

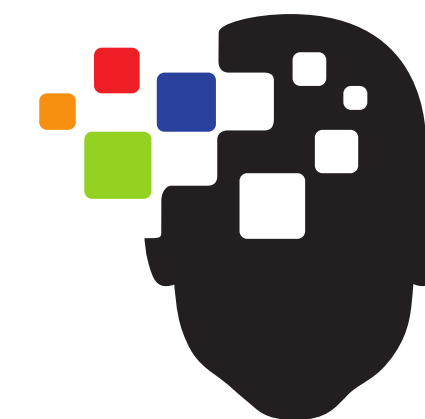


UNIVERSITÀ
DI TORINO



Politecnico
di Torino

INTESA
SANPAOLO



ICML

International Conference
On Machine Learning

SAEmnesia: Erasing Concepts in Diffusion Models with Supervised Sparse Autoencoders



Enrico Cassano*



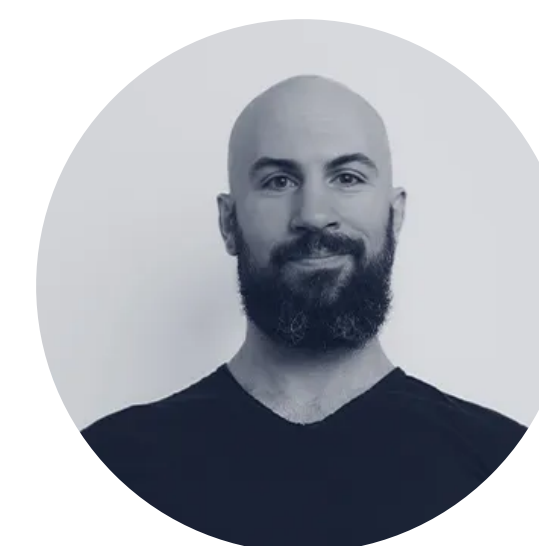
Riccardo Renzulli*



Marco Nurisso



Mirko Zaffaroni



Alan Perotti



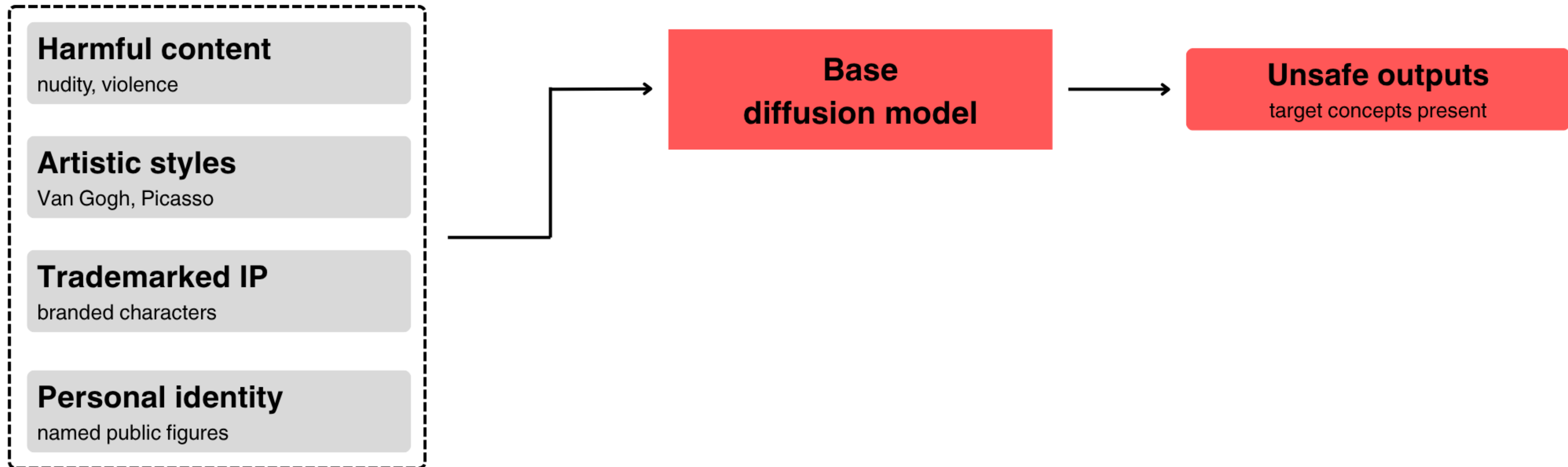
Marco Grangetto

*Equal contribution

Machine Unlearning

Generative models (text/images) can inadvertently generate **harmful** or **undesirable content**, raising significant ethical and safety concerns.

