Sparse Autoencoder as a Zero-Shot Classifier for Concept Erasing in Text-to-Image Diffusion Models

Zhihua Tian [† 1] Sirun Nan [2] Ming Xu [2] Shengfang Zhai [3] Wenjie Qu [2] Jian Liu [1] Ruoxi Jia [* 4]
Jiaheng Zhang [* 2]

Abstract

Text-to-image (T2I) diffusion models have achieved remarkable progress in generating high-quality images but also raise people's concerns about generating harmful or misleading content. While extensive approaches have been proposed to erase unwanted concepts without requiring retraining from scratch, they inadvertently degrade performance on normal generation tasks. In this work, we propose **Interpret then Deactivate** (`ItD`), a novel framework to enable precise concept removal in T2I diffusion models while preserving overall performance. `ItD` first employs a sparse autoencoder (SAE) to interpret each concept as a combination of multiple features. By permanently deactivating the specific features associated with target concepts, we repurpose SAE as a zero-shot classifier that identifies whether the input prompt includes target concepts, allowing selective concept erasure in diffusion models. Moreover, we demonstrate that `ItD` can be easily extended to erase multiple concepts without requiring further training. Comprehensive experiments across celebrity identities, artistic styles, and explicit content demonstrate `ItD` 's effectiveness in eliminating targeted concepts without interfering with normal concept generation. Additionally, `ItD` is also robust against adversarial prompts designed to circumvent content filters. Code is available at: `https://github.com/NANSirun/Interpret-then-deactivate`.

[*] Equal advising [†] Work done while visiting NUS [1]Zhejiang University [2]National University of Singapore [3]Peking University [4]Virginia Tech. Correspondence to: Jian Liu <liujian2411@zju.edu.cn>.

1. Introduction

Text-to-image (T2I) diffusion models have achieved remarkable success in generating images that faithful reflect the input text descriptions (Dhariwal & Nichol, 2021; Saharia et al., 2022; Ruiz et al., 2023; Rombach et al., 2022) while simultaneously raising concerns about being used to generate images containing inappropriate content such as offensive, pornographic, copyrighted, or Not-Safe-For-Work (NSFW) content (Schramowski et al., 2023; Rando et al., 2022; Gandikota et al., 2023). To mitigate the issue, concept erasing has been proposed to prevent T2I diffusion models from generating images relevant to the unwanted concepts without requiring retraining from scratch.

A recent line of research proposes fine-tuning model parameters to remove the unwanted knowledge learned by diffusion models (Gandikota et al., 2023; Kumari et al., 2023; Fan et al., 2023; Huang et al., 2023; Gandikota et al., 2024; Orgad et al., 2023). However, even only modifying the cross-attention (CA) layers within diffusion models would **inadvertently degrade the generation quality of normal concepts** (Gandikota et al., 2023; Huang et al., 2023; Zhang et al., 2024b; Bui et al., 2024b). For instance, erasing the concept of "nudity" could impair the model's ability to generate an image of a person. To mitigate the issue, some approaches incorporate regularization techniques during fine-tuning to preserve the generation quality of normal concepts (Kumari et al., 2023; Fan et al., 2023; Ko et al., 2024). However, fine-tuning with a subset of concepts may introduce new biases into the model, leading to unpredictable performance degradation when generating other concepts (Bui et al., 2024b).

As an alternative solution, some approaches integrate customized modules into the model to enable concept erasing without modifying the original model parameters(Lyu et al., 2024; Lu et al., 2024; Lee et al., 2025). However, each module may also affect the generation of normal concepts due to its limited generalization capability. Moreover, erasing new concepts requires training additional modules, resulting in substantial computational overhead.

In this work, we aim to overcome the above limitation by introducing a novel framework, Interpret-then-Deactivate (`ItD`), to enable **precise** and **expandable** concept erasure in T2I diffusion models. **Precise** refers to the erasure only influences the generation of the target concept while **Expandable** means that the approach can be easily extended to erase multiple concepts without further training.

`ItD` employs sparse autoencoder (SAE) (Olshausen & Field, 1997), an unsupervised model, to learn the sparse features that constitute the semantic space of the text encoder. Within this space, we interpret each concept as a linear combination of sparse features, which may overlap with the feature sets between the target and normal concepts. We hypothesize that this overlap is a key factor causing unintended effects on the normal concepts during erasure.

To this end, we propose to selectively erase the features unique to the target concept, enabling the precise erasure. This can be achieved by first encoding the text embedding of the concept into the sparse feature space, deactivating the specific features, and then decoding it back into the embedding space. To enable expandability in erasing multiple concepts, we can deactivate concept-specific features without requiring additional retraining.

In summary, we make the following contributions: (1) We propose `ItD` , a novel framework to enable **precise** and **expandable** concept erasure in T2I diffusion models. (2) To the best of our knowledge, we are the first to adopt SAE to concept erasing tasks in T2I diffusion models. (3) Extensive experiments across various datasets demonstrate that `ItD` effectively erases target concepts while preserving the diversity of remaining concepts, outperforming baselines by large margins.

2. Related Works

2.1. Sparse Autoencoder (SAE)

The internal of the neural network (NN) is hard to explain due to its polysemanticity nature, where neurons appear to activate in multiple, semantically distinct contexts. Recently, SAE has emerged as an effective tool for interpreting mechanisms of NN by breaking done the intermediate results into features interpretable to specific concepts (Huben et al., 2024; Kissane et al., 2024). Let $\mathbf{x} \in \mathbb{R}^{d_{in}}$ denote the input vector, the decoder and encoder of SAE can be formalized as:

$$\begin{aligned}\mathbf{z} &= \mathrm{ReLU}(W_{\text{enc}}\mathbf{x} + \mathbf{b}) \\ \hat{\mathbf{x}} &= W_{\text{dec}}\mathbf{z} \\ &= \sum_{i=0}^{d_{\text{hid}}-1} z_i \mathbf{f}_i\end{aligned}$$

where $W_{\text{enc}} \in \mathbb{R}^{d_{\text{hid}} \times d_{\text{in}}}$ and $W_{\text{dec}} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{hid}}}$ are learned matrices of encoder and decoder. $\mathbf{b} \in \mathbb{R}^{d_{\text{hid}}}$ is the learned bias. The loss to train the autoencoder is $\mathcal{L}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha\mathcal{L}_{aux}$, where $\mathcal{L}_{aux}$ is the loss to control the sparsity of the reconstruction (Huben et al., 2024) or to prevent dead features that do not been fired on a large number of training samples (Gao et al., 2024), scaled with coefficient α.

In our work, $\mathcal{L}_{aux}$ is the reconstruction error using only the largest K_{aux} feature activations following (Gao et al., 2024). We adopt SAE to identify and deactivate features specific to the target concepts to disable the diffusion model to generate related images.

