ICML International Conference On Machine Learning

SAEURON

Warsaw University of Technology
IDEAS Research Institute

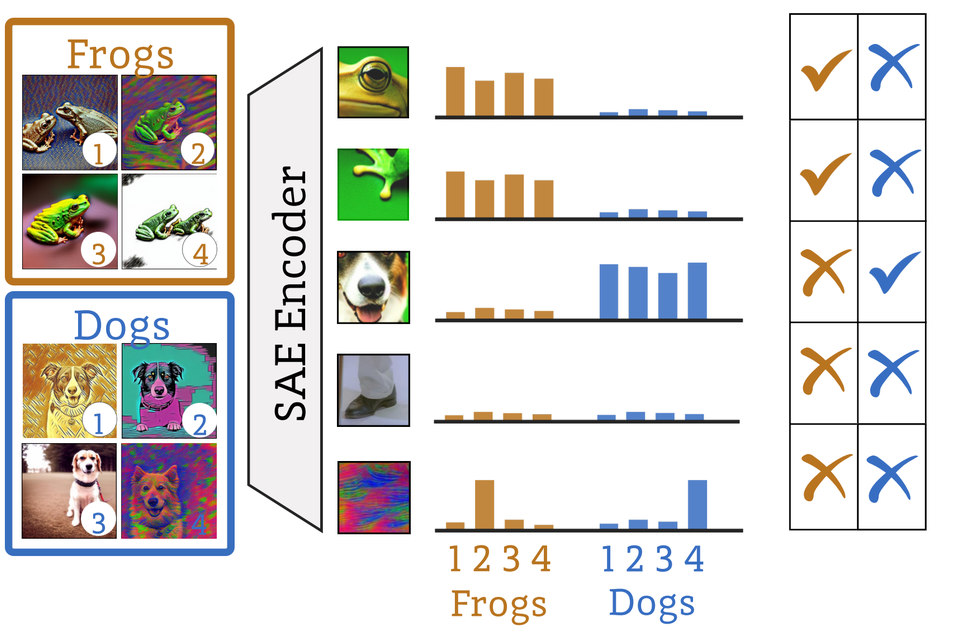
Interpretable Concept Unlearning in Diffusion Models with Sparse Autoencoders

