Evaluating Machine Unlearning in Text-to-Image Diffusion Models Using Sparse Autoencoders

Paul Ellement

Abstract—Text-to-image diffusion models have improved rapidly, complementing language models with their high-quality visual generation. However, this capability also introduces risks of unwanted outputs, such as private, inappropriate, or copyrighted content. Machine unlearning has emerged as a way to mitigate these risks, an area of research which attempts to erase specific concepts from trained models. Existing evaluation benchmarks primarily assess unlearning in terms of image-space behavior and reveal a trade-off between erasure strength and model utility. That is, concept removal is often imprecise, which results in unintentionally suppressing neighbouring concepts and degrading overall performance. This work proposes an internal evaluation based on sparse autoencoders. Instead of only examining generated images, this approach analyzes activations from unlearned diffusion models to probe how internal representations of the erased concept and related concepts change. Experiments indicate that latent directions associated with the target concept and its neighbours broaden and collapse into more generic directions, suggesting a loss of uniqueness. This internal, white-box evaluation perspective complements existing black-box benchmarks and may help guide the design of more precise and reliable unlearning methods in text-to-image models.

Index Terms—Machine Unlearning, Mechanistic Interpretability

1 Introduction

Text-to-image diffusion models have revolutionized generative AI, allowing unprecedented control over visual content creation with natural language prompts. However, their potential to generate copyrighted material, harmful content, or biased representations poses concerns about safety, intellectual property rights, and ethical deployment. Machine unlearning has emerged as a promising research field to address these risks, with the objective being the development of techniques to remove specific concepts from models without having to fully retrain them. Current unlearning methods in diffusion models are often unsuccessful due to erased concepts being partially recoverable through adversarial prompting and the generation of neighbouring concepts (i.e., similar concepts that were not intended to be erased) being negatively impacted.

Determining the effectiveness of an unlearning method typically involves behavioural evaluations such as generation quality metrics with various test prompts. While these methods are helpful to assess the efficacy of erasure, they do not reveal what internal knowledge persists or changes in the model. Understanding residual representations of erased concepts is crucial for developing unlearning methods that truly forget the targeted concepts without over-erasing related knowledge.

Sparse autoencoders (SAEs), which decompose dense, polysemantic activations into sparse, interpretable features, offer a mechanistic interpretability lens for analyzing internal model activations. Recent work has applied SAEs to language models and vision encoders, revealing monosemantic features that have causal effects on outputs. SAEs have also been used to perform ablation in machine unlearning due to their ability to attribute model activations to interpretable concepts. However, they remain underexplored for analyzing unlearned diffusion models, revealing a gap at the intersection of interpretability and model safety.

This paper bridges these domains by introducing a novel evaluation framework for machine unlearning: training SAEs on unlearned diffusion models to probe residual concept representations post-erasure. The hypothesis is that the SAE's latent space can capture traces of erased concepts and enable quantitative assessment of the completeness and the collateral effects of the erasure. The experimental pipeline applies state-of-the-art unlearning methods to erase

Fig. 1: Example image outputs before (left) and throughout unlearning "cat" with strength increasing to the right [1]. Image generation of tiger, lion, and leopard are unintentionally impacted.

visual concepts such as "Van Gogh" style and "cat" from Stable Diffusion, then trains SAEs to extract interpretable features from U-Net attention activations.[1]

Key contributions include:

1) The first SAE-based analysis of unlearned diffusion models, offering a new way to identify persistent concept subspaces.
2) Latent direction analysis, comparing targeted concept residuals from related or general category representations (e.g., Van Gogh vs. art).
3) Empirical evidence of over-erasure, where neighbouring concepts are unintentionally affected.

The results demonstrate that SAEs reveal nuanced knowledge structures, providing a mechanistic angle absent in related work. This advances mechanistic interpretability and machine unlearning by extending SAE applications to post-erasure and providing metrics to verify erasure beyond image output evaluations. This in turn paves the way for more robust and verifiable concept removal in generative systems where safety is critical.