2.2. Concept Erasing in T2I Diffusion Model

Diffusion models can be used to generate inappropriate content (Zhang et al., 2025; Schramowski et al., 2023). To address this issue, several approaches have been proposed, such as dataset censoring (Face & CompVis, 2023b), post-generation filtering (Face & CompVis, 2023a), and safety-guided generation (Schramowski et al., 2023). However, these methods either demand significant computational resources (Face & CompVis, 2023b), introduce new biases, or remain vulnerable to adversarial prompts (Yang et al., 2024; Rando et al., 2022).

To address these limitations, fine-tuning-based approaches have been extensively explored to erase target concepts. FMN (Zhang et al., 2024a) efficiently erases target concepts by re-steering cross-attention (CA) layers. UCE (Gandikota et al., 2024) and TIME (Orgad et al., 2023) modify the projection layer within CA layers with a closed-form solution. ESD (Gandikota et al., 2023) reduces the probability of generating images that are labeled as target concepts, and AC (Kumari et al., 2023) aligns the distribution of target concepts with surrogate concepts for concept erasure. To improve the effectiveness of erasing, SalUn (Fan et al., 2023) and Scissorhands (Wu & Harandi, 2025) identify the most sensitive neurons related to target concepts and update only those neurons.

2.3. Forgetting on Remaining Concepts

While the above approaches demonstrate good performance in erasing target concepts, they inadvertently degrade the generation of remaining concepts (Zhang et al.). To alleviate the issue, many approaches (Huang et al., 2023; Heng & Soh, 2024; Ko et al., 2024; Zhang et al., 2024b) utilize a regularization loss on remaining concepts to preserve their generation capability. EAP (Bui et al., 2024b) further investigate the impact of selecting different concepts for regularization, and OTE (Bui et al., 2024a), which employs a similar training objective to AC (Kumari et al., 2023), adaptively selecting the optimal surrogate concept during erasing. However, the effectiveness of regularization on unseen concepts remains unclear due to the enormous scale of normal concepts.

As an alternative solution, some approaches incorporate customized modules into the intermediate layers of the diffusion model without modifying original model parameters. we summarize them inference-based approaches. SPM (Lyu et al., 2024) applies one-dimensional LoRA (Hu et al., 2021) to the intermediate layers of diffusion models and proposes an anchoring loss for distant concepts. MACE (Lu et al., 2024) propose using LoRA for erasing each target concept and introduced a loss integrating LoRAs from multiple target concepts, enabling the massive concept erasing while mitigating forgetting of remaining concepts. CPE (Lee et al., 2025) introduces a residual attention gate, a module inserted into each CA layer of the diffusion model to control whether erasure should be applied to a given concept. However, erasing multiple concepts requires training separate modules for each target concept, resulting in significant computational overhead.

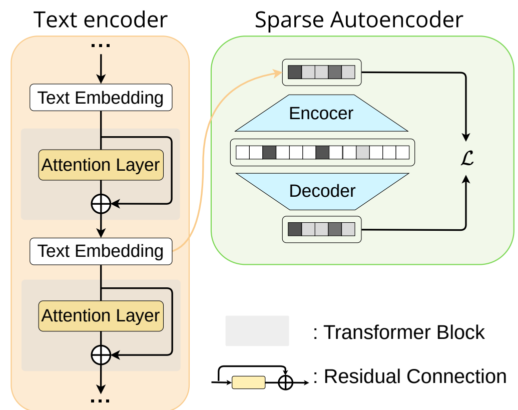


Figure 1. Unsupervised training of SAE, which takes a token embedding obtained from the residual streamer in text encoder as an input and aims to reconstruct it with sparse features

3. Sparse Autoencoder for T2I Diffusion Models

In this section, we introduce the training of an SAE for T2I diffusion model. We first discuss where to apply SAE for effective and efficient concept erasing in diffusion models in Section 3.1. We then introduce how an SAE is trained in Section 3.2.

3.1. Where to apply Spare Autoencoder.

A T2I diffusion model comprises multiple modules that work jointly to generate an image. As a short preliminary, we take the Latent DM (LDM) (Rombach et al., 2022) as an example, wherein the text encoder and U-net are two main modules. The text encoder transforms input text prompts into embeddings E, which is used to guide the image generation process. The U-net module works as a noise predictor, taking the text embedding E, timestep t, and the noised latent representation x_t as inputs to predict the noise added at time t. Specifically, with an initial noise $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the generation process iteratively performs denoising operations on x_T, ultimately producing the final image x_0. Take DDPM (Ho et al., 2020) as an example, the denoise step can be formulated as:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \epsilon_\theta \left(x_t, E, t \right) \right) + \sigma_t \epsilon, \quad (1)$$

where ϵ_θ is realized by an U-net model, $\alpha_t, \bar{\alpha}_t, \sigma_t$ are predefined values and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

To disable the model's ability to generate unwanted images, numerous approaches propose modifying the U-net module to remove unwanted knowledge (Zhang et al., 2024a; Kumari et al., 2023; Gandikota et al., 2023; Fan et al., 2023; Huang et al., 2023; Bui et al., 2024b; Lu et al., 2024; Li et al., 2024; Lee et al., 2025). However, due to (1) the complexity of the U-Net architecture and (2) the iterative nature of the denoising process, which typically requires multiple steps (e.g., 50 in DDIM), even minor modifications to the U-Net can lead to unexpected outcomes, potentially degrading overall performance.

To mitigate the issue, we propose applying SAE to the text encoder to remove unwanted knowledge from the text embedding before feeding it into the U-Net. The key motivations behind this are as follows:

- During the image generation process, the text embedding E plays a dominant role in encoding the semantic information to the generated images (cf. Equation 1). Consequently, erasing concepts at the text embedding level is sufficient to prevent their appearance in generated images.
- While previous works perform concept erasure within U-Net, their modifications mainly target text information processed within the CA layers (i.e., Key and Value projections, as detailed in the Appendix) (Lu et al., 2024; Lee et al., 2025), which demonstrates the effectiveness of performing erasure on text embedding.
- Prior studies demonstrate that performing unlearning on the text encoder achieves the best robustness against adversarial attacks (Zhang et al., 2024b).

In the following section, we detail the training process of SAE.