Bartosz Cywiński, Kamil Deja

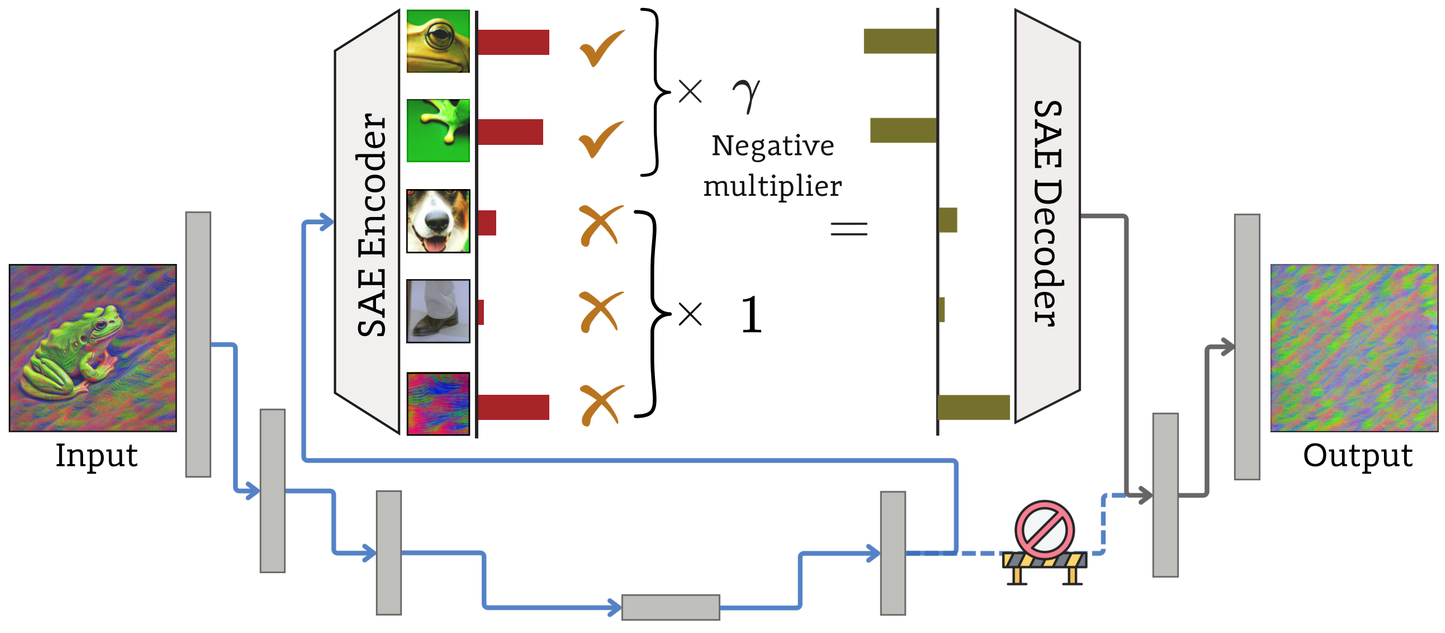
Paper

Recipe for Unlearning in Diffusion Models using SAEs 👨‍🍳

Select features based on importance



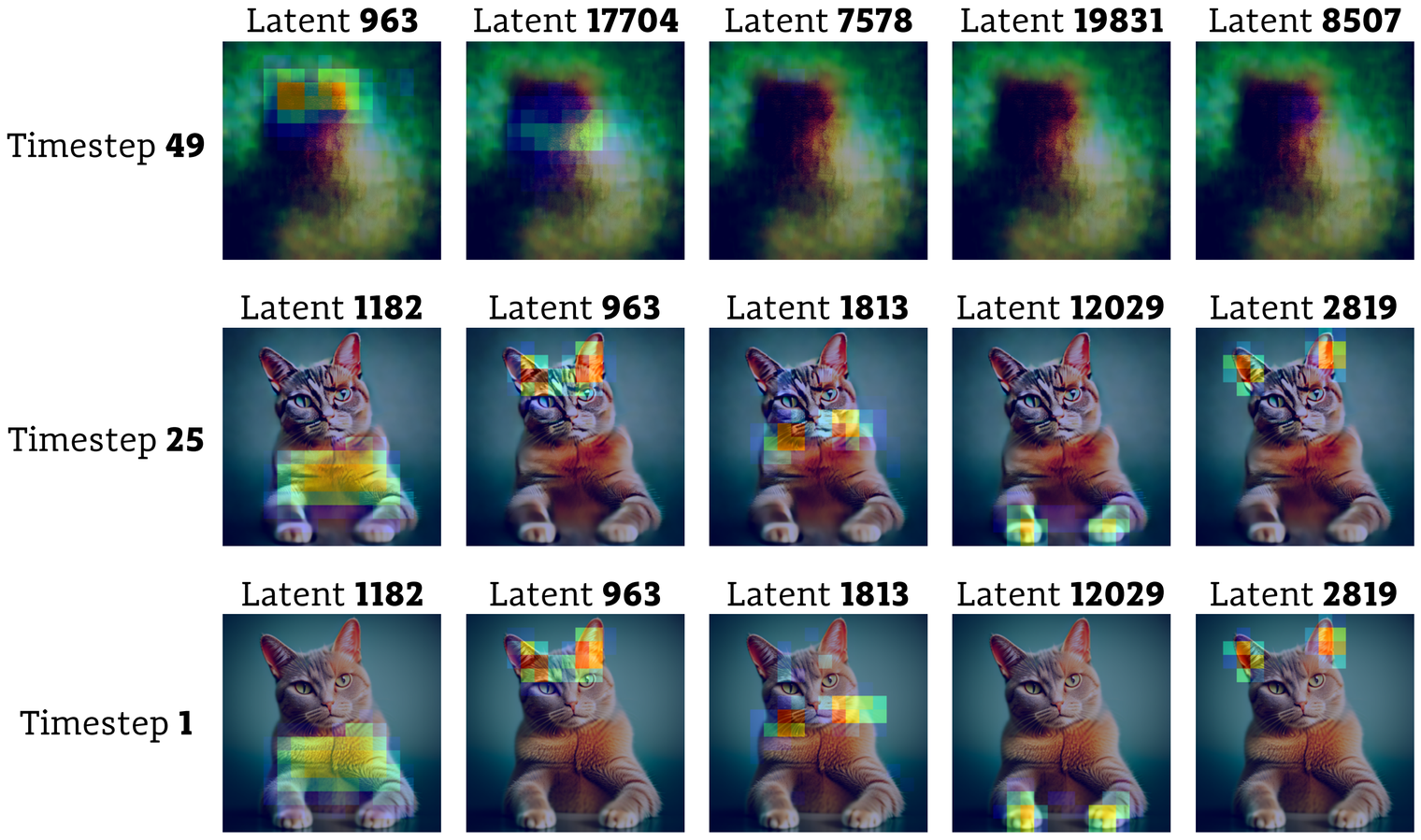
Remove selected features during the inference