2 Related Work

Machine unlearning and mechanistic interpretability have both recently emerged as research domains, with the former focused on removing concepts from models and the latter on understanding how concepts are internally represented. Recent work on concept erasure in diffusion models has introduced powerful methods for preventing certain model outputs, while benchmarks have begun to evaluate the side effects and limitations of these techniques. In parallel, sparse autoencoders have emerged as a promising tool for encoding model activations into interpretable features and concept directions.

2.1 Machine Unlearning

Machine unlearning attempts to remove certain concepts from trained models without having to fully retrain. Recent work has proposed general frameworks that formalize the trade-offs between completeness and side effects of erasure.

2.1.1 Erasure Methods

Novello et al. introduce an f-divergence notion of unlearning, showing how different divergences control the strength and precision of forgetting, and introducing a trade-off between strong erasure and preservation of model utility [2]. Chen et al. propose Score Forgetting Distillation (SFD), a data-free approach that aligns score functions between "safe" and "unsafe" concepts in diffusion models, enabling efficient concept removal without large retraining datasets [3].

For text-to-image models, Gandikota et al. present Erased Stable Diffusion (ESD), which fine-tunes diffusion models to forget specific concepts such as artist styles, objects, or inappropriate images while attempting to preserve overall visual quality [4]. More recent analyses challenge whether or not such methods truly erase knowledge. George et al. argue in *The Illusion of Unlearning* that many unlearning procedures leave latent correlations intact, so erased knowledge can potentially be recovered through adversarial prompting or distribution shifts [5]. Probing Residual Knowledge in Unlearned Models formalizes this by introducing metrics for quantifying lingering concept traces post-unlearning [6]. This work is aligned with this line of inquiry, treating ESD as a representative erasure method and focusing on examining what remains in the model's internal activations after unlearning.

2.1.2 Evaluation Benchmarks

Benchmarking unlearning has recently grown traction as behavioral tests alone are not robust enough. Probing Residual Knowledge in Unlearned Models introduces quantitative metrics to measure residual concept traces, highlighting that models can pass simple prompt tests while still retaining substantial internal representations of erased concepts [6]. EraseBench proposes a systematic protocol to study the "ripple effects" of concept erasure, measuring both under-erasure (concept leakage) and over-erasure (collateral damage to neighbouring concepts) across multiple methods and concepts [7]. This project analysis fits into this evaluation landscape: like EraseBench, the focus is side effects and residual knowledge, but using SAE latent directions and projection distributions rather than image outputs.

2.2 Mechanistic Interpretability

Sparse autoencoders (SAEs) have emerged as a powerful tool for mechanistic interpretability, decomposing dense, polysemantic activations into sparse, more interpretable features. Early work applied SAEs to language models to discover monosemantic neurons with causal influence on model behavior [8]. Further research extended these ideas to vision and text-to-image models. H-Space Sparse Autoencoders propose an optimized latent space for finding interpretable features in text-to-image systems, improving localization and interpretability of concept features [9]. SAeUron bridges SAEs with concept unlearning in diffusion models, showing that SAE feature ablations can be used to suppress unwanted concepts in a targeted manner [1].

Instead of using SAEs as an active unlearning mechanism, I train SAEs on already-unlearned models and use the resulting latent directions to probe residual concept knowledge and over-erasure effects. This method combines work from ESD-style concept erasure [4], SAE-based interpretability [1], and benchmark evaluations such as EraseBench [7], treating SAE latent geometry as an evaluation metric for concept unlearning in diffusion models.

3 Proposed Method

The goal of this project is to evaluate what remains from erased concepts inside a text-to-image diffusion model after concept unlearning. The pipeline has three stages: (1) erasing target concepts from Stable Diffusion using Erased Stable Diffusion (ESD), (2) training an SAE on activations from the unlearned model, and (3) analyzing SAE latents for targeted, related, and random prompts by analyzing concept directions and projections.