3.2. Training of SAE

The text encoder comprises a series of transformer blocks (cf. Figure 1). Denote $\mathcal{T}_l$ as the l-th transformer block, the text encoder with L blocks can be roughly formulated as $\text{TextEncoder} = \mathcal{T}_L \circ ... \circ \mathcal{T}_1$. To train an SAE for concept erasure, we focus on residual stream (Elhage et al., 2021), which is the output of a transformer block. The residual stream of l-th layer can be represented as:

$$\mathbf{e}_l = \mathcal{T}_l \circ ... \circ \mathcal{T}_1(\mathbf{e}_0),\ \mathbf{e}_l \in \mathbb{R}^{H \times d_{\text{in}}} \tag{2}$$

where $\mathbf{e}_0$ is the embedding of the tokenized prompt, H is the number of tokens composing the prompt, and d_{in} is the output dimension. We assume that d_{in} are the same across different layers for notation simplicity. We train an SAE that aims to learn the sparse features for each token embedding $\mathbf{e}_l^h, h \in [1, .., H]$. Therefore, for a prompt with H tokens, we get H samples to train SAE.

In our work, we adopt K-sparse autoencoder (K-SAE) (Makhzani & Frey, 2013), which could explicitly control the number of active latents by only keeping the K largest activations and zeros the rest for reconstruction. Let $\mathbf{e} \in \mathbb{R}^{d_{\text{in}}}$ refers to a single SAE training example. The encoder and decoder within K-SAE are then defined as follows:

$$\begin{aligned} \mathbf{z} &= \mathrm{TopK}\left(W_{\text{enc}}\left(\mathbf{e} - \mathbf{b}_{pre}\right)\right) \\ \hat{\mathbf{e}} &= W_{\text{dec}}\mathbf{z} + \mathbf{b}_{\text{pre}} \end{aligned} \tag{3}$$

where $W_{\text{enc}} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{hid}}}$ and $W_{\text{dec}} \in \mathbb{R}^{d_{\text{hid}} \times d_{\text{in}}}$ are learned matrix of encoder and decoder, b_{pre} is the bias term. d_{hid} is significantly larger than d_{in} to enforce the sparsity of learned features.

The training objective is:

$$\mathcal{L}(\mathbf{x}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \alpha \mathcal{L}_{aux}, \tag{4}$$

where $\mathcal{L}_{aux}$ is the reconstruction error using top K_{aux} ($K_{aux} > K$) feature activations, to prevent dead features that have not been fired on a large number of training samples (Gao et al., 2024), scaled with coefficient α.

We refer to the activation of the ρ-th learned feature as $z^\rho \in \mathbb{R}$. Its associated feature vector $f_\rho \in \mathbb{R}^{d_{in}}$ is a column in the decoder matrix $W_{\text{dec}} = \left(\mathbf{f}_1 | \cdots | \mathbf{f}_{n_f}\right) \in \mathbb{R}^{d_{\text{hid}} \times d_{\text{in}}}$. As a result, we can represent each token embedding as a sparse sum

$$\mathbf{e} \approx \sum_{\rho=1}^{d_{hid}} z^\rho \mathbf{f}_\rho, \text{ with } ||\mathbf{z}||_0 \leq K. \tag{5}$$

The details of training SAE are presented in Appendix.

4. Methodology

In this section, we introduce `ItD` , a framework to erase multiple concepts in a pretrained T2I diffusion model.

4.1. Overview

With a well-trained SAE, a straightforward approach to erasing unwanted knowledge within the text embedding is to deactivate features associated with the target concepts during the inference of the text encoder. However, target concepts often share certain features learned by SAE with normal concepts (e.g., *"nudity"* and *"person"* both contain information related to the *"human body"*). As a result, indiscriminately deactivating all related features could affect the generation of normal concepts.

To solve the problem, we propose a simple yet effective contrast-based method to identify features specific to the target concepts (Section 4.2). By only deactivating carefully selected features, `ItD` could effectively remove unwanted knowledge while having limited effects on normal concepts. Furthermore, we find that SAE can be exploited as a zero-shot classifier to distinguish between target and remaining concepts. This enables selective concept erasure, further reducing the impact on normal concepts (Section 4.3). Figure 2 depicts the workflow of `ItD` . Overall, it is designed to meet the following criteria:

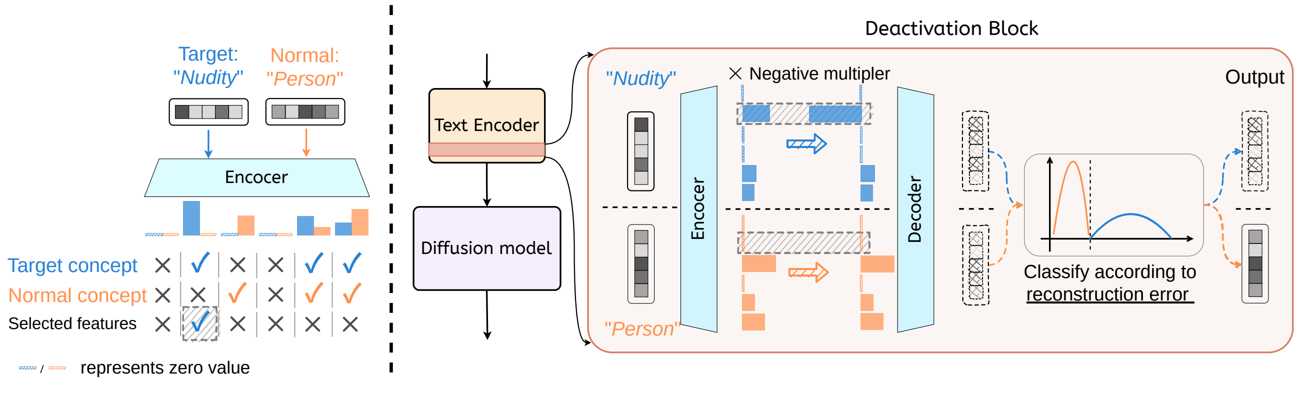
- Effectiveness: The unlearned model could not generate images of target concepts even when prompted with texts related to target concepts.
- Robustness: The model should also prevent the generation of images that are semantically related to synonyms of the targeted concepts, ensuring that erasure is not restricted to exact prompt wording.
- Specificity: The erasure should target only the specified concepts, with minimal or no impact on the remaining concepts.
- Expandability: When erasing new concepts, the algorithm can be easily extended without the need for additional training.

