Security and Privacy of Machine Learning, 2025 Critique G7: Machine Unlearning – (1) SAeUron: Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders (2) Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

CRITIQUE:(1) SAEURON: INTERPRETABLE CONCEPT UNLEARNING IN DIFFUSION MODELS WITH SPARSE AUTOENCODERS

I. SUMMARY

The paper presents **SAeUron**, an activation-space unlearning framework for text-to-image diffusion models. Instead of fine-tuning weights, the authors train *k*-sparse autoencoders (SAEs) on internal cross-attention activations of Stable Diffusion across multiple denoising steps to discover sparse, interpretable features. At inference, they ablate a small set of concept-correlated latent features (selected via a score contrasting in/out-of-concept average activations over timesteps), thereby suppressing unwanted styles or objects. On the UnlearnCanvas benchmark [1], SAeUron achieves SOTA average performance for style unlearning and competitive results for object unlearning, with low memory/storage overheads. It also reduces nudity on I2P and exhibits robustness to adversarial prompt attacks. The approach takes inspiration from mechanistic interpretability advances with SAEs [2] and is complementary to the diffusion modeling literature [3], [4], while contrasting with weight-editing concept-erasure methods such as ESD [5].

II. STRENGTHS

1) **Clear conceptual pivot:** Moves unlearning from parameter space to activation space. This avoids catastrophic interference typical of fine-tuning while enabling transparent control over *which* features are removed.
2) **Interpretability-first design:** Training SAEs on cross-attention activations yields monosemantic, human-aligned features (e.g., paws/ears/texture patches). Heatmaps and per-feature inspection make failure analysis feasible—a rare property in safety tooling [2].
3) **Robustness evidence:** The method remains resilient under UnlearnDiffAtk-style prompt optimization, suggesting it is less about *masking* and more about *removing* the targeted generative factors.
4) **Practicality:** Only two SAEs (objects/styles) trained once and reused across many concepts; low GPU memory at inference and small storage footprint compared with multiple per-concept finetunes.
5) **Multi-concept scaling:** Sequential and simultaneous unlearning (even extreme settings) show graceful degradation, a known weakness of weight-editing pipelines [1].

III. WEAKNESSES

1) **Block specificity & transferability:** The approach hinges on empirically chosen cross-attention blocks (e.g., `up.1.1`, `up.1.2`). It is unclear how robust block selection is across model families (SD 2.x/SDXL, non-UNet architectures) or after distillation; portability claims would benefit from broader cross-arch studies.
2) **Per-concept hyperparameter tuning:** Object unlearning requires concept-specific (τ_c, γ_c); this introduces manual effort and a potential new attack surface (over/under-suppression). An automated, reliable tuning protocol is not yet demonstrated.
3) **Similarity entanglement:** The method can unintentionally impact visually similar non-target classes (e.g., cats vs. dogs). This reflects partial feature overlap and raises questions about fine-grained disentanglement limits of single-block SAEs.
4) **Abstract-safety limits:** Activation-level removal excels on concrete visual concepts, but struggles on abstract categories (e.g., hate/violence semantics). This caps its applicability as a one-stop safety solution.

5) **Operational constraints:** Because filtering is inserted at inference, open-weight deployments could disable it. The claims primarily fit *closed* or API-served systems.