Test prompt: "An image of Frogs in Color Fantasy style."

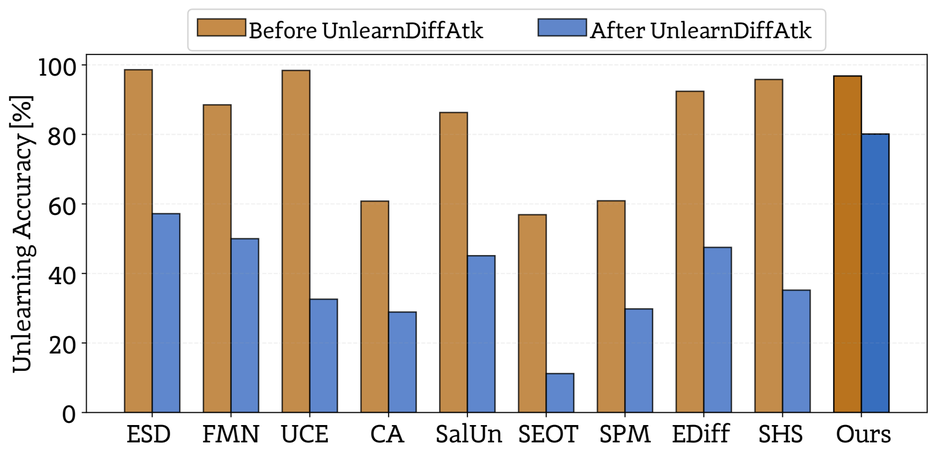
Interpretable features

We can visualize latent activations to see which features are selected at each denoising timestep.