4.2. Feature Selection.

Select Features for a Concept. A concept C is typically composed of multiple tokens, denoted as $C_1, \ldots, C_H$ where $H \geq 1$. For example, "Bill Clinton" consists of the tokens "Bill" and "Clinton". SAE decomposes the embedding of each token into a set of features. To select features that can represent the concept, we collect features from all token embedding and select the top K_{sel} features based on their activation values s^ρ. Formally, we denote F as the set of indices of features specific to the concept C,

$$\begin{aligned} F = \{\rho | s_C^\rho \in \mathrm{TopK}(s_C^1, ..., s_C^{d_{hid}})\}, \\ \text{where } s_C^\rho = \max(s_1^\rho, ..., s_H^\rho). \end{aligned} \tag{6}$$

Select Features for Concept Erasing. After selecting features associated with target concepts, we propose using the features that are specifically activated by the target concepts for erasure. To achieve that, we suggest a simple yet effective contrast-based approach. Specifically, given a retain set $\mathcal{C}_{\text{retain}}$ comprising normal concepts, the features specific to the target concept (i.e. feature to deactivate) $\hat{F}_{\text{tar}}$



(a) Feature selection

(b) Feature Deactivation during inference

Figure 2. (a) With a well-trained SAE, we identify unique features of target concepts by contrast with normal concepts; (b) we wrap SAE as a deactivation block and insert it into the text encoder for concept eraser.

can be found by eliminating features that can be activated by normal concepts in $\mathcal{C}_{\text{retain}}$:

$$\hat{F}_{\text{tar}} = F_{\text{tar}} \setminus \bigcup_{C_r \in C_{\text{retain}}} F_{C_r}. \tag{7}$$

In our experiments, we use the concepts employed for the utility-preserving regularization term in previous works (Kumari et al., 2023; Zhang et al., 2024b; Fan et al., 2023; Lu et al., 2024) as the retain set for comparison. Notably, our approach does not require fine-tuning on these concepts, thereby avoiding the introduction of new biases and making more effective use of them.

When erasing multiple concepts, we select features for erasing as the union of the specific features associated with each target concept. Specifically, let $\mathcal{C}_{\text{tar}}$ denote the set of target concepts.

$$F_{\text{erase}} = \bigcup_{C \in \mathcal{C}_{\text{tar}}} \hat{F}_C. \tag{8}$$

4.3. Concept Erasing

To erase knowledge of target concepts within the text embedding, we first encode the text embeddings into their activations s using SAE. We then modify each activation component as follows:

$$\hat{s}^\rho = \begin{cases} s^\rho \cdot \tau, & \text{if } \rho \in F_{\text{erase}} \\ s^\rho, & \text{otherwise} \end{cases} \tag{9}$$

where F_{erase} represents the set of features selected before, and τ is a scaling factor controlling the degree of deactivation. Finally, the decoder reconstructs the modified embedding as $\hat{\mathbf{e}} = \text{Dec}(\hat{s})$.

To build an unlearned model, we wrap SAE as a deactivation block inserted into the intermediate layers of the text encoder (Figure 2 (b)). Any text embedding inputted to the block would erase the information of target concepts. As a result, the generated image guided by the text embedding would not contain target concepts. The block is plug-and-play, which means we do not need to fine-tune the diffusion model, making it highly efficient. Moreover, since the selected features are not activated by normal concepts, Equation 10 does not influence normal concepts.

Selective feature deactivation The inherent reconstruction loss of SAE may still influence the generation of normal concepts. To alleviate the issue, we propose a mechanism to selectively apply SAE for concept erasure. Given a text embedding $\mathbf{e}$ and its reconstructed version $\hat{\mathbf{e}}$, we construct a classifier G to identify whether $\mathbf{e}$ contains information about target concepts. The classification is based on the reconstruction loss between $\mathbf{e}$ and $\hat{\mathbf{e}}$:

$$G(\mathbf{e}) = \begin{cases} 1, & \text{if } ||\mathbf{e} - \hat{\mathbf{e}}||^2 < \tau \\ 0, & \text{if } ||\mathbf{e} - \hat{\mathbf{e}}||^2 \geq \tau \end{cases} \tag{10}$$

where τ is the threshold. Finally, the deactivation block outputs $\mathbf{e}$ for subsequent computation if it does not contain target concept information; otherwise, it outputs $\hat{\mathbf{e}}$.

Figure 3 demonstrates the effectiveness of the classifier. We select 50 celebrities as target concepts for erasure. For the remaining concepts, we include 100 different celebrities, 100 artistic styles, and the COCO-30K dataset. It shows that there is a clear boundary in the reconstruction loss between the target concepts and the remaining concepts.

5. Experiments

In this section, we conduct comprehensive experiments to evaluate the effectiveness of our approach.

Table 1. Quantitative results on celebrities erasure. We use CLIP Score (CS) and GCD accuracy (ACC) for target celebrities. We measured CS and FID for COCO-30K and DiffusionDB-10K, or KID for the other remaining concepts.

Methods	Target Concepts		Remaining Concepts							Unpredictable Concepts	
	50 Celebrities		100 Celebrities			100 Artistic Styles		COCO-30K		DiffusionDB-10K	
	CS ↓	ACC% ↓	CS ↑	ACC% ↑	KID(×100) ↓	CS ↑	KID(×100) ↓	CS ↑	FID(×100) ↓	CS ↑	FID ↓
ESD-x (Gandikota et al., 2023)	24.41	7.30	26.23	10.39	2.66	28.23	0.01	29.55	14.40	28.94	9.46
AC (Kumari et al., 2023)	33.63	90.16	33.98	87.04	1.86	28.47	0.38	30.91	16.91	31.28	8.55
AdvUn (Zhang et al., 2024b)	16.71	0.00	17.61	2.84	14.80	19.29	10.29	18.25	47.38	14.53	63.61
Receler (Huang et al., 2023)	**12.84**	0.00	13.37	86.45	18.09	24.80	5.23	29.98	15.85	27.85	18.96
MACE (Lu et al., 2024)	24.60	3.29	34.39	84.64	0.23	27.75	0.47	30.38	**14.03**	29.59	11.10
CPE (Lee et al., 2025)	20.79	0.37	34.82	88.26	0.08	29.01	0.01	30.86	14.62	31.84	4.18
ItD (Ours)	19.65	**0.00**	**34.87**	**89.12**	**0**	**29.02**	**0**	**31.02**	14.72	**32.06**	**0.91**
SD1.4 (Rombach et al., 2022)	34.49	90.56	34.87	89.12	-	29.02	-	31.02	14.73	32.17	-

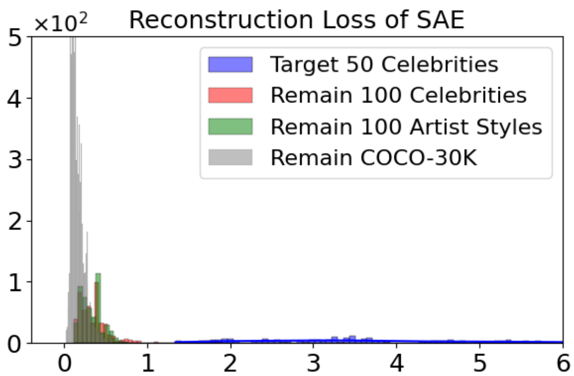


Figure 3. Histogram of reconstruction MSE loss on target concepts and remaining concepts.

