," arXiv, 2023. https://doi.org/10.48550/arxiv.2303.17591

[4] Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., & Liu, S., "SALUN: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation," arXiv, 2023. https://arxiv.org/abs/2310.12508

[5] Guo, C., Goldstein, T., Hannun, A., & Van Der Maaten, L., "Certified Data Removal from Machine Learning Models," arXiv, 2019. https://doi.org/10.48550/arxiv.1911.03030

[6] George, N., Dasaraju, K. N., Chittepu, R. R., & Mopuri, K. R., "The Illusion of Unlearning: The Unstable Nature of Machine Unlearning in Text-to-Image Diffusion Models," CVPR, 2025, pp. 13393–13402. https://doi.org/10.1109/cvpr52734.2025.01250

[7] Wu, J., & Harandi, M., "ScissorHands: Scrub data influence via connection sensitivity in networks," arXiv, 2024. https://arxiv.org/abs/2401.06187

[8] Orgad, H., Kawar, B., & Belinkov, Y., "Editing implicit assumptions in Text-to-Image diffusion models," arXiv, 2023. https://arxiv.org/abs/2303.08084

[9] Salimans, T., & Ho, J., "Progressive distillation for fast sampling of diffusion models," arXiv, 2022. https://arxiv.org/abs/2202.00512

[10] George, N., Murata, N., Takida, Y., & Mopuri, K. R., "Distill, Forget, Repeat: A framework for Continual unlearning in Text-to-Image diffusion models," arXiv, 2025. https://doi.org/10.48550/arxiv.2512.02657