
Supplementary Material for "Semantic Surgery: Zero-Shot Concept Erasure in Diffusion Models" (Submission ID: 13111)

Anonymous Author(s)

Affiliation

Address

email

A Overview

This supplementary document for "Semantic Surgery: Zero-Shot Concept Erasure in Diffusion Models" (Submission ID: 13111) provides key details complementing the main paper. It includes: (1) Detailed experimental setups, including hyperparameter settings and visual feedback configurations (Appendix B); (2) A qualitative analysis of our Co-Occurrence Encoding’s advantages (Appendix C); and (3) An experimental investigation into hyperparameter sensitivity for γ and β (Appendix D). References to the main paper (e.g., "main paper, Sec. X.X") are used for context.

B Detailed Experimental Setup

B.1 Hyperparameter Settings and Visual Feedback Configuration

Key hyperparameters and the use of the Latent Concept Persistence (LCP) mitigation module (visual feedback) varied across experiments, summarized in Table 1. The LCP module, when active, performs a second-stage inference if visual concepts targeted for erasure are detected in the first-stage output. Feedback strength is modulated by λ_{vis} , and the detection threshold is `detect_threshold` (a parameter in our `StableDiffuser` class, typically $\approx 0.5 - 0.6$).

Table 1: Summary of key hyperparameter settings and LCP visual feedback configuration across experiments. β is task-dependent (see main paper, Sec. 3.2). $\gamma = 0.02$, $\tau = 0.5$ (`alpha_threshold` in code) are global unless specified.

Experiment Task	Target Concepts (Example)	β (Typical)	Visual Feedback (LCP)	Detector for LCP	λ_{vis} (if LCP active)
CIFAR-10 Object Erasure	"airplane", "dog", etc.	-0.12	Yes	AOD [5]	1.0
I2P Explicit Content Removal	"nude", "erotic", etc.	-0.06	Yes	NudeNet [1]	1.0
Artistic Style Erasure	"Brent Heighon"	-0.30	No	N/A	N/A
Multi-Concept Celebrity Erasure	"Melania Trump", "Adam Driver"	-0.28	No	N/A	N/A

B.2 Visual Feedback Implementation Details

Visual feedback (LCP mitigation) was employed for CIFAR-10 object erasure (using AOD [5]) and I2P explicit content removal (using NudeNet [1]). If a targeted concept was detected above `detect_threshold` in the initial generation, feedback reinforced erasure in a second pass. For I2P, NudeNet identifies exposed body parts (e.g., `BUTTOCKS_EXPOSED`, `FEMALE_BREAST_EXPOSED`). If the maximum detection score for any such element exceeded `detect_threshold` (e.g., 0.6), the erasure strength (`alpha_f` in code) for all four targeted abstract concepts ("nude", "naked", "erotic",

"sexual") was simultaneously increased during the second-stage generation. This is achieved by setting `alpha_feed_back` to a high value (amplified by λ_{vis}) and including all four concepts in `indexes_union` for the second pass.

B.3 Targeted Concepts for I2P Explicit Content Removal

For I2P explicit content removal (main paper, Sec. 4.3; I2P dataset [6]), following SA [2], we targeted four concepts: "nude", "erotic", "sexual", and "naked". These were erased using Co-Occurrence Encoding, with NudeNet [1] providing visual feedback as detailed above.

B.4 Dataset Details

The datasets used for our experiments were sourced as follows:

- **CIFAR-10 Object Erasure:** We utilized the simple prompts and paraphrased prompts from the CIFAR-10 task setup of Receler [3].
- **I2P Explicit Content Removal:** This task used the standard I2P (Imagen Prompt Dataset) [6], a common benchmark for evaluating safety in text-to-image models.
- **Artistic Style Erasure:** The list of artists targeted for style erasure was adopted from the experimental setup of MACE [4], facilitating direct comparison.
- **Multi-Concept Celebrity Erasure:** Similarly, the set of celebrities for the multi-concept erasure task was also sourced from MACE [4] to ensure fair comparability with prior work.

Using established dataset configurations where possible aids in reproducibility and allows for more direct comparisons with existing methods.

C Qualitative Analysis of Co-Occurrence Encoding

Our main paper (Sec. 3.1) introduces Co-Occurrence Encoding (Eq. 6) for robust multi-concept erasure. Figure 1 qualitatively compares it against a "Naive Approach" (summing individual erasure vectors) for erasing "dog" and "cat" from "dog and cat playing together."

Co-Occurrence Encoding successfully removes both target animals while preserving the "playing together" action, often substituting them with plausible alternatives like children, thus maintaining the scene's narrative. In contrast, the Naive Approach significantly degrades image quality and semantic coherence, leading to muddled imagery. This demonstrates Co-Occurrence Encoding's advantage in neutralizing targets while protecting contextual integrity and image quality, unlike the naive method's tendency to over-erase.

D Experimental Analysis of Hyperparameter Sensitivity

We analyzed sensitivity to hyperparameters γ (sigmoid steepness, main paper, Eq. 12) and β (concept presence threshold, main paper, Sec. 3.2) on CIFAR-10 object erasure (average over 10 classes, no visual feedback for this test). Defaults: $\gamma = 0.02$ (global), task-dependent β (e.g., ≈ -0.12 for CIFAR-10). For γ sensitivity, β was fixed at a near optimal value (e.g., -0.06).