Setting. We consider four domains for concept erasing tasks: celebrities, artistic styles, COCO-30K captions, and DiffusionDB-10K captions, where the last dataset contains 10K captions collected from DiffusionDB (Wang et al., 2022). These captions are constructed by collecting chat messages from Stable Diffusion Discord channels, representing the user prompts in a real-world setting. Next, we conduct experiments on the removal of explicit contents and evaluate the efficacy on I2P dataset (Schramowski et al., 2023). We also evaluate the robustness against adversarial prompts using the red-teaming tools UnlearningDiff (Zhang et al., 2025). We generate images using SD1.4 (Rombach et al., 2022) with DDIM in 50 steps.

SAE Training Setting For celebrity and artistic style erasure, we train an SAE using 200 celebrity names and 200 artist names from MACE (Lu et al., 2024). Additionally, we incorporate the full captions from COCO-30K. The celebrity and artist names are embedded into 100 and 35 prompt templates, respectively, following the setup in CPE (Lee et al., 2025). This results in a total of 57,000 text samples for training the SAE. For nudity erasure, we collect 10,000 captions from DiffusionDB (Wang et al., 2022) with an NSFW score above 0.8, along with all captions from COCO-30K, to train the SAE. Further details on the training process can be found in the Appendix B.

Baselines. We compare with six baselines including four fine-tuning-based approaches: ESD (Gandikota et al., 2023), AC (Kumari et al., 2023), AdvUn (Zhang et al., 2024b), Receler (Huang et al., 2023). and two inference-based approaches: MACE (Lu et al., 2024), and CPE (Lee et al., 2025). The implementation details are provided in Appendix A.

Metrics. We adopt CLIP score (Hessel et al., 2021) to evaluate the quality of generated images, where a lower score indicates an effective erasure for target concepts, and a higher score refers to better preservation for the remaining concepts. Additionally, for celebrity erasure experiments, we utilize the GIPHY Celebrity Detector (Nick Hasty & Korduban, 2025) to measure the top-1 accuracy (**ACC**) of generated celebrity images. A lower accuracy is better for target concepts, and a high accuracy is better for remaining concepts. For COCO-30K and DiffusionDB-10K captions, we evaluate their Frechet Inception Distance (**FID**) (Heusel et al., 2017). A lower value is better. Following (Lee et al., 2025), we evaluate the Kernel Inception Distance (**KID**) for other remaining concepts, which is more stable and reliable for a smaller number of images.

5.1. Celebrity Erasure

We select 50 celebrities as the target concepts from the list of celebrities provided by (Lu et al., 2024), which consists of 200 celebrities. For the remaining concepts, we consider two domains: 100 celebrities and 100 artistic styles. We generated 25 images using 5 prompt templates with 5 random seeds, resulting in 2,500 images for each remaining domain. We also use COCO-30K and DiffusionDB-10K as remaining concepts. Notably, the captions in DiffusionDB-10K are not used for training SAE, allowing us to evaluate the effectiveness of our approach on unpredictable concepts.

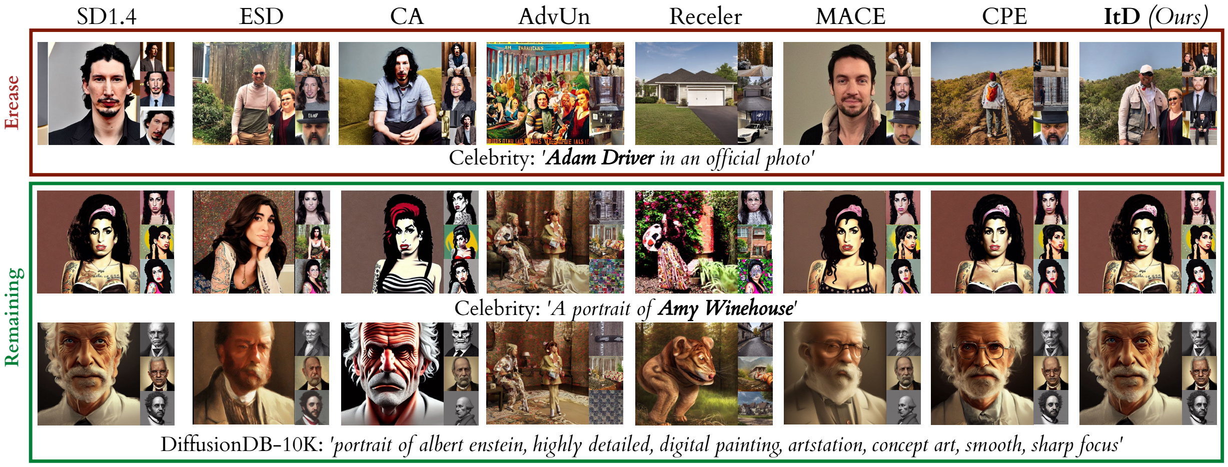


Figure 4. Qualitative results of our `ItD` and baselines on multiple concepts erasing. We erased 50 celebrities at once. The remaining celebrity concepts serve as surrogate concepts in the baselines and as training data for SAE in `ItD` , whereas the concepts in DiffusionDB-10K are used solely during the generation process.

Figure 4 presents the qualitative results of erasing multiple celebrities. The results indicate that `ItD` and all baselines, except CA, effectively remove the target concepts. CA struggles in this scenario due to the large number of target concepts, exceeding its capacity for erasure. For the remaining concepts, AdvUn and Receler cause significant degradation in image quality, likely because adversarial unlearning introduces broader disruptions to the generation process. In contrast, MACE and CPE preserve the quality of remaining celebrity images, as they are trained within the same domain. However, images from DiffusionDB-10K exhibit noticeable deviations from those generated by the original model. Among all methods, `ItD` is the only approach that maintains the same generation quality as the original model.

Table 1 presents quantitative results on erasing target concepts while evaluating the impact on remaining concepts. The results demonstrate that `ItD` has the least effect on remaining concepts and outperforms all baselines by a large margin in preserving them, while still achieving strong erasure performance.

5.2. Artistic Styles Erasure

We select 100 artistic styles as the target concepts following (Lu et al., 2024) and utilize the same remaining concepts in Section 5.1, including 100 celebrity, 100 artistic styles, COCO-30K, and DiffusionDB-10K. We utilize the same SAE used for celebrity erasure. Figure 5 shows the qualitative results of erasing multiple artist styles. Among all baselines, Receler fails to erase the target artistic style, possibly because it adopts a mask to identify pixels for erasure within the image, which fails when erasing styles, leading to unsuccessful erasure. Table 2 shows results for artistic style erasure that are consistent with those from celebrity erasure. `ItD` effectively distinguishes between target artistic styles, remaining artistic styles, and celebrities, ensuring strong erasure performance while preventing degradation of remaining concepts. Additionally, it has the least impact on DiffusionDB-10K and outperforms all baselines by a large margin.