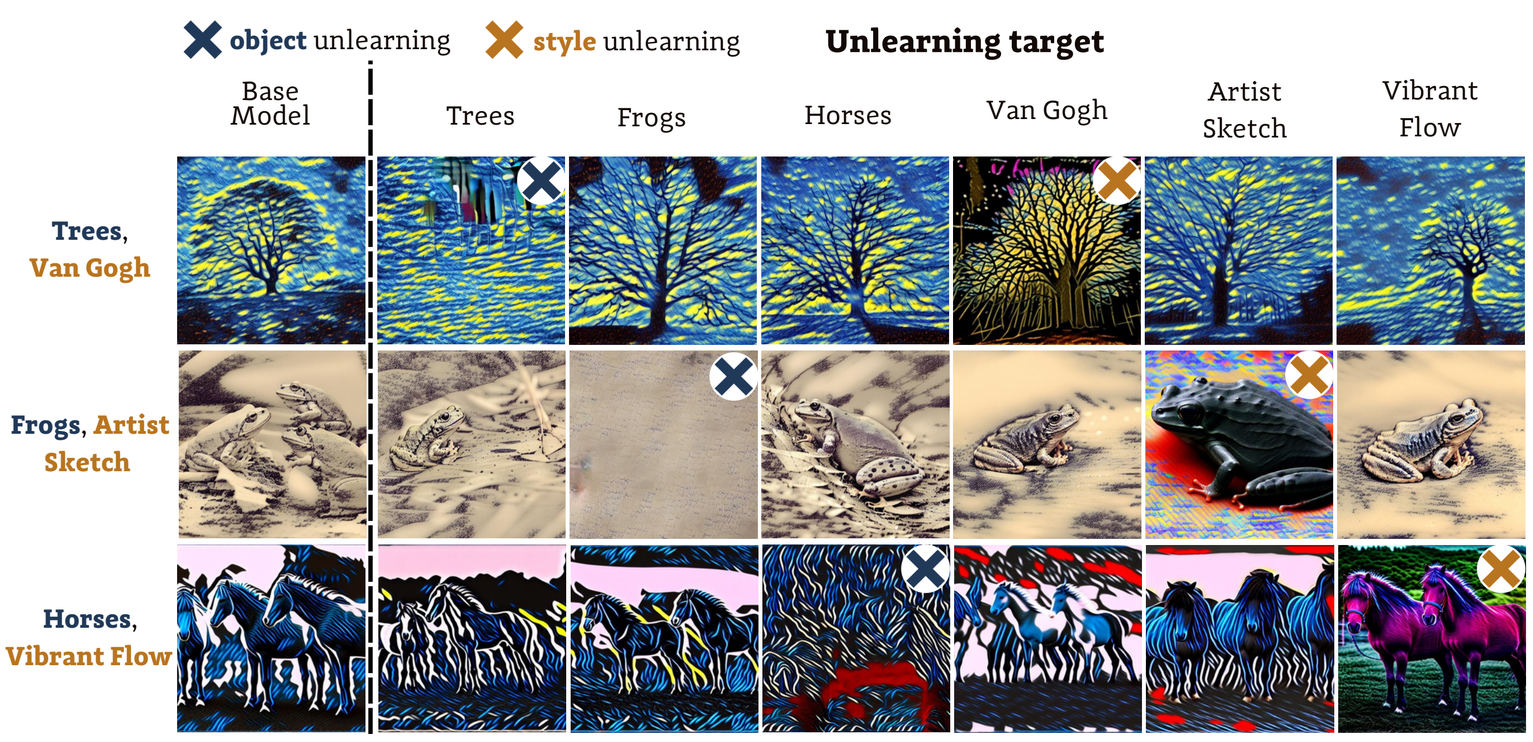
Robustness against adversarial attacks

Our approach selects and removes the correct features, even with adversarial prompts as input.



Precise unlearning

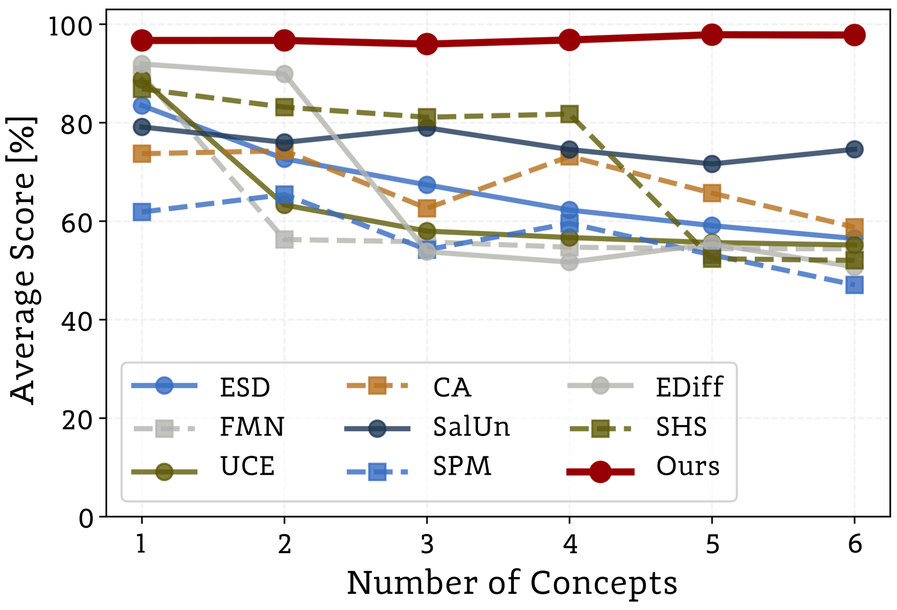
Features targeted for unlearning are highly concept-specific, so the overall generation capabilities of the model are preserved.



Test prompt: "An image of **{object}** in **{style}** style."

Multiple concept unlearning

All you need to do is select more features for unlearning! No performance drop!



Actionable Interpretability Workshop @ ICML 2025