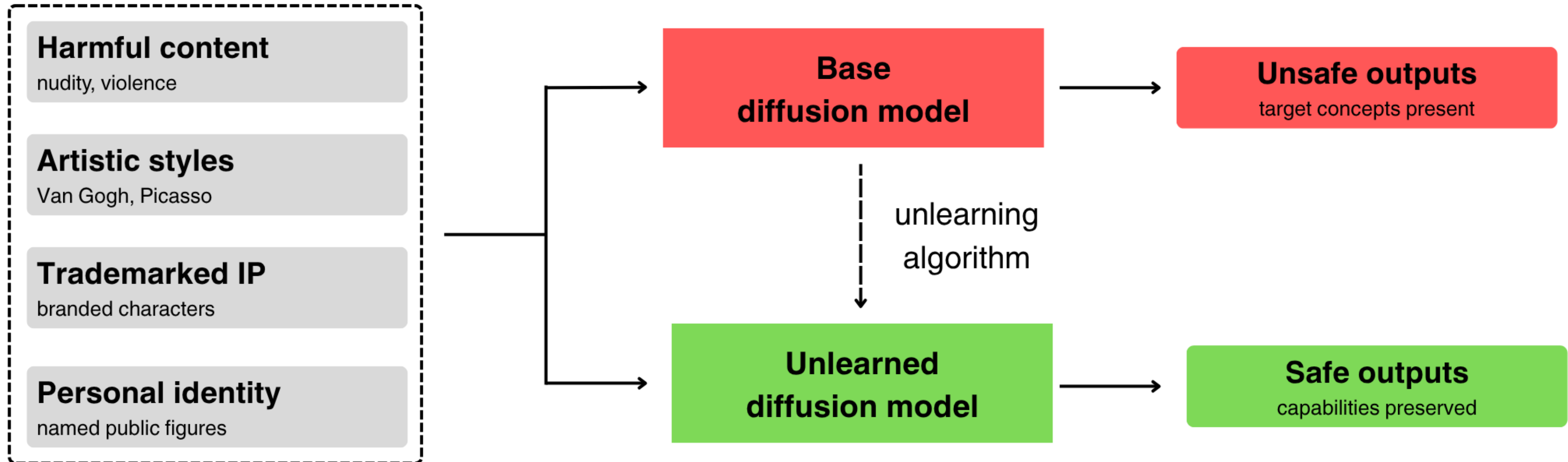
Concepts to erase



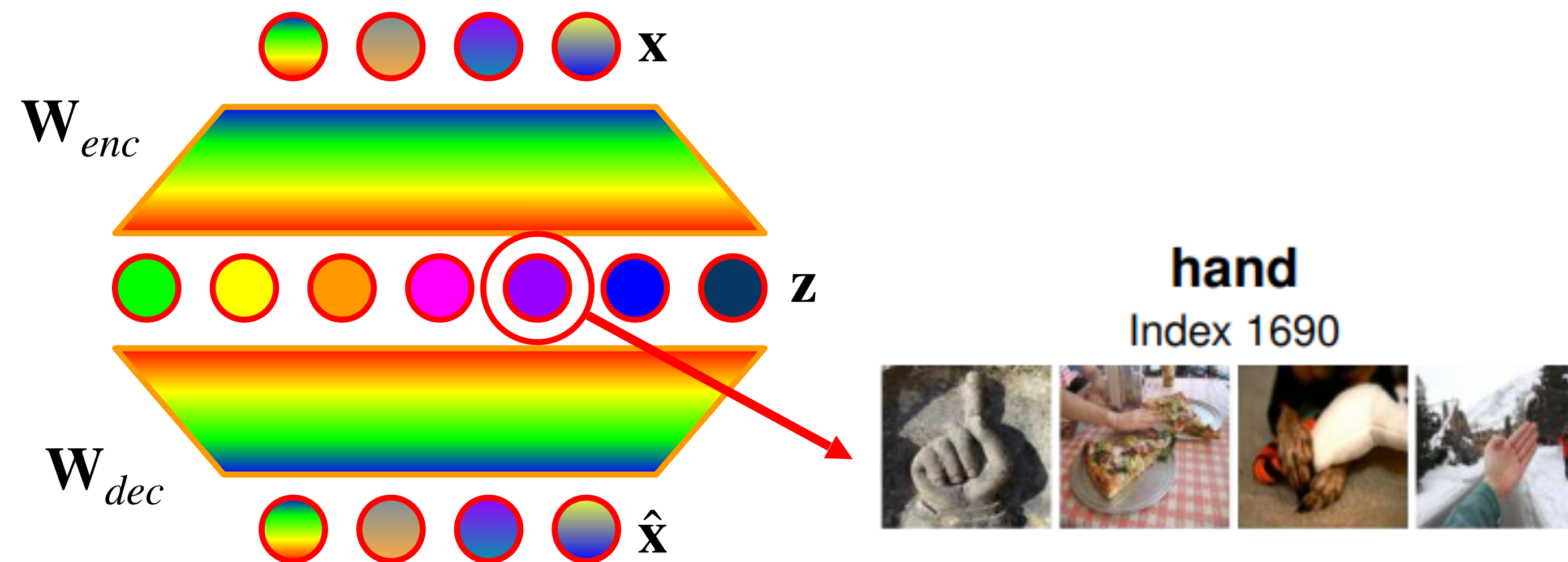
Machine Unlearning

Generative models (text/images) can inadvertently generate **harmful** or **undesirable content**, raising significant ethical and safety concerns.

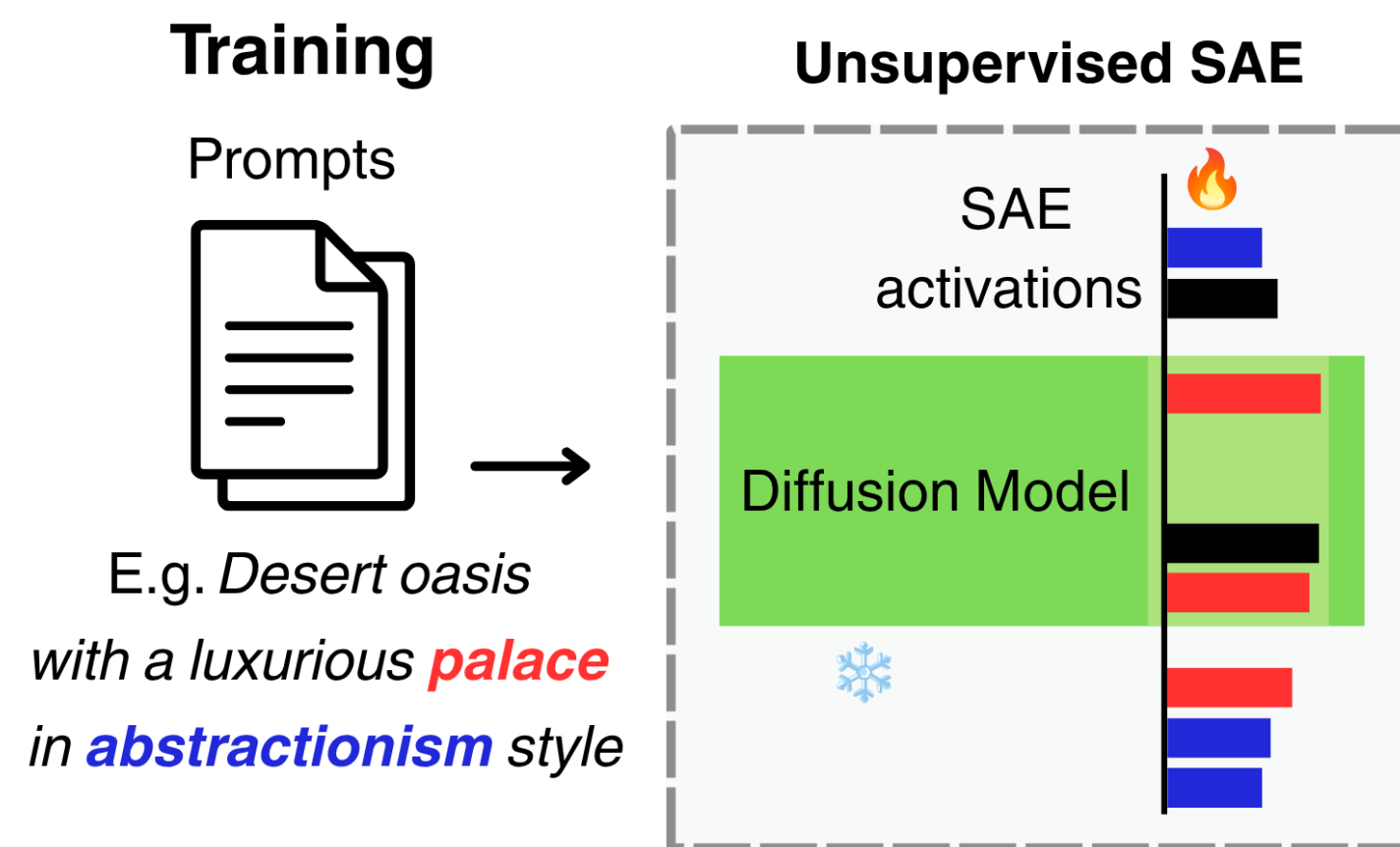
Concepts to erase



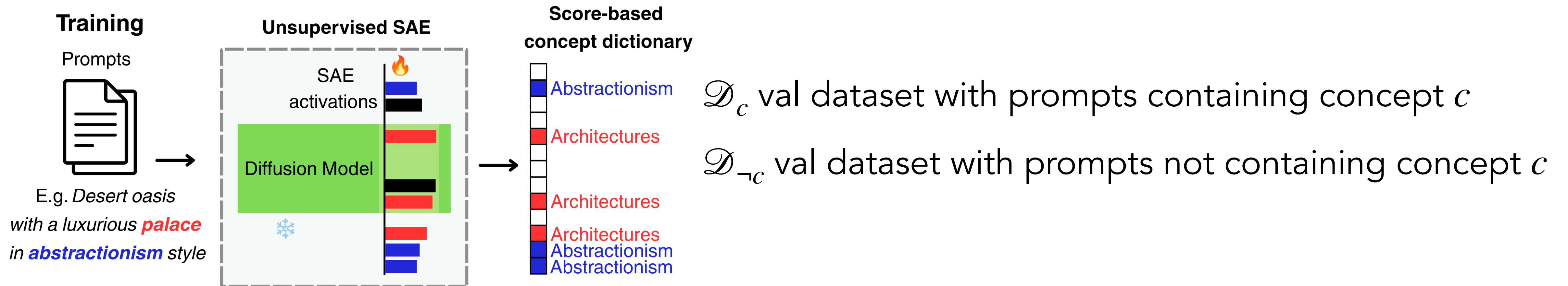
Can we exploit sparse autoencoders (SAEs) to unlearn concepts in diffusion models?