5.3. Explicit Content Erasure

We evaluate the effectiveness of erasing explicit contents on the I2P dataset (Schramowski et al., 2023). It consists of 4,703 prompts without inappropriate words but would generate explicit images with stable diffusion. To erase explicit concepts, we adopt four keywords as target concepts to select features following (Lu et al., 2024): 'nudity', 'naked', 'erotic', and 'sexual'. We employ the NudeNet detector (Bedapudi, 2025) to measure the frequency of explicit content. The threshold is set to 0.6 (Lu et al., 2024). For preservation performance evaluation, we utilize COCO-30K. We train an SAE using the captions of COCO-30K. Table 4 shows the number of explicit contents detected by the NudeNet detector. The results show that `ItD` resulted in the fewest detected explicit contents compared with baselines.

5.4. Robustness

To evaluate the robustness against the adversarial attack prompts, we utilize UnlearnDiff (Zhang et al., 2025) as the red-teaming tool to verify the robustness of our `ItD` . We evaluate on I2P dataset and report the attack success rate (ASR) as the evaluation metric to measure the ratio of gen-

Figure 5. Qualitative results of our `ItD` and baselines on multiple concepts erasing. We erased 100 artist styles at once. The remaining artist style concepts serve as surrogate concepts in the baselines and as training data for SAE in `ItD` , whereas the concepts in DiffusionDB-10K are used solely during the generation process.

Table 2. Quantitative results on artistic styles erasure. We used CLIP Score (CS)for target artistic styles. We measured CS and FID for COCO-30K and DiffusionDB-10K, or KID for the other remaining concepts.

Methods	Target Concepts	Remaining Concepts							Unpredictable Concepts	
	100 Artistic Styles	100 Artistic Styles		100 Celebrities			COCO-30K		DiffusionDB-10K	
	CS ↓	CS ↑	KID(×100) ↓	CS ↑	ACC% ↑	KID(×100) ↓	CS ↑	FID(×100) ↓	CS ↑	FID ↓
ESD-x (Gandikota et al., 2023)	20.89	28.90	0.65	30.42	81.86	0.81	29.52	15.19	28.11	11.54
AC (Kumari et al., 2023)	28.91	28.33	1.25	34.77	**93.71**	0.25	30.97	16.19	31.21	10.20
AdvUn (Zhang et al., 2024b)	**18.94**	19.28	9.65	17.78	0.0	13.90	18.11	43.24	12.63	62.84
Receler (Huang et al., 2023)	24.90	24.96	2.11	32.76	86.68	0.44	29.29	16.25	26.83	22.00
MACE (Lu et al., 2024)	22.59	28.95	0.25	26.87	10.79	0.25	29.51	**12.71**	25.77	16.70
CPE (Lee et al., 2025)	20.67	28.95	0.01	34.81	89.80	0.04	30.95	14.77	31.60	5.72
ItD (Ours)	19.88	**29.02**	**0.00**	**34.87**	89.12	**0.00**	**31.02**	14.71	**31.91**	**1.07**
SD1.4 (Rombach et al., 2022)	29.72	29.02	-	34.87	89.12	-	31.02	14.73	32.17	-

Table 3. Robust concept erasure against adversarial attack: UnlearnDiff (Zhang et al., 2025)

Method	ASR (↓)
FMN(Zhang et al., 2024a)	97.89
ESD (Gandikota et al., 2023)	76.05
UCE (Gandikota et al., 2024)	79.58
MACE (Lu et al., 2024)	66.90
AdvUnlearn (Zhang et al., 2024b)	21.13
`ItD` (Ours)	**12.61**

erated images containing explicit content. The results are shown in Table 3.

From Table 3, we can find that `ItD` demonstrates robust erasure of target concepts competitive to recent robust methods AdvUnlearn (Zhang et al., 2024b) and MACE (Lu et al., 2024). In particular, `ItD` successfully defends attacks by UnlearnDiff, significantly outperforming the existing approaches, verifying its robustness.

6. Conclusion

In this work, we show that only fine-tuning CA layers for concept erasing in diffusion models could sometimes fail in preserving remaining concepts. As one solution, we proposed our framework, `ItD` (Interpret-then-deactivate), simple and effective approach to remove the regulation constraint constraint aiming to erase target concepts while maintaining diverse remaining concepts. We integrate the Sparse Autoencoder (SAE) into the the diffusion models, making it capable of adaptively adjusting the text embeddings. To robustly erase target concepts without forgetting on remaining concepts, we also comprehensively compare the unique features unique to the target concepts, and deactive them, removing the effect on the remaining concepts. Through extensive experiments on erasure of celebrities, artistic styles, and explicit concepts, the empirical results ensure the robust deletion of target concepts and protection of diverse remaining concepts.

Table 4. Results of the detected number of explicit contents using NudeNet detector on I2P and preservation performance on COCO 30K with CS, FID.

Methods	Number of nudity detected on I2P (Detected Quantity)									COCO-30K	
	Armpits	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Total	CS ↑	FID ↓
FMN (Zhang et al., 2024a)	43	117	12	59	155	17	19	2	424	30.39	13.52
ESD-x (Gandikota et al., 2023)	59	73	12	39	100	4	30	8	315	30.69	14.41
AC (Kumari et al., 2023)	153	180	45	66	298	22	67	7	838	**31.37**	16.25
AdvUn (Zhang et al., 2024b)	8	**0**	**0**	13	1	1	**0**	**0**	28	28.14	17.18
Receler (Huang et al., 2023)	48	32	3	35	20	0	17	5	160	30.49	15.32
MACE (Lu et al., 2024)	17	19	2	39	16	0	9	7	111	29.41	**13.40**
CPE (Lee et al., 2025)	10	8	2	8	6	1	3	2	40	31.19	13.89
ItD (Ours)	**0**	2	3	**3**	**0**	**0**	**0**	10	**18**	30.42	14.64
SD v1.4 (Rombach et al., 2022)	148	170	29	63	266	18	42	7	73	31.02	14.73
SD v2.1 (Face & CompVis, 2023b)	105	159	17	60	177	9	57	2	586	31.53	14.87

7. Impact Statements:

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, for example, improving the safety and ethical use of generative models by preventing the creation of harmful or explicit content. This could help mitigate the spread of inappropriate or harmful imagery, ensuring that AI systems align more closely with societal values and legal standards, particularly in areas such as content moderation, education, and creative industries.

References

Bedapudi, P. Nudenet: Neural nets for nudity detection and censoring., 2025. URL `https://nudenet.notai.tech/`.

Bui, A., Vu, T., Vuong, L., Le, T., Montague, P., Abraham, T., and Phung, D. Fantastic targets for concept erasure in diffusion models and where to find them. *Preprint*, 2024a.