IV. Potential Improvements / Extensions

1) **Adaptive, timestep-aware ablation:** Learn a policy (or schedule) over timesteps for (τ_c, γ_c) to minimize collateral damage while maintaining UA. This could be optimized on a small validation set with multi-objective criteria (UA, IRA/CRA, FID).
2) **Multi-block compositional SAEs:** Train lightweight SAEs for multiple cross-/self-attention bottlenecks and fuse their attributions causally. This can reduce reliance on one block and improve handling of similar classes by exploiting complementary features.
3) **Automatic hyperparameter selection:** Calibrate (τ_c, γ_c) using uncertainty-aware selection (e.g., cross-validated UA retention curves or PAC-inspired bounds), then lock them per concept to avoid manual tuning and improve reproducibility.
4) **Abstract-concept bridging:** Couple activation filtering with *text-encoder* SAE steering or concept bottlenecks, enabling alignment to semantic attributes that lack crisp visual correlates while preserving the core activation-level benefits.
5) **Model-agnostic validation:** Evaluate portability across SD 2.x/SDXL and emerging non-UNet backbones; include distillation (Turbo) and inpainting/editing variants to stress-test generalization.
6) **Bypass-resistance hardening:** Ship SAeUron as part of a secure execution path (e.g., server-side graph rewriting, signature checks) and audit that no alternate forward path can skip the SAE gating.

V. Questions for the Authors

1) How stable are the discovered features across seeds/datasets/models? Could you quantify feature alignment (e.g., CKA or transport distances) to show cross-model consistency?
2) Can you formalize conditions under which activation ablation guarantees non-regeneration under adversarial prompting (beyond empirical UnlearnDiffAtk results)?
3) What is the trade-off frontier between UA and fine-grained collateral damage (e.g., subtle texture loss) measured with human and automated audits?
4) Could a small number of concept-agnostic "safety latents" explain most safety wins (suggesting a compact universal safety SAE), or are features inherently concept-specific?
5) Can you close the loop to *learn* the feature-scoring rule jointly with the SAE (e.g., via multi-task objectives) instead of post-hoc scoring?

Critique: (2) Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models

Summary

This paper targets the fragility of concept-erased diffusion models (DMs) to adversarial prompts. Building on latent diffusion [6] and unlearning by ESD [5], the authors cast *robust concept erasure* as a bi-level game: an upper-level unlearner versus a lower-level adversarial prompt optimizer. Directly adding adversarial training (AT) degrades image quality, so they propose (i) a **utility-retaining regularization** that penalizes deviation from the original model on a curated, LLM-filtered retain prompt set (inspired by semi-supervised AT [7]) and (ii) **module-wise robustification** that optimizes the *text encoder* rather than the UNet. They also provide a one-step FGSM variant for efficiency (fast AT). Experiments spanning nudity, style, and object erasure show sizable ASR reductions against UnlearnDiffAtk [8] and other attacks, while keeping FID/CLIP competitive. A learned robust text encoder transfers plug-and-play to other SD backbones.

Strengths

1) **Principled framing.** The bi-level formulation fits the attacker–defender nature of prompt jailbreaks, connecting generative unlearning to the TRADES-style robustness–utility trade-off in discriminative models [9]. This anchors design choices in known theory rather than ad-hoc heuristics.
2) **Robustness–utility engineering.** The retain-set regularizer is a simple, effective mechanism to counter the fidelity collapse typically seen when injecting AT into DMs. Using an LLM to filter retain prompts is pragmatic and scalable.
3) **Modularity insight.** Moving optimization to the *text encoder* is a key contribution: better robustness, fewer trainable parameters, and a reusable, plug-in encoder across DMs. This is a practical step toward deployment-ready unlearners.
4) **Broad empirical coverage.** The paper evaluates multiple concept families (nudity, styles, objects), considers several attack types (discrete and embedding-space), and includes ablations (retain-set source, weight γ, fast vs. standard AT), building a credible case for the method's generality.
5) **Clear takeaways.** The work surfaces actionable lessons (e.g., prefix-based adversarial prompting during training works best; deeper text-encoder layers matter for global concepts like nudity) that other practitioners can re-use.

Weaknesses

1) **Metric coupling and proxy bias.** Robustness is primarily assessed via ASR using automated classifiers (e.g., NudeNet) and specific attack generators. These proxies can both under- and over-estimate safety; improvements might partly reflect better "playing to the grader."

Human-in-the-loop judgments or multi-annotator audits are absent.