1. The code for this project can be found at https://github.com/paulellement/Unlearning-CSC2529Proj.git

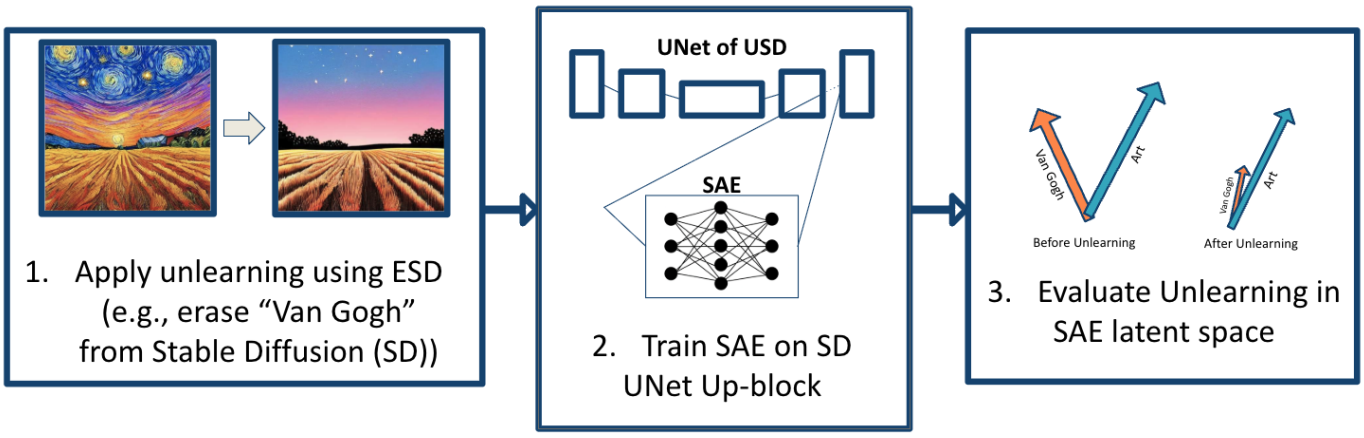

Fig. 2: Overview of my method: (1) Two concepts are erased separately to make two unlearned models: "Van Gogh" and "cat". (2) An SAE is trained on each unlearned model from its U-Net activations. (3) Latent analysis is performed on prompts categorized into those containing the target concepts, related to the target concepts, or unrelated to the target concepts. This identifies side effects of erasure such as neighbouring concepts being impacted.

3.1 Concept Erasure

Concepts were removed from Stable Diffusion v1.4 using the Fine-Tune ESD method [4]. ESD fine-tunes the U-Net of a pretrained diffusion model so that prompts with the target concept (e.g., "cat" or "Van Gogh") are distanced from that concept in score space, while other prompts are constrained to have similar reconstruction to the original. In practice, the method repeatedly:

- samples images using the original model for prompts that explicitly mention the target concept,
- computes a noise-prediction update that reduces alignment with the target concept, and
- regularizes the edited model towards the original model to limit side effects.

This produces an "unlearned" model intended to prevent generations that contain the target concept, while preserving overall image quality and behavior on other prompts. All analysis in this paper is performed on ESD unlearned models.

3.2 SAE Training on the Unlearned Model

To analyze model activations, an SAE was trained on activations from a single U-Net layer, `up_blocks.1.attentions.1`, in the unlearned model. This layer has been used in related work because it encodes mid-level features that influence global style and object structure. The SAE implementation and training pipeline follow the SAeUron framework [1].

Training involves two stages. First, U-Net activations at the specified layer are collected for a large, diverse prompt dataset, a portion of which are related to the erased concept. These activations form the training dataset. Second, an SAE is fit to reconstruct these activations under sparsity constraints. The input dimension is $d_{\text{in}} = 2048$, matching the SD layer width, and an expansion factor of 8 gives a hidden dimension of 16384.