Bui, A., Vuong, L., Doan, K., Le, T., Montague, P., Abraham, T., and Phung, D. Erasing undesirable concepts in diffusion models with adversarial preservation. *arXiv preprint arXiv:2410.15618*, 2024b.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Du, Y., Li, S., and Mordatch, I. Compositional visual generation with energy based models. *Advances in Neural Information Processing Systems*, 33:6637–6647, 2020.

Du, Y., Li, S., Sharma, Y., Tenenbaum, J., and Mordatch, I. Unsupervised learning of compositional energy concepts. *Advances in Neural Information Processing Systems*, 34:15608–15620, 2021.

Efron, B. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Face, H. and CompVis. Stable diffusion safety checker, 2023a. URL `https://huggingface.co/CompVis/stable-diffusion-safety-checker`.

Face, H. and CompVis. Stable diffusion 2, 2023b. URL `https://huggingface.co/stabilityai/stable-diffusion-2`.

Fan, C., Liu, J., Zhang, Y., Wong, E., Wei, D., and Liu, S. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*, 2023.

Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

Gandikota, R., Orgad, H., Belinkov, Y., Materzyńska, J., and Bau, D. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5111–5120, 2024.

Gao, L., la Tour, T. D., Tillman, H., Goh, G., Troll, R., Radford, A., Sutskever, I., Leike, J., and Wu, J. Scaling and evaluating sparse autoencoders, 2024. URL `https://arxiv.org/abs/2406.04093`.

Heng, A. and Soh, H. Selective amnesia: A continual learning approach to forgetting in deep generative models. *Advances in Neural Information Processing Systems*, 36, 2024.

Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., and Choi, Y. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2021.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Huang, C.-P., Chang, K.-P., Tsai, C.-T., Lai, Y.-H., and Wang, Y.-C. F. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. *arXiv preprint arXiv:2311.17717*, 2023.

Huben, R., Cunningham, H., Riggs, L., Ewart, A., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL `https://openreview.net/forum?id=F76bwRSLeK`.

Jin, M., Yu, Q., Huang, J., Zeng, Q., Wang, Z., Hua, W., Zhao, H., Mei, K., Meng, Y., Ding, K., et al. Exploring concept depth: How large language models acquire knowledge at different layers? *arXiv preprint arXiv:2404.07066*, 2024.

Kinga, D., Adam, J. B., et al. A method for stochastic optimization. In *International conference on learning representations (ICLR)*, volume 5, pp. 6. San Diego, California;, 2015.

Kissane, C., Krzyzanowski, R., Bloom, J. I., Conmy, A., and Nanda, N. Interpreting attention layer outputs with sparse autoencoders, 2024. URL `https://arxiv.org/abs/2406.17759`.

Ko, M., Li, H., Wang, Z., Patsenker, J., Wang, J. T., Li, Q., Jin, M., Song, D., and Jia, R. Boosting alignment for post-unlearning text-to-image generative models. *arXiv preprint arXiv:2412.07808*, 2024.

Kumari, N., Zhang, B., Wang, S.-Y., Shechtman, E., Zhang, R., and Zhu, J.-Y. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22691–22702, 2023.

Lee, B. H., Lim, S., Lee, S., Kang, D. U., and Chun, S. Y. Concept pinpoint eraser for text-to-image diffusion models via residual attention gate. In *International Conference on Learning Representations*, 2025. URL `https://openreview.net/forum?id=ZRDhBwKs7l`.

Li, X., Yang, Y., Deng, J., Yan, C., Chen, Y., Ji, X., and Xu, W. Safegen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 4807–4821, 2024.

Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55. Springer, 2024.

Lu, S., Wang, Z., Li, L., Liu, Y., and Kong, A. W.-K. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024.

Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., and Ding, G. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7559–7568, 2024.

Makhzani, A. and Frey, B. K-sparse autoencoders. *arXiv preprint arXiv:1312.5663*, 2013.

Nick Hasty, Ihor Kroosh, D. V. and Korduban, D. Giphy celebrity detector, 2025. URL `https://github.com/Giphy/celeb-detection-oss`.

Olshausen, B. A. and Field, D. J. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

Orgad, H., Kawar, B., and Belinkov, Y. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7053–7061, 2023.

Rando, J., Paleka, D., Lindner, D., Heim, L., and Tramèr, F. Red-teaming the stable diffusion safety filter. *arXiv preprint arXiv:2210.04610*, 2022.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., and Aberman, K. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22500–22510, 2023.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Schramowski, P., Brack, M., Deiseroth, B., and Kersting, K. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22522–22531, 2023.

Wang, Z. J., Montoya, E., Munechika, D., Yang, H., Hoover, B., and Chau, D. H. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. URL `https://arxiv.org/abs/2210.14896`.

Wu, J. and Harandi, M. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pp. 367–384. Springer, 2025.

Yang, Y., Hui, B., Yuan, H., Gong, N., and Cao, Y. Sneakyprompt: Jailbreaking text-to-image generative models. In *2024 IEEE symposium on security and privacy (SP)*, pp. 897–912. IEEE, 2024.

Zhang, G., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1755–1764, 2024a.

Zhang, Y., Fan, C., Zhang, Y., Yao, Y., Jia, J., Liu, J., Zhang, G., Liu, G., Kompella, R. R., Liu, X., et al. Unlearncanvas: Stylized image dataset for enhanced machine unlearning evaluation in diffusion models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhang, Y., Chen, X., Jia, J., Zhang, Y., Fan, C., Liu, J., Hong, M., Ding, K., and Liu, S. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *arXiv preprint arXiv:2405.15234*, 2024b.

Zhang, Y., Jia, J., Chen, X., Chen, A., Zhang, Y., Liu, J., Ding, K., and Liu, S. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now. In *European Conference on Computer Vision*, pp. 385–403. Springer, 2025.

A. Preliminaries on baseline approaches

We consider six baseline approaches, including 4 fine-tuning-based approaches:ESD (Gandikota et al., 2023), AC (Kumari et al., 2023), AdvUn (Zhang et al., 2024b), Receler (Huang et al., 2023), and two inference-based approaches MACE (Lu et al., 2024), CPE (Lee et al., 2025).

A.1. Cross-Attention in T2I Diffusion Model

We first introduce the fundamentals of the cross-attention (CA) layer in T2I diffusion models, which is widely utilized in previous works. The CA layer serves as a key mechanism for integrating textual information, represented by text embeddings $\mathbf{E}$, with image features, represented as image latents x_t, within T2I diffusion models. Specifically, the image feature $\phi(x_t)$ is linearly transformed into query matrix $Q = l_Q(\phi(x_t)$, while the text embedding is mapped into key matrix $K = l_K(\mathbf{E})$ and value matrix $V = l_V(\mathbf{E})$ through separate linear transformations. The attention map can be calculated as:

$$M = \mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \in \mathbb{R}^{u\times(h\times w)\times n},$$

where d is the projection dimension, u is the number of attention heads, $h \times w$ represents the spatial dimension of the image, and n is the number of text tokens. The final output of the cross-attention is MV, representing the weighted average of the values in V.