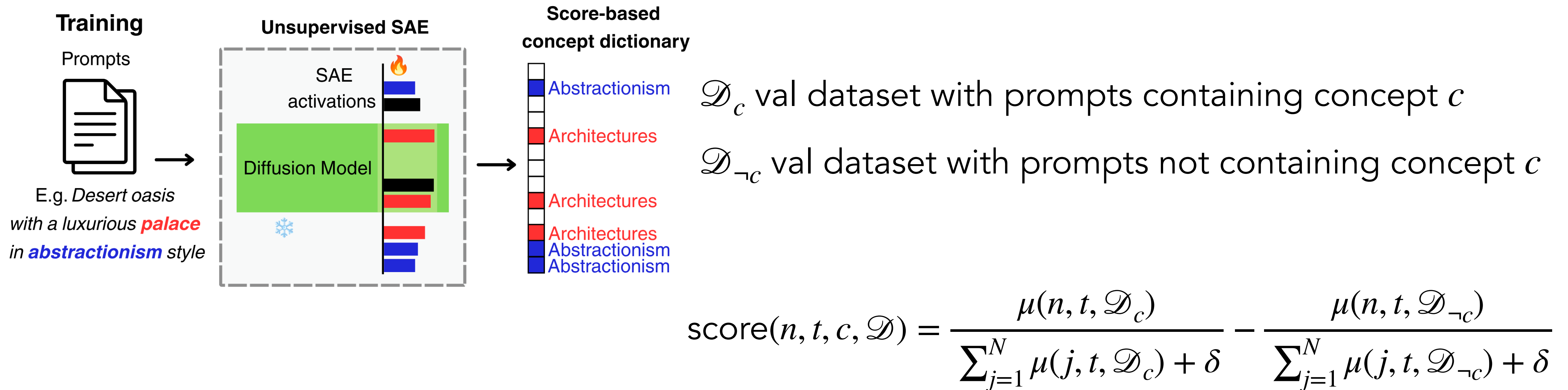
Mechanistic unlearning in diffusion models: SAeUron



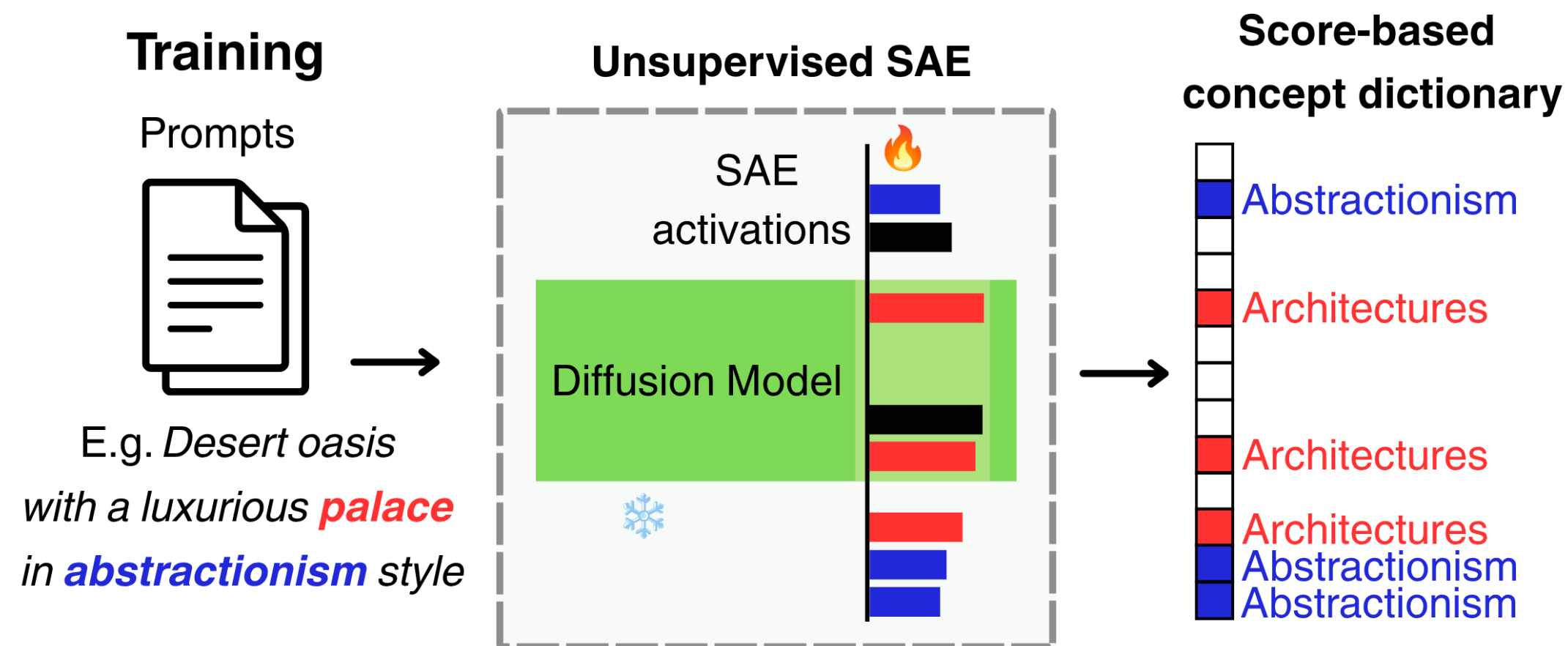
Mechanistic unlearning in diffusion models: SAeUron



Mechanistic unlearning in diffusion models: SAeUron



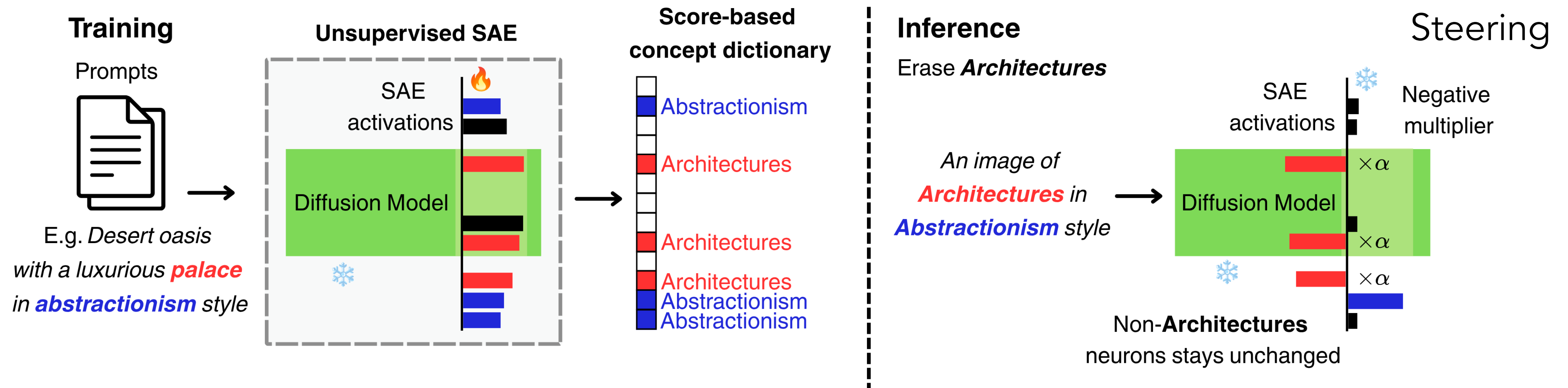
Mechanistic unlearning in diffusion models: SAeUron



$$\mathcal{F}_c := \{i \mid i \in \text{Top-}\tau_c(\{\text{score}(i, t, c, \mathcal{D})\})\}$$

