

---

# NeurIPS Supplemental Materials

---

Anonymous Author(s)

Affiliation

Address

email

## 1 Licenses and Third-Party Assets

We acknowledge the use of the following models, datasets, and codebases in our experiments:

- **Stable Diffusion v1.4:** Released by CompVis under the CreativeML Open RAIL-M license.
- **CLIP ViT-B/32:** Released by OpenAI under the MIT license.
- **ResNet-50 Classifier (ImageNet/Imagenette):** Pretrained models used under the BSD 3-Clause (torchvision) or Apache 2.0 (fastai Imagenette) licenses.
- **UnlearnDiffAtk:** We use software provided by the OPTML Group under the MIT License. Copyright © 2023 OPTML Group.
- **Hugging Face Diffusers and Transformers libraries, includes Textual Inversion:** Used under the Apache 2.0 license.

## 2 Code and Data Availability

We release all code, data, and scripts used in our experiments at:

<https://github.com/koyfish445/when-are-concepts-erased-anon>

The repository includes:

- All prompt datasets and concept lists used for erasure evaluation.
- Implementations of optimization-based, context-based, and trajectory-based probes.
- CLIP scoring, classification evaluation scripts, and visualizations.

We hope this release facilitates future work in evaluating and developing concept erasure methods for diffusion models.

## 2 Model Access

To facilitate reproducibility, we have released all 91 concept-erased diffusion models covering 13 concepts across the 7 erasure methods.

We will make these models publicly available after paper decision.

## 3 Full Results with Standard Deviations

We include here the full quantitative results with standard deviations across runs, complementing Table 1, 2, and 3 from the main paper. These tables report the mean and standard deviation for

both CLIP similarity scores and classification accuracies across multiple erasure evaluation settings. Including standard deviation helps illustrate the consistency and robustness of each erasure method under different probing strategies.

Eval Metric	Base	GA	UCE	ESD-x	ESD-u	TaskVec	STEREO	RECE
<b>Erased Concepts (↓)</b>								
CLIP	–	24.3 ± 2.7	22.4 ± 5.2	21.1 ± 4.1	20.9 ± 3.4	23.1 ± 3.0	<b>19.6 ± 2.3</b>	21.2 ± 4.0
<b>Text Inversion (↓)</b>								
CLIP	–	<b>22.7 ± 2.5</b>	30.7 ± 2.0	30.6 ± 2.4	28.0 ± 3.4	25.1 ± 2.6	24.5 ± 2.9	29.2 ± 2.8
<b>UnlearnDiffAtk</b>								
CLIP	–	<b>26.0 ± 2.2</b>	28.3 ± 3.2	28.7 ± 2.2	27.8 ± 2.8	27.1 ± 1.7	26.1 ± 2.8	27.9 ± 2.3
<b>Unrelated Concepts (↑)</b>								
CLIP	–	28.8 ± 2.8	<b>31.2 ± 2.3</b>	30.8 ± 2.5	30.7 ± 3.3	29.4 ± 2.6	29.0 ± 3.0	30.5 ± 2.7
<b>Inpainting (↓)</b>								
CLIP	29.5 ± 2.2	24.8 ± 2.4	26.9 ± 3.3	26.8 ± 3.1	23.9 ± 3.1	25.9 ± 2.6	<b>22.7 ± 3.0</b>	26.3 ± 2.7
<b>Diffusion Completion <math>t = 5</math> (↓)</b>								
CLIP	30.2 ± 2.1	24.0 ± 2.4	27.7 ± 2.8	27.2 ± 3.1	26.9 ± 2.9	<b>23.8 ± 2.4</b>	23.9 ± 2.7	28.8 ± 2.5
<b>Diffusion Completion <math>t = 10</math> (↓)</b>								
CLIP	30.2 ± 2.1	<b>24.5 ± 2.3</b>	29.6 ± 2.3	28.7 ± 2.9	27.5 ± 2.8	24.9 ± 2.3	27.8 ± 2.6	28.8 ± 2.5

Table 1: CLIP scores (mean ± std) across concept erasure methods. Lower scores (↓) indicate better erasure of the target concept, while higher scores (↑) reflect stronger retention of unrelated concepts. Rows cover adversarial and in-context evaluations including inpainting and diffusion completion at denoising steps  $t = 5$  and  $t = 10$ . For each evaluation, 13 concepts were assessed for each model with 100 images generated per concept.

Eval Metric	Base	GA	UCE	ESD-x	ESD-u	TaskVec	STEREO	RECE
<b>Erased Concepts (↓)</b>								
Acc. (%)	–	0.6 ± 0.48	4.4 ± 1.1	3.6 ± 1.3	1.0 ± 0.69	2.2 ± 1.0	<b>0.0 ± 0.00</b>	4.0 ± 1.2
<b>Text Inversion (↓)</b>								
Acc. (%)	–	<b>0.6 ± 0.59</b>	71.2 ± 2.3	65.9 ± 2.9	31.8 ± 3.6	6.2 ± 1.8	6.3 ± 1.6	58.2 ± 3.1
<b>UnlearnDiffAtk</b>								
Acc. (%)	–	6.5 ± 1.5	26.8 ± 2.8	21.0 ± 2.6	16.6 ± 2.3	10.3 ± 2.1	<b>3.7 ± 1.0</b>	7.2 ± 1.7
<b>Unrelated Concepts (↑)</b>								
Acc. (%)	–	52.2 ± 2.7	<b>75.0 ± 1.9</b>	71.3 ± 2.2	70.4 ± 2.4	60.4 ± 2.6	52.8 ± 2.9	71.7 ± 2.1
<b>Inpainting (↓)</b>								
Acc. (%)	77.7 ± 1.5	<b>61.7 ± 2.4</b>	69.1 ± 1.8	69.1 ± 1.9	68.5 ± 1.7	66.8 ± 1.6	63.8 ± 2.0	68.2 ± 1.8
<b>Diffusion Completion <math>t = 5</math> (↓)</b>								
Acc. (%)	78.0 ± 1.4	<b>1.1 ± 0.58</b>	42.7 ± 2.7	37.8 ± 3.0	32.5 ± 3.2	2.4 ± 0.94	3.2 ± 1.2	36.5 ± 3.3
<b>Diffusion Completion <math>t = 10</math> (↓)</b>								
Acc. (%)	78.0 ± 1.4	<b>3.2 ± 1.1</b>	62.1 ± 2.3	54.8 ± 2.7	36.9 ± 3.0	6.1 ± 1.5	21.2 ± 2.2	45.4 ± 2.8

Table 2: Classification accuracy (% , mean ± std) across seven concept erasure methods and the original Stable Diffusion model (**Base**). Lower values (↓) on erased concepts, textual inversion, UnlearnDiffAtk, inpainting, and diffusion completion indicate more effective removal of the target concept. Higher values (↑) on unrelated concepts reflect successful preservation of general generation capabilities.