A.2. Fine-tuning-based approaches

ESD is inspired by energy-based composition (Du et al., 2020; 2021). It aims to reduce the probability of generating images belonging to the target concept. Specifically, it aims to minimize

$$P_\theta(x) \propto \frac{P_{\theta_o}(x)}{P_{\theta_o}(c \mid x)^\eta}, \tag{11}$$

where $P_{\theta_o}(x)$ represents the distribution generated by the original model, c is the target concept to erase, and η is scale parameter. Inspired by Equ. 11, and Tweedie's formula (Efron, 2011)

$$\tilde{\epsilon}_\theta\left(x_t \mid c_e\right) \leftarrow \epsilon_{\theta_o}\left(x_t \mid \emptyset\right) - \eta\left(\epsilon_{\theta_o}\left(x_t \mid c\right) - \epsilon_{\theta_o}\left(x_t \mid \emptyset\right)\right), \tag{12}$$

The training loss to optimize θ is:

$$\min_\theta \ell_{\mathrm{ESD}}(\theta, c) := \mathbb{E}\left[\left\|\epsilon_\theta\left(x_t \mid c\right) - \left(\epsilon_{\theta_0}\left(x_t \mid \emptyset\right) - \eta\left(\epsilon_{\theta_0}\left(x_t \mid c\right) - \epsilon_{\theta_0}\left(x_t \mid \emptyset\right)\right)\right)\right\|_2^2\right]. \tag{13}$$

Empirical results show that updating only the CA layers (ESD-x) effectively erases target concepts while preserving the overall performance of remaining concepts. In contrast, fine-tuning all parameters (ESD-u) can lead to significant degradation of remaining concepts. In our implementation of ESD, we mainly implement ESD-x, as it demonstrates good concept erasure while minimizing its impact on remaining concepts.

AdvUn is designed based on ESD. It proposes a bi-level optimization approach that is robust to adversarial attacks. The upper-level optimization aims to minimize the unlearning objective, while lower-level optimization aims to find the optimized adversarial prompt c^*:

$$\begin{array}{lll} \underset{\boldsymbol{\theta}}{\operatorname{minimize}} & \ell_{\mathrm{u}}\left(\boldsymbol{\theta}, c^*\right) & \text{[Upper-level]} \\ \text{subject to} & c^* = \arg\min_{\|c^*-c\|_0 \le \rho} \mathbb{E}\left[\left\|\epsilon_\theta\left(\mathbf{x}_t \mid c^*\right) - \epsilon_{\theta^*}\left(\mathbf{x}_t \mid c\right)\right\|_2^2\right]. & \text{[Lower-level]} \end{array} \tag{14}$$

To maintain the utility on remaining concepts, it adopts a regularization term to penalize the discrepancy between the original model and optimized model on a retain set C_{retain}. As a result, the upper-level optimization objective is:

$$\ell_{\mathrm{u}}\left(\boldsymbol{\theta}, c^*\right) = \ell_{\mathrm{ESD}}\left(\boldsymbol{\theta}, c^*\right) + \gamma \mathbb{E}_{\tilde{c} \sim \mathcal{C}_{\text{retain}}}\left[\left\|\epsilon_{\boldsymbol{\theta}}\left(\mathbf{x}_t \mid \tilde{c}\right) - \epsilon_{\boldsymbol{\theta}_{\mathrm{o}}}\left(\mathbf{x}_t \mid \tilde{c}\right)\right\|_2^2\right]$$

AdvUn tried to perform erasing on different layers within the text encoder and U-net in SD v1-4 (Rombach et al., 2022). Empirical results show that erasing within the text encoder has the best robustness against adversarial attacks. In our implementation of AdvUn, we perform erasing on the text encoder.

Receler is also designed based on ESD and adopts adversarial prompt learning to ensure erasure robustness. To reduce the impact on remaining concepts, it adopts a concept-localized regularization for erasing locality:

$$\ell_{Reg} = \frac{1}{L} \sum_{l=1}^{L} \left\| o^l \odot (1 - M) \right\|^2 \tag{15}$$

where L is the number of U-Net's layers, $\odot$ is the element-wise product, o^l is the output of the eraser in the l-th layer, and M is the mask of target concept in image and generated using GroundingDINO (Liu et al., 2024)

AC is another approach for unlearning. It prevents the model from generating unwanted images by mapping target concepts to an anchor concept, which can be either a generic concept, such as 'dog' to replace 'English springer' or a null concept, such as an empty text prompt. Moreover, it also utilizes a regularization term to penalize the discrepancy between the original model and the optimized model on a set of retained concepts. The training objective can be formulated as follows:

$$\min_{\theta} \ell_{\text{AC}}(\theta, c) := \mathbb{E}\left[\left\|\epsilon_\theta\left(x_t \mid c\right) - \epsilon_{\theta_o}\left(x_t \mid c_a\right)\right\|_2^2\right] + \gamma \mathbb{E}_{\tilde{c} \sim \mathcal{C}_{\text{retain}}}\left[\left\|\epsilon_{\boldsymbol{\theta}}\left(\mathbf{x}_t \mid \tilde{c}\right) - \epsilon_{\boldsymbol{\theta}_o}\left(\mathbf{x}_t \mid \tilde{c}\right)\right\|_2^2\right], \tag{16}$$

where c_a is the surrogate concept.

A.3. Inference-based approaches

MACE adopts a closed-form solution to refine CA layers within U-net to erase unwanted knowledge. Briefly, it finds the linear projections W' of *keys* and *values* in the cross-attention layers such that:

$$\mathbf{W}_{\text{new}} = \arg\min_{\mathbf{W}'} \sum_{n=1}^{N} \left\|\mathbf{W}'\mathbf{E}_{\text{tar}}^n - \mathbf{W}_{\text{o}}\mathbf{E}_{\text{sur}}^n\right\|_F^2 + \lambda \sum_{m=1}^{M} \left\|\mathbf{W}'\mathbf{E}_{\text{retain}}^m - \mathbf{W}_{\text{o}}\mathbf{E}_{\text{retain}}^m\right\|_F^2, \tag{17}$$

where $\mathbf{W}_{\text{o}}$ is the original key/value projection. $\mathbf{E}_{\text{tar}}, \mathbf{E}_{\text{sur}}$ and $\mathbf{E}_{\text{retain}}$ represent the text embedding of target, surrogate, and retain concepts, respectively. To mitigate the impact on the overall parameters, MACE inserts LoRA modules into the CA