The SAE is trained under two main constraints: reconstruction quality and sparsity. Reconstruction is enforced with an ℓ_2 loss between the SAE output and the original activations. Sparsity is achieved by having an ℓ_1 penalty on the latent activations and by top-k masking: at each forward pass, only the k largest latent activations (here, typically $k = 64$ or $k = 96$) are retained and the rest are zeroed out. This encourages each concept to be represented by a small subset of neurons, causing more interpretable features. Further discussion of SAE evaluation and reconstruction quality can be found in the limitations section.

Once trained, the SAE defines an encoding from U-Net activations to a latent representation space for the unlearned model. Residual representations of erased concepts can be analyzed in this latent space.

3.3 Prompt Selection for Evaluation

To evaluate the unlearning method in SAE latent space, three categories of prompts are used:

1) **Target prompts:** prompts that explicitly include the erased concept (e.g., containing "Van Gogh" or "cat").
2) **Neighbouring prompts:** prompts about related concepts but do not mention the target (e.g., other artists or other animals).
3) **Random prompts:** generic prompts that are unrelated to the erased concept.

This allows assessment of how much the model still represents the concept it was meant to erase and how much nearby concepts are unintentionally affected.

3.4 Analyzing SAE Latents

For each test prompt, activations at the U-Net layer `up_blocks.1.attentions.1` are extracted from the unlearned model and passed through the SAE. The resulting latent activations are averaged over spatial and temporal dimensions, resulting in a single latent vector per prompt, which serves as the SAE's representation of that prompt.

3.4.1 Concept Directions

Given latent vectors for different prompt groups, concept directions are computed using a difference-of-means method. For a target concept with mean latent μ_{target} and a control concept (e.g., generic prompts) with mean latent μ_{ctrl}, the direction is

$$d_{\text{target}} = \mu_{\text{target}} - \mu_{\text{ctrl}}$$

This subtracts out generic structure and isolates what is most specific to the target concept. Directions are calculated for each prompt group of interest (i.e., target, related, and generic concepts), and cosine similarity between directions is used to determine their alignment.

3.4.2 Latent Projections Along Directions

To measure how strongly a given prompt corresponds a concept, each prompt's latent vector z is projected onto a concept direction d using a dot product $z \cdot d$. Larger projections indicate stronger alignment with that concept direction. For each prompt group, the distribution of projection values along a direction is visualized using kernel density estimates (KDEs). Comparing these distributions before and after unlearning, and across prompt categories, reveals how concept directions and latent representations change as a result of erasure.

4 Experimental Results

SAE latent analysis was performed with two different erased concepts. The first is "Van Gogh", which can be categorized as a style erasure, where outputs should not have an artistic style. The second is "cat", which involves an object removal, where outputs should not contain a certain object.

4.1 Artistic Style Removal

After erasing "Van Gogh" from Stable Diffusion and training an SAE on this modified model, three categories of test prompts were used:

1) Prompts containing "Van Gogh" (e.g., "Almond Blossoms by Vincent van Gogh")
2) Prompts about art, excluding Van Gogh (e.g., "A sunset over a beach, with the soft brushstrokes and pastel colors that characterized Monet's late work.")
3) Generic prompts from COCO-30k (e.g., "A room with blue walls and a white sink and door")

4.1.1 Getting Concept Directions

The direction of Van Gogh d_{VG} and of art d_{Art} were taken with the difference of means of their latents (μ_{VG} and μ_{Art} respectively) with the mean latents of the generic prompts μ_{gen}:

$$d_{\text{VG}} = \mu_{\text{VG}} - \mu_{\text{gen}}$$

$$d_{\text{Art}} = \mu_{\text{Art}} - \mu_{\text{gen}}$$

The cosine similarity between these two directions was 0.9947, indicating a high degree of similarity between the directions. The high cosine similarity implies that the concept of Van Gogh has been reduced to the concept of art after unlearning[2]. That is, prompts about Van Gogh are represented similarly to prompts about art, indicating that ESD effectively erased his unique style.