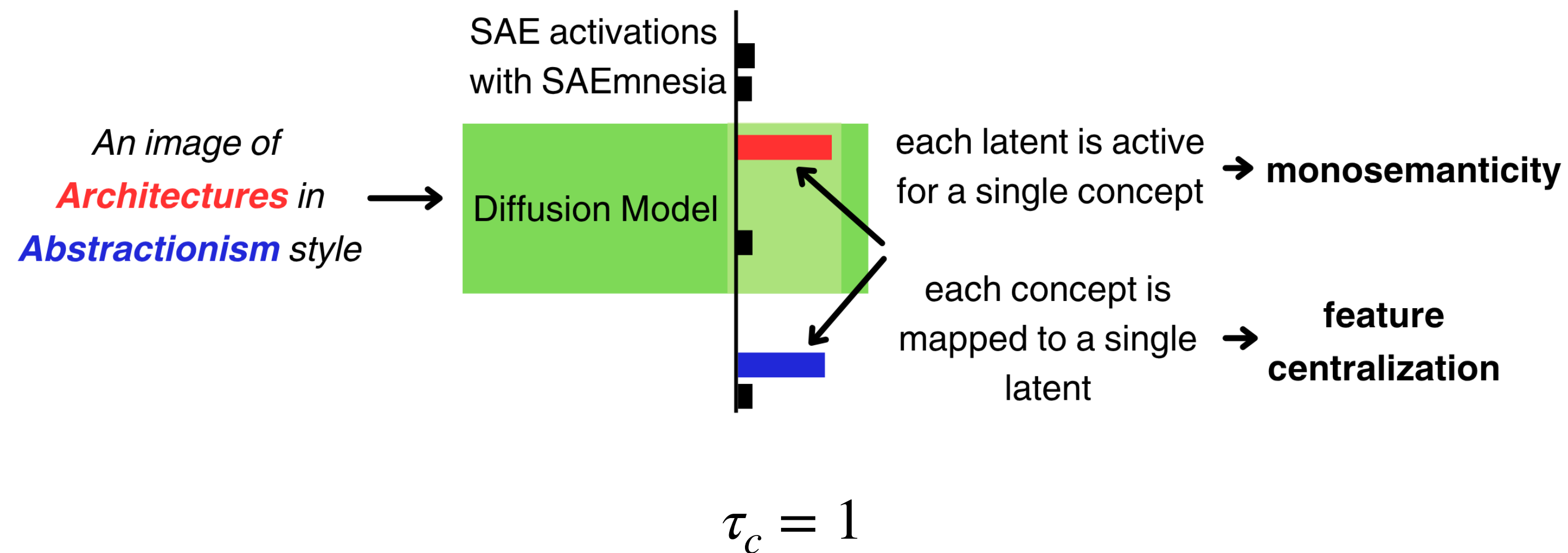
The number τ_c of top-scoring latents selected for concept c is a hyperparameter that needs to be carefully tuned

Mechanistic unlearning in diffusion models: SAeUron

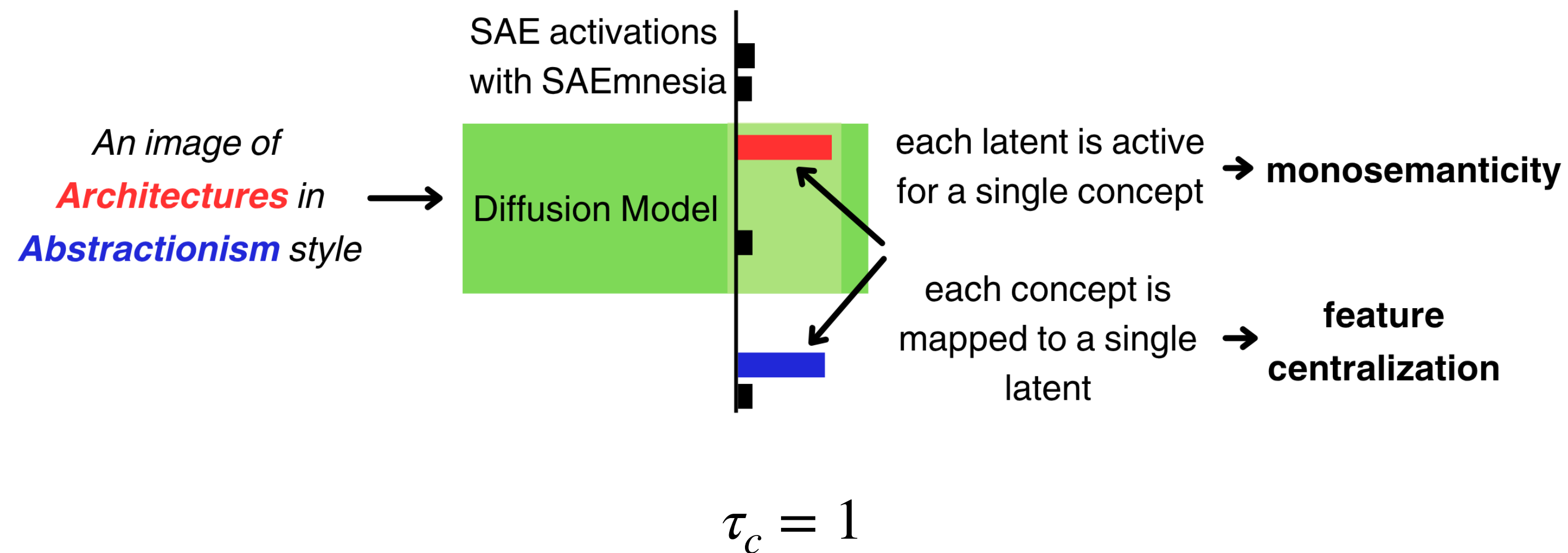


Since we know in advance the concepts to remove, we can train SAE in a **supervised** way to increase **feature centralization**

Since we know in advance the concepts to remove, we can train SAE in a **supervised** way to increase **feature centralization**