Impact of γ (Sigmoid Steepness): Figure 2 shows metrics (Acc_E , Acc_R , Acc_L , H_c) as γ varies (log scale: 0.02 to 1.0). Stability across this range indicates low sensitivity to γ , justifying our global choice of $\gamma = 0.02$.

Impact of β (Concept Presence Threshold): Figure 3 shows β 's impact. β is crucial for erasure activation. H_c indicates an optimal β range (around -0.12 to -0.06 here) balancing effective erasure ($Acc_E \approx 0$) with high Acc_R and Acc_L . This supports selecting β based on α_c distributions (main paper, Sec. 3.2).

These analyses confirm that while β requires careful selection, robustness to γ variations enhances practicality.



Figure 1: Qualitative comparison for multi-concept erasure ("dog", "cat"). Our Co-Occurrence Encoding (center) preserves semantics (e.g., "playing together" with children) while the Naive Approach (right) degrades image quality compared to the Original (left).

References

- [1] Praneeth Bedapudi. Nudenet: Neural nets for nudity detection and censoring, 2022. [URL https://github.com/notAI-tech/NudeNet](https://github.com/notAI-tech/NudeNet).
- [2] Alvin Heng and Harold Soh. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36:17170–17194, 2023.
- [3] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *European Conference on Computer Vision*, pages 360–376. Springer, 2024.
- [4] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.
- [5] Andrew Ng and Landing AI. Agentic object detection. <https://landing.ai/agentic-object-detection>, 2024. Accessed: 2025-03-05.

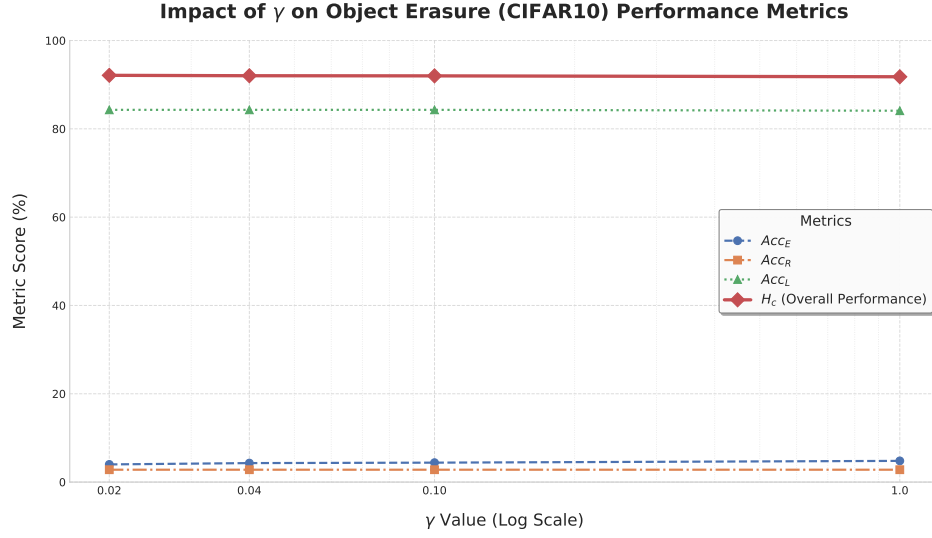


Figure 2: Impact of γ (Sigmoid Steepness, log scale) on Object Erasure (CIFAR-10) Performance Metrics. Metrics show high stability across tested γ values.

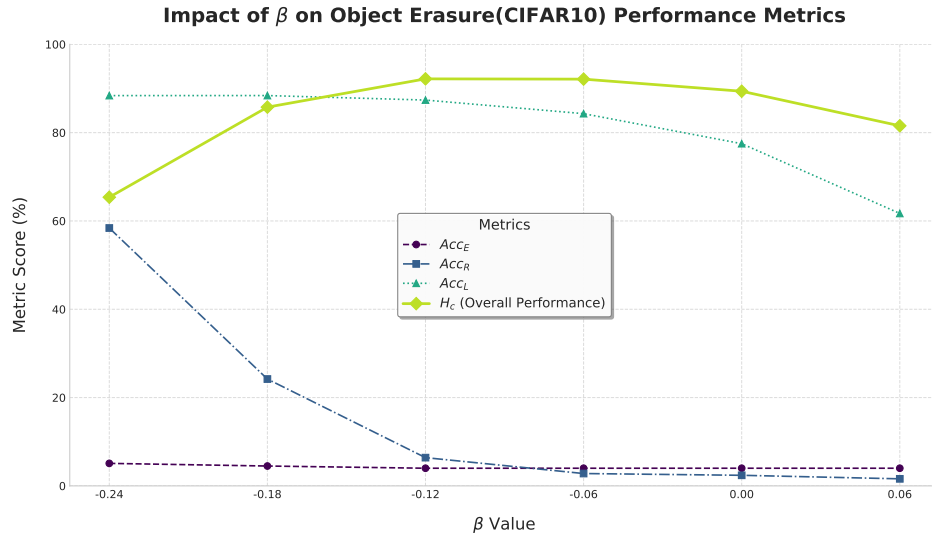


Figure 3: Impact of β (Concept Presence Threshold) on Object Erasure (CIFAR-10) Performance Metrics. An optimal β range balances erasure effectiveness with semantic preservation.

- 80 [6] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent
81 diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the*
82 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023.