4.1.2 Latent Projections Along Concept Directions

Projections of the latents of prompts from the three groups along the Van Gogh and art directions are visualized as probability density estimates in Fig. 3. The similarity between the two plots shows the minimal difference between how the SAE represents these concepts, as expected from their high cosine similarity. The art prompts exhibit a wider density, showing higher variance. That is, some art prompts are close to Van Gogh style, while others are not. In contrast, generic prompts, which are unrelated to Van Gogh, have a narrower distribution.

4.2 Object Removal

After erasing "cat" from Stable Diffusion and training an SAE on this modified model, six categories of test prompts were used:

1) Cat (e.g., "A chef cat cooking a gourmet fish dish")
2) Tiger
3) Lion
4) Dog
5) Animal (e.g., "An elephant as a construction worker operating a crane"; prompts related to many different animals)
6) Generic (i.e. random, unrelated to animals)

4.2.1 Getting Concept Directions

As before, the direction of cat d_{Cat}, tiger d_{Tiger}, lion d_{Lion}, dog d_{Dog}, and animal d_{Animal} were calculated with the difference of means of their latents μ with the mean latents of the generic prompts μ_{gen}:

$$d_{\text{Cat}} = \mu_{\text{Cat}} - \mu_{\text{gen}}$$

$$d_{\text{Tiger}} = \mu_{\text{Tiger}} - \mu_{\text{gen}}$$

$$d_{\text{Lion}} = \mu_{\text{Lion}} - \mu_{\text{gen}}$$

$$d_{\text{Dog}} = \mu_{\text{Dog}} - \mu_{\text{gen}}$$

$$d_{\text{Animal}} = \mu_{\text{Animal}} - \mu_{\text{gen}}$$

In order to compare to a baseline, these directions were calculated both prior to unlearning[3] and after unlearning. The cosine similarities between these directions are shown in Table 1.

All cosine similarities increased after unlearning, indicating that internal representations of many animals have merged into a broader "animal" subspace, despite intent to only erase cat from the model. This is especially the case for the feline animals (cat, lion, and tiger), but is also unexpectedly the case for dogs.

2. Under the assumption that the cosine similarity of these directions is lower prior to unlearning, which could be verified by training an SAE on the original Stable Diffusion model.

3. Latents are taken from the SAE trained on the model with Van Gogh erased, which assumes that the impact of this erasure on animal representations is negligible.

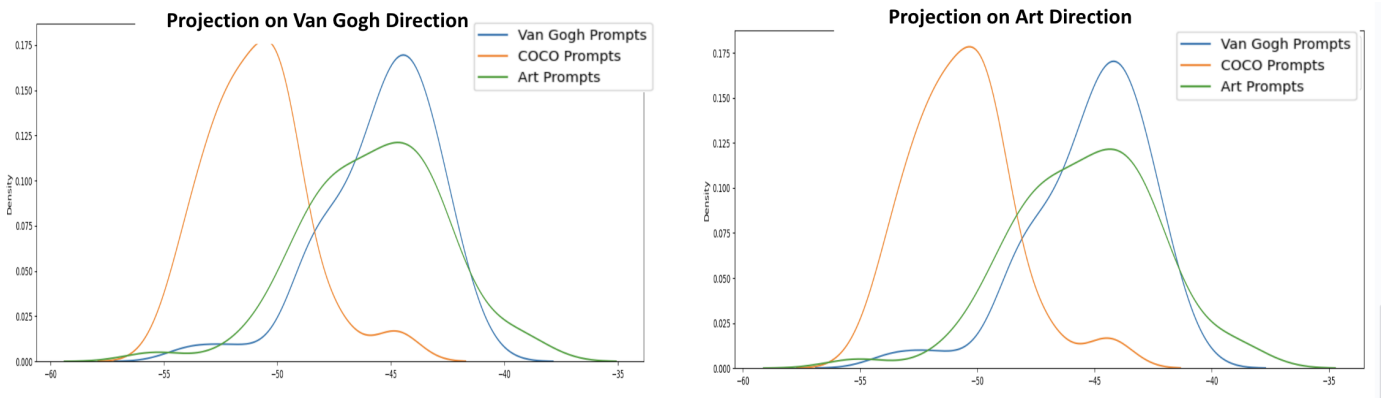

Fig. 3: Kernel Density Estimates (KDEs) of projections of the SAE latents of three groups of prompts onto the art and Van Gogh directions: prompts containing Van Gogh, prompts about art, and generic prompts from COCO-30k. Projection distributions further to the right are more related to the corresponding concept direction. Similarity between plots implies art and Van Gogh directions align very well.

TABLE 1: Cosine Similarity Between Animal-related Directions with Cat Unlearned

Directions	Before Unlearning	After Unlearning
Cat vs. Dog	0.6795	0.8322
Cat vs. Tiger	0.9261	0.9740
Cat vs. Lion	0.8707	0.9539
Lion vs. Tiger	0.9351	0.9723
Animal vs. Dog	0.5871	0.8688
Animal vs. Cat	0.9032	0.9773
Animal vs. Tiger	0.9351	0.9739
Animal vs. Lion	0.9204	0.9626

4.2.2 *Latent Projections Along Concept Directions*

Projections of the latents of prompts from each group along the cat and animal directions are shown in Fig. 4. Curves are wider and more similar to one another after unlearning, showing a convergence of various individual animal directions to a single general animal direction. This is evidence of neighbouring concepts being impacted by erasure; if cat were the only concept erased (as is intended), it would be the only curve flattened in the projections along the animal direction. This is because neighbouring concepts such as "tiger" and "lion" should still have unique representations after unlearning instead of becoming closer to the "animal" direction.

Furthermore, the plots for projections along the cat direction and the animal direction are more similar after unlearning, which provides evidence that the cat direction has become more similar to the animal direction, indicating successful erasure.

5 Discussion

5.1 Limitations

5.1.1 *SAE Capacity and Training Setup*

The SAE was trained using SAeUron under time and storage constraints. As a result, the training configuration deviated from their default setup: a smaller subset of activations was used, and the expansion factor (hidden dimensionality) was reduced by half. These design choices likely limited the expressive capacity of the SAE and its ability to capture detailed structure in the U-Net activations. These conclusions should therefore be interpreted as a lower bound on what a higher-capacity, fully trained SAE might reveal.

To assess reconstruction quality, images generated by the unlearned model with and without the SAE in the loop were compared. Concretely, a hook was attached to the `up_blocks.1.attentions.1` layer, activations were passed through the SAE, and reconstructed activations were fed back into the remaining U-Net. Visually, the resulting images were incoherent: for example, a "dog" prompt produced blobs of fur, eyes, and paws instead of a well-formed dog. Nevertheless, the outputs remained semantically related to the prompts, and the SAE training losses converged. This suggests that the SAE captured some meaningful structure, but not with enough fidelity to support image-space interventions.

5.1.2 *Single-Layer and Single-Method Focus*

Another limitation is that the analysis is restricted to a single U-Net layer and a single unlearning method. Concept representations in diffusion models are typically distributed across multiple layers and blocks; focusing only on `up_blocks.1.attentions.1` may miss residual knowledge encoded elsewhere. Similarly, all experiments use one ESD configuration per concept. It remains unclear how robust the observed effects are across layers, architectures, or alternative unlearning methods such as SFD or f-divergence-based approaches. A more complete experiment would involve multi-layer SAEs and a broader set of erasure techniques.

5.1.3 *Size of Experiment*

The experiments in this project were conducted on relatively small prompt sets and with a limited number of erased concepts. Consequently, the estimated concept directions, which are defined as differences of mean latents over each dataset, are subject to sampling noise and may not fully reflect the underlying representation geometry. A more exhaustive evaluation with larger, more diverse prompt collections would likely yield more stable and reliable directions.

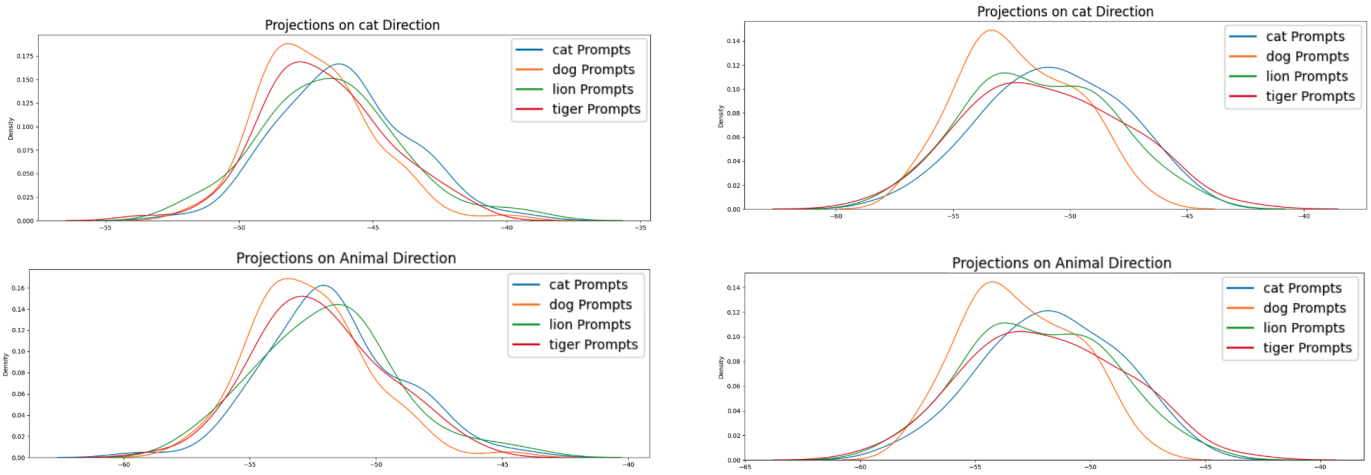

Fig. 4: KDEs of latent projections before (left) and after (right) unlearning 'cat' for multiple animal prompt categories. Right-shifted curves indicate stronger alignment with the corresponding direction; wider curves indicate higher variance across prompts. After unlearning, cat, lion, and tiger distributions widen, suggesting less consistent concept latents, implying successful cat unlearning with side effects on related animals.

For instance, the "animal" direction may be biased in this setting: before unlearning "cat", its cosine similarity with feline directions (cat, lion, tiger) is higher than with the dog direction. With more data and a more balanced prompt distribution across species, the animal direction might align more uniformly across related animal categories.

Despite these scale limitations, the observed patterns, such as the collapse of Van Gogh style onto a broader art subspace and the entanglement of feline directions after cat erasure, are consistent and interpretable. The results should therefore be viewed as an initial case study that motivates follow-up experiments with larger datasets and additional erasure examples to confirm, refine, or qualify the claims made here.

5.1.4 Latent Geometry Representing Concepts

Finally, the evaluation relies on cosine similarities between directions and projection distributions as a sign of concept presence. While this provides a representational view, it does not directly measure downstream harms or user-facing failures. For example, a concept direction might remain detectable in latent space even if it is no longer practically recoverable in images, or vice versa. Bridging this gap between latent metrics and behavioral or societal risk remains an open challenge.

5.2 Future Directions

This work focuses on using SAEs to probe side effects of unlearning, rather than on evaluating the robustness of erasure. In particular, it was not tested whether residual concept information can be actively recovered by adversarial prompting or targeted attacks. Integrating SAE-based analysis with adversarial evaluation, where an attacker is allowed to optimize prompts or latent manipulations, would provide a stronger notion of effective forgetting.

A natural extension is to combine the SAE-based concept directions with steering or activation editing. Given a latent direction associated with an erased concept, one could deliberately amplify that direction at inference time to test whether the unlearned model still contains recoverable traces of the target concept. Such robustness experiments would complement the current, more diagnostic analysis and provide a stronger guarantee of unlearning.

Beyond robustness, there are several possible directions. Training higher-capacity SAEs (larger expansion factors, deeper stacks), exploring convolutional or multi-layer SAEs tailored to vision architectures, and fitting SAEs jointly across multiple U-Net layers could improve reconstruction fidelity and interpretability. Comparing latent-space indicators across different unlearning methods would also help isolate algorithm-specific artifacts from general effects of concept erasure.

From a broader perspective, this line of work points toward representation-level probing as a component of responsible deployment. As unlearning becomes increasingly relevant for copyright, privacy, and safety regulation, tools that expose what a model still "knows" after erasure, rather than only what it visibly generates, may play an important role in providing transparency and calibrating trust in generative systems.

6 Conclusion

This paper introduced an SAE-based framework for evaluating the side effects of machine unlearning in text-to-image diffusion models. By training SAEs on models where specific concepts such as "Van Gogh" and "cat" were erased with ESD, then analyzing latent directions and projections for targeted, related, and generic prompts, this project explored how concept representations change post-unlearning. The results show that erased concepts are not simply removed: stylistic knowledge of Van Gogh largely collapses onto a broader "art" subspace, and object-level erasure of "cat" shifts and entangles nearby animal concepts such as lions and tigers. These findings provide empirical evidence of over-erasure and the potential for residual knowledge within internal representations, contributing an

alternate angle of evaluation compared to existing benchmarks.

From a mechanistic interpretability perspective, this work extends SAE analysis to post-unlearning and demonstrates that SAE latents can serve as a tool for identifying concept subspaces, their overlap, and how they are deformed by unlearning interventions. For the machine unlearning community, these results highlight that behavioral evaluations alone may be insufficient: erasing methods may still leave recoverable structure in activation space, or inadvertently distort neighboring concepts. Incorporating representation-level audits into unlearning benchmarks can help distinguish genuine forgetting from redistributed or concealed knowledge.

Societally, these findings underscore both the promise and the limits of unlearning as a governance tool for generative models. The AI industry increasingly looks to unlearning to address copyright, safety, and privacy concerns, but this analysis suggests that current techniques may give a false sense of erasure if not carefully evaluated. At the same time, tools like SAEs can make unlearning more transparent by revealing what a model still "knows" after erasure, supporting more trustworthy and accountable deployment of generative systems.

Ultimately, I hope this line of work will contribute to unlearning methods that are not only empirically effective, but also interpretable, auditable, and aligned with emerging societal and legal expectations for responsible AI.

REFERENCES

[1] B. Cywiński *et al.*, "Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders," *arXiv preprint arXiv:2501.18052*, 2025.

[2] L. Novello *et al.*, "A unified f-divergence framework for machine unlearning," *arXiv preprint arXiv:2509.21167*, 2025.

[3] T. Chen *et al.*, "Score forgetting distillation: A swift, data-free method for machine unlearning in diffusion models," *arXiv preprint arXiv:2409.11219*, 2024.

[4] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[5] N. George *et al.*, "The illusion of unlearning: The unstable nature of machine unlearning in text-to-image diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[6] H. Xuan *et al.*, "Verifying robust unlearning: Probing residual knowledge in unlearned models," *arXiv preprint arXiv:2504.14798*, 2025.

[7] I. Amara, A. I. Humayun, I. Kajic, Z. Parekh, N. Harris, S. Young, C. Nagpal, N. Kim, J. He, C. N. Vasconcelos, D. Ramachandran, G. Farnadi, K. Heller, M. Havaei, and N. Rostamzadeh, "Erasebench: Understanding the ripple effects of concept erasure techniques," *arXiv preprint arXiv:2501.09833*, 2025.

[8] T. Bricken, C. Wild, N. Elhage, N. Nanda, T. Henighan *et al.*, "Towards monosemanticity: Decomposing language models with dictionary learning," *Transformer Circuits*, 2023, anthropic interpretability team.

[9] A. Ijishakin *et al.*, "H-space sparse autoencoders for interpretable text-to-image models," *OpenReview*, 2024, openReview preprint. [Online]. Available: https://openreview.net/forum?id=zKhpACcPdb