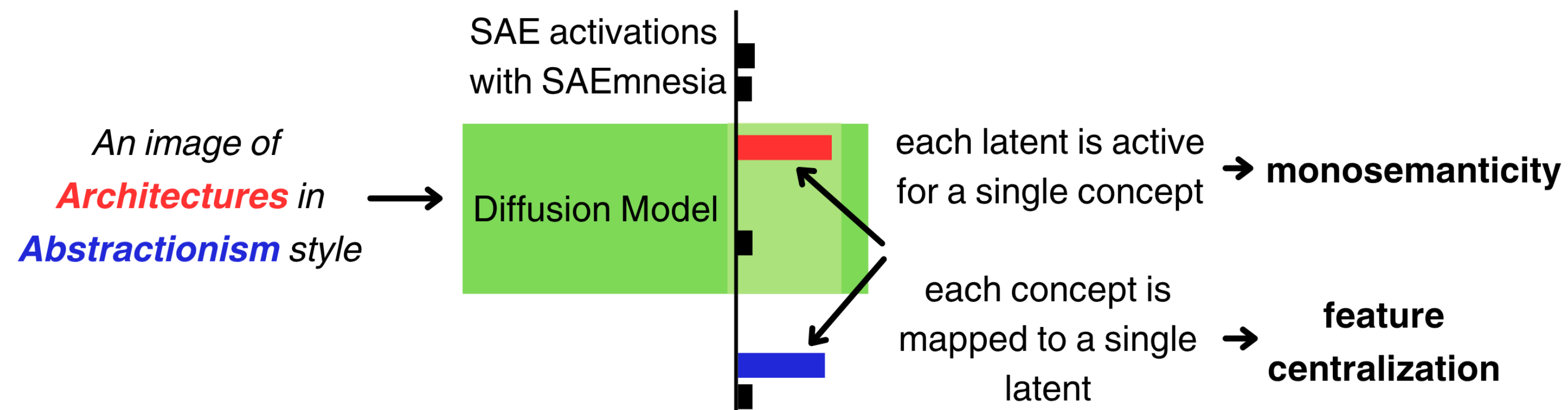


Since we know in advance the concepts to remove, we can train SAE in a **supervised** way to increase **feature centralization**



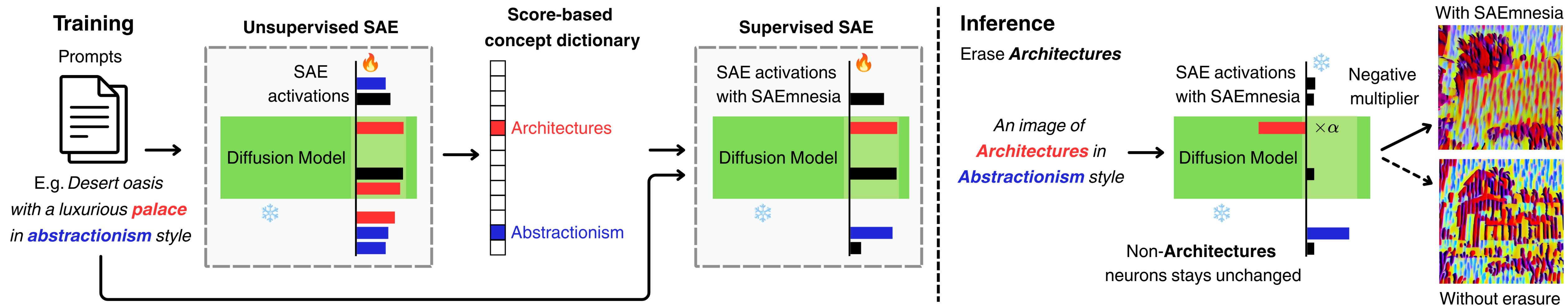
We don't need extensive hyperparameters search

Since we know in advance the concepts to remove, we can promote **disentanglement** across multiple **macro-categories of concepts**

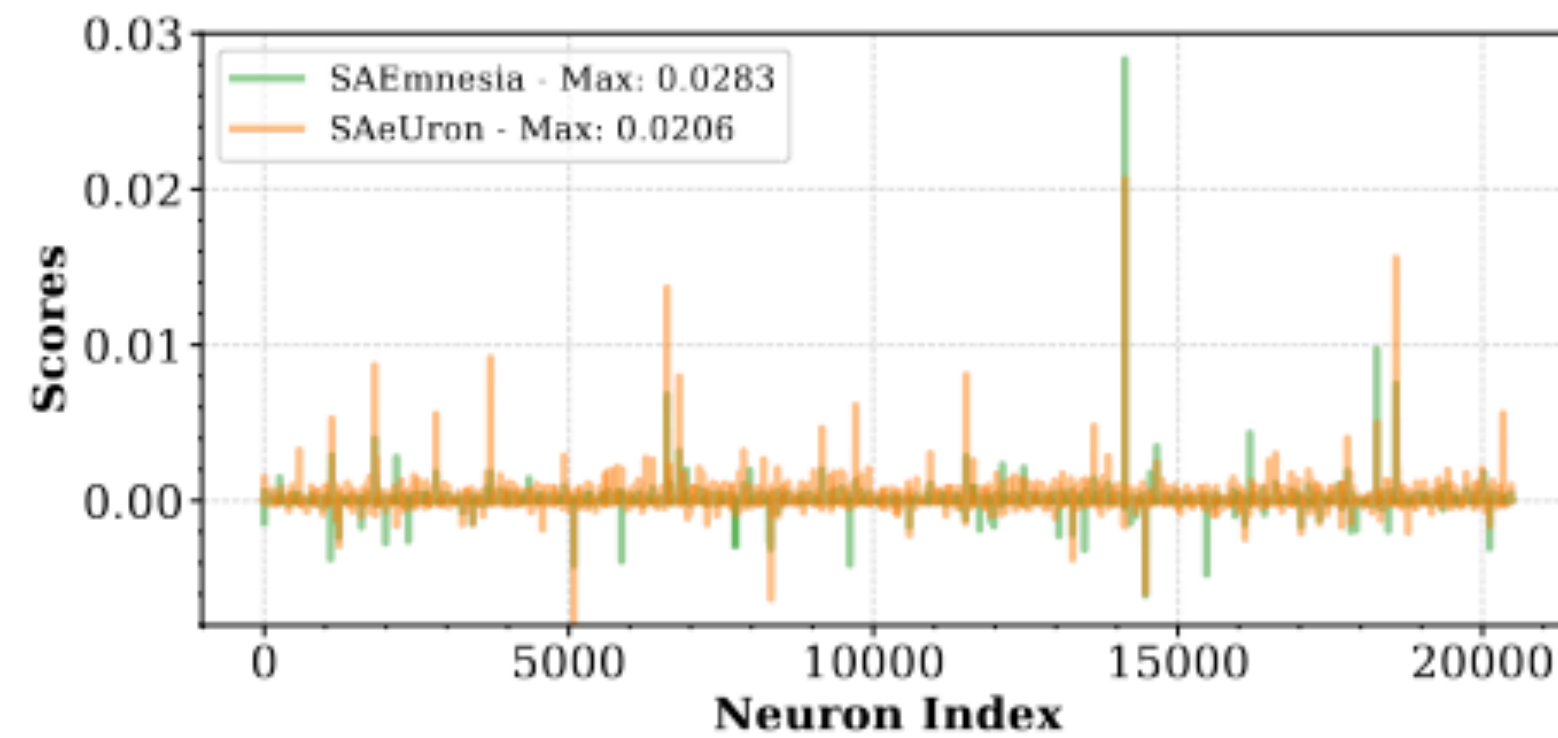


E.g. Decorrelating objects and style neurons

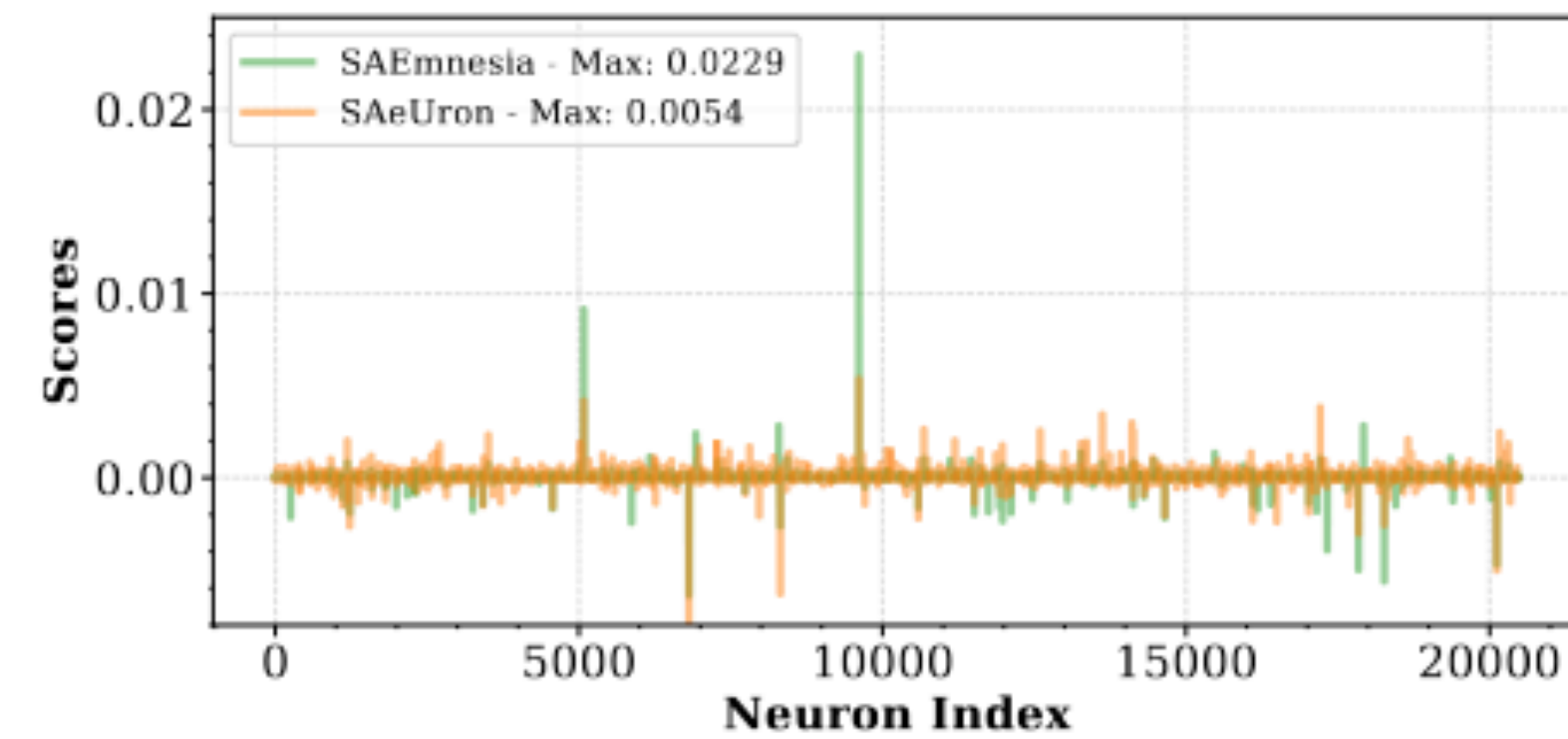
Mechanistic unlearning in diffusion models: SAEEmnesia



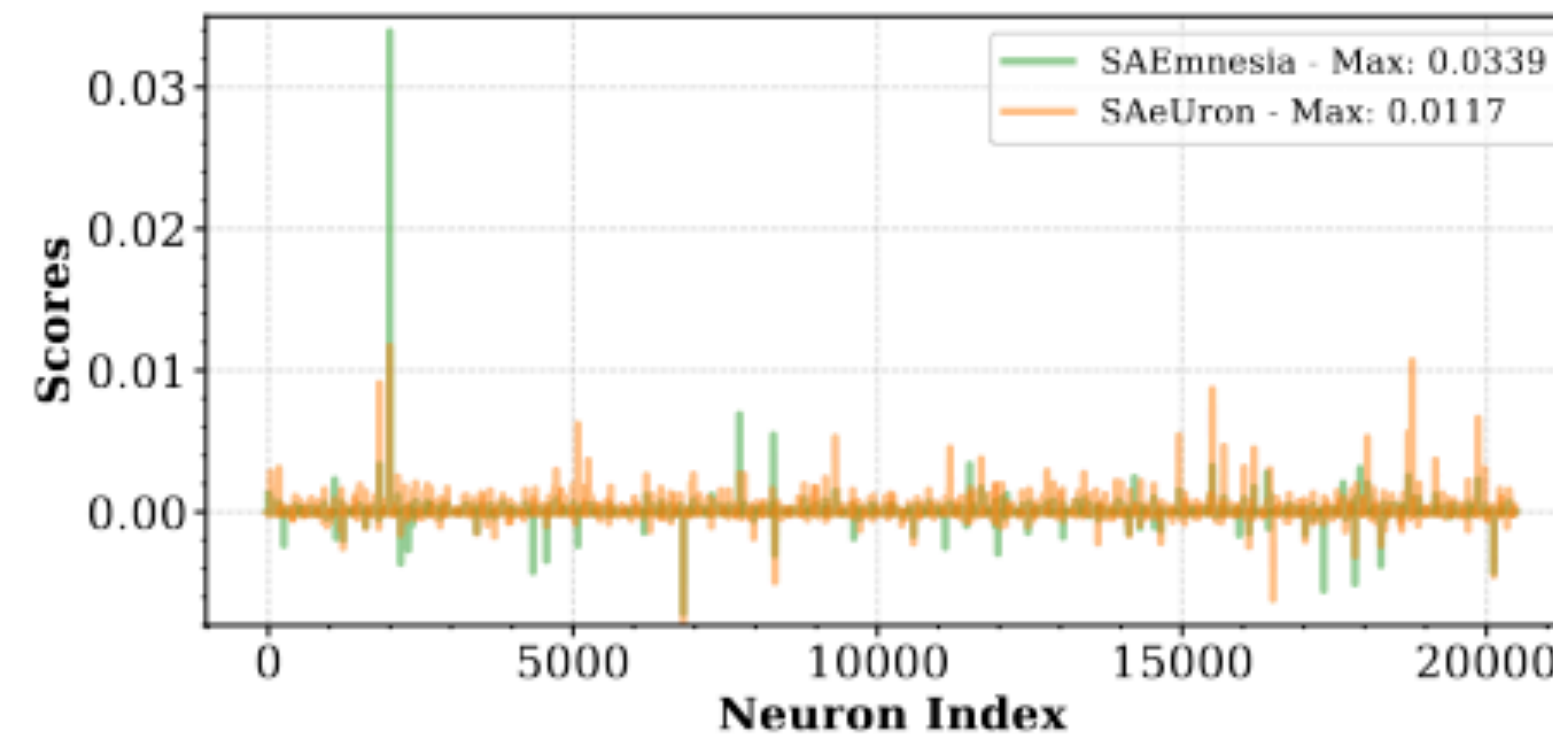
Mechanistic unlearning in diffusion models: SAEmnnesia



(b) Rabbits



(a) Architectures



(c) Sea

SAEmnesia mitigates feature splitting in SAEUron

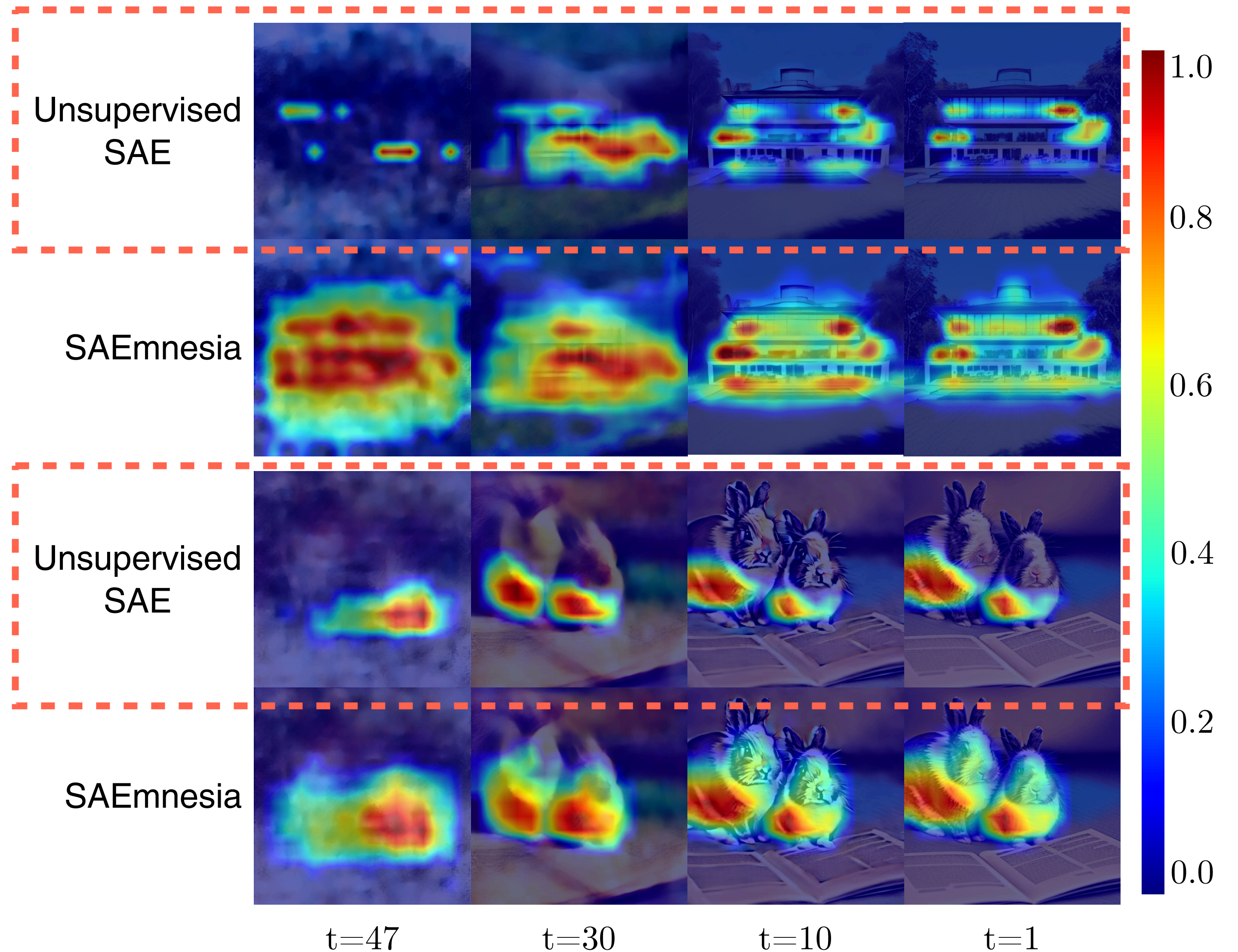
Mechanistic unlearning in diffusion models: SAEEmnesia

Visualization of patches for **SAE latent with highest score** for given concept

Unsupervised SAE top scoring latent is **spatially incomplete** and **temporally inconsistent**



No single latent captures the full concepts, behavior consistent with feature splitting

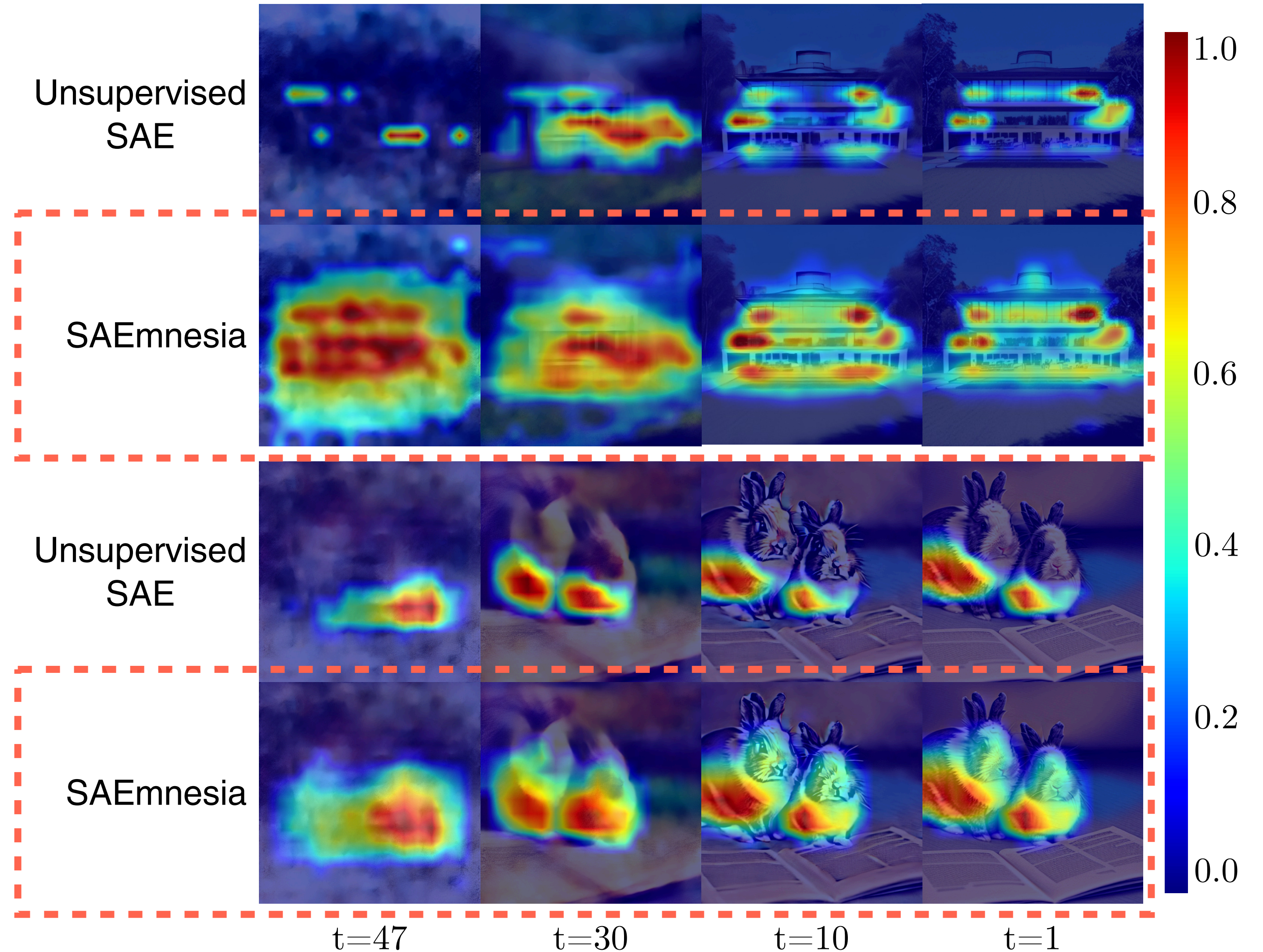


Mechanistic unlearning in diffusion models: SAEEmnesia

SAEmnesia provides **concentrated representation**, where object-level semantics are encoded in a **single, highly interpretable latent**



More precise concept **manipulation** and **unlearning** through targeted interventions



Mechanistic unlearning in diffusion models: SAEtnnesia

- **Unlearning Accuracy (UA)** → proportion of samples from target concept prompts that are misclassified (i.e., successful unlearning)
- **In-domain Retain Accuracy (IRA)** → measuring classification accuracy on retained concepts within the same domain
- **Cross-domain Retain Accuracy (CRA)** → assessing accuracy on concepts from different domains (i.e., style accuracy during object unlearning)

Mechanistic unlearning in diffusion models: SAEmnnesia

- **Unlearning Accuracy (UA)** → proportion of samples from target concept prompts that are misclassified (i.e., successful unlearning)
- **In-domain Retain Accuracy (IRA)** → measuring classification accuracy on retained concepts within the same domain
- **Cross-domain Retain Accuracy (CRA)** → assessing accuracy on concepts from different domains (i.e., style accuracy during object unlearning)

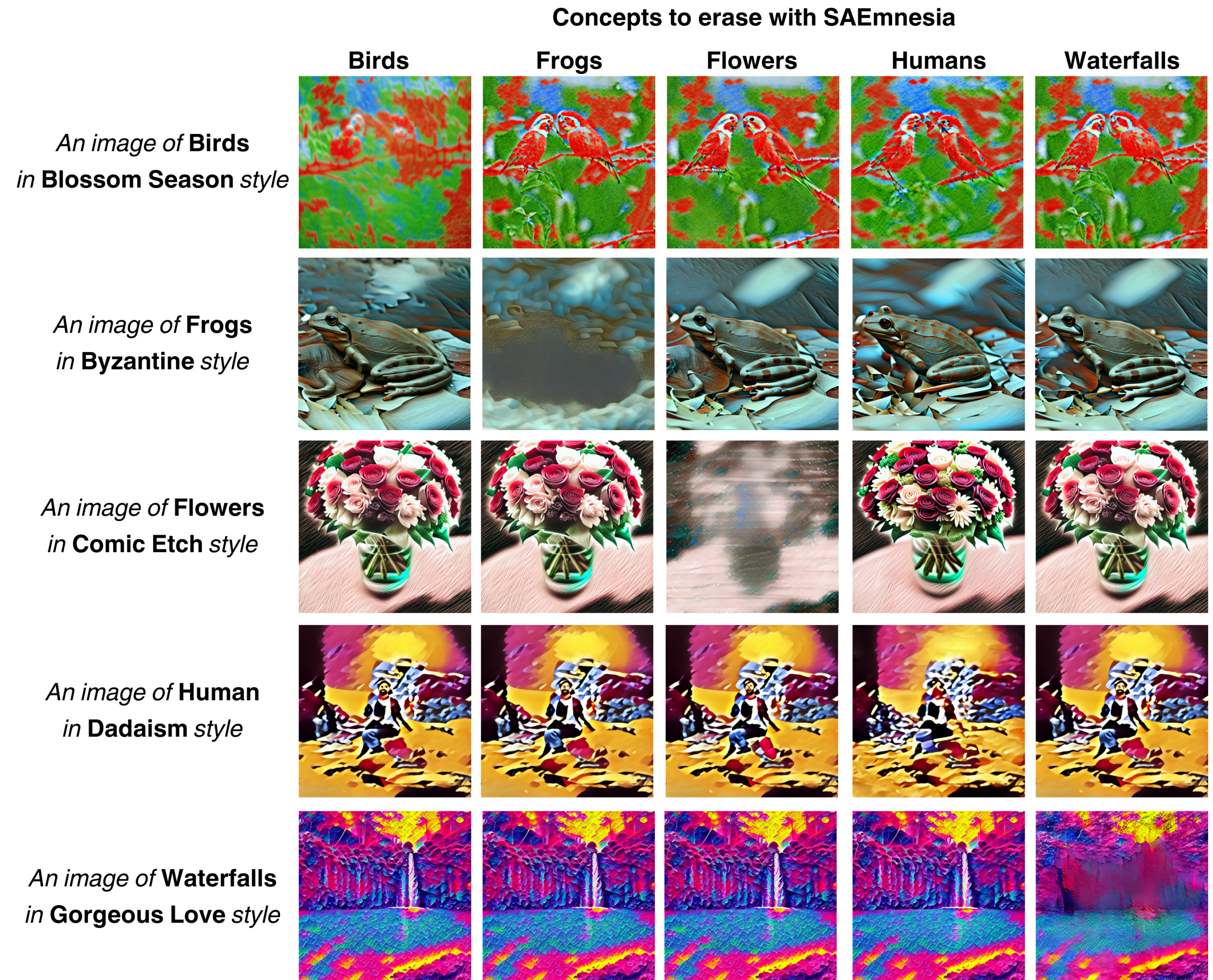
Method	Effectiveness						Efficiency			
	Style Unlearning			Object Unlearning			Avg. (↑)	FID (↓)	Memory (GB) (↓)	Storage (GB) (↓)
	UA (↑)	IRA (↑)	CRA (↑)	UA (↑)	IRA (↑)	CRA (↑)				
ESD (Gandikota et al., 2023)	98.58	80.97	93.96	92.15	55.78	44.23	77.61	65.55	17.8	4.3
FMN (Zhang et al., 2024a)	88.48	56.77	46.60	45.64	90.63	73.46	66.93	131.37	17.9	4.2
UCE (Gandikota et al., 2024)	<u>98.40</u>	60.22	47.71	<u>94.31</u>	39.35	34.67	62.45	182.01	<u>5.1</u>	1.7
CA (Kumari et al., 2023)	60.82	96.01	92.70	46.67	90.11	81.97	78.05	54.21	10.1	4.2
SalUn (Fan et al., 2023)	86.26	90.39	95.08	86.91	96.35	99.59	<u>92.43</u>	61.05	30.8	4.0
SEOT (Li et al., 2024)	56.90	94.68	84.31	23.25	<u>95.57</u>	82.71	72.91	62.38	7.34	0.0
SPM (Lyu et al., 2024)	60.94	92.39	84.33	71.25	90.79	81.65	80.23	59.79	6.9	0.0
EDiff (Wu et al., 2024)	92.42	73.91	98.93	86.67	94.03	48.48	82.41	81.42	27.8	4.0
SHS (Wu & Harandi, 2024)	95.84	80.42	43.27	80.73	81.15	67.99	74.90	119.34	31.2	4.0
SAeUron (Cywiński & Deja, 2025)	95.80	99.10	99.40	87.16	85.57	74.14	90.10	62.69	2.8	<u>0.2</u>
SAEMNESIA (ours)	96.60	<u>98.67</u>	<u>99.30</u>	94.65	91.39	<u>88.48</u>	94.85	<u>56.15</u>	2.8	<u>0.2</u>

Results on UnlearnCanvas

Mechanistic unlearning in diffusion models: SAEEmnesia

Qualitative results with concept multipliers of $\gamma_c = -1$

We restrict the application of SAEEmnesia to the final 25 denoising steps to reduce artifacts





Thanks for watching!

Paper



Code



Enrico Cassano enrico.cassano@unito.it

Riccardo Renzulli riccardo.renzulli@unito.it