2) **Threat model coverage.** Although multiple attacks are tested, the space remains narrow: no black-box, adaptive, or mixed *prompt-engineering + image-space* attacks; no reinforcement or evolutionary prompt searches; limited study of transfer attacks across encoders/backbones.
3) **Erasure semantics.** The paper focuses on surface failure (visual presence) but not *latent knowledge leakage* (e.g., re-emergence under compositional prompts, paraphrases, or multi-concept blends). A notion of *certified* or *statistical* erasure is missing.
4) **LLM-judge dependence.** The retain-set filtering relies on a proprietary LLM with unspecified criteria. Misfiltering can reintroduce erased semantics or over-prune semantically adjacent but benign content, risking utility skew.
5) **Compute and stability.** Even with FGSM, bi-level training is costly; stability under hyperparameter drift (e.g., guidance scale, classifier-free guidance, scheduler) is not deeply probed, and results on larger backbones (e.g., SDXL) are not shown.

Potential Improvements / Extensions

1) **Robustness auditing beyond proxies.** Combine ASR with structured human evaluation, multi-annotator labeling, and failure taxonomies (content severity, context, partial exposure). Include calibration metrics for safety classifiers to quantify proxy uncertainty.
2) **Stronger adversaries.** Train-time ensembles of attacks (gradient-based, discrete, black-box), expectation-over-transformations (prompt paraphrasing, multilingual variants), and RL/evolutionary adversaries. Evaluate transfer/black-box attacks crafted on different encoders/backbones.
3) **Erasure guarantees.** Explore probabilistic certificates for concept absence (e.g., randomized smoothing over prompt perturbations) and *compositional* robustness (concept mixes, negations, style+object blends).
4) **Multi-concept and hierarchical unlearning.** Jointly erase concept families (e.g., nudity attributes, style clusters) with capacity control to mitigate collateral forgetting; leverage group sparsity or task vectors.
5) **Retain-set governance.** Replace opaque LLM filtering with auditable criteria: lexical/embedding distance to erased concepts, controllable exclusion radii, and diversity constraints; report how retain-set size/entropy impacts FID and ASR systematically.
6) **Scale and deployment.** Validate on SDXL and instruction-tuned systems; study inference-time defenses (prompt sanitizers, guidance modulation) complementary to AdvUnlearn; test downstream UX impacts (false positives on benign art/fashion).

Questions for the Authors (toward a 10/10 review)

1) How does robustness change under *composite* prompts (erased concept + distractors + style), negations ("no X"), and multilingual paraphrases?
2) What is the sensitivity of results to the safety classifier and its thresholding? Does swapping classifiers materially change ASR conclusions?
3) Could you report Pareto fronts (FID vs. ASR) across γ, retain-set size, and number of encoder layers to clarify the optimal operating region?
4) Does the plug-in encoder preserve robustness when the UNet/backbone is later fine-tuned for personalization (LoRA/Textual Inversion)?
5) Any evidence that utility-regularization pulls the model back toward the pre-erasure distribution in a way that subtly reintroduces erased content under rare prompts?

References

[1] Y. Zhang, C. Fan, Y. Zhang, Y. Yao, J. Jia, J. Liu, G. Zhang, G. Liu, R. R. Kompella, X. Liu, and S. Liu, "Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models," in *NeurIPS Datasets and Benchmarks Track*, 2024.

[2] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," Transformer Circuits Thread, 2023, https://transformer-circuits.pub/2023/monosemantic-features/.

[3] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning (ICML)*, 2015, pp. 2256–2265.

[4] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[5] R. Gandikota, J. Materzyńska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," *arXiv:2303.07345*, 2023.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022, pp. 10 684–10 695.

[7] Y. Carmon, A. Raghunathan, L. Schmidt, P. Liang, and J. C. Duchi, "Unlabeled data improves adversarial robustness," *NeurIPS*, 2019.

[8] Y. Zhang, J. Jia, X. Chen, A. Chen, Y. Zhang, J. Liu, K. Ding, and S. Liu, "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now," in *ECCV*, 2024.

[9] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. El Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *ICML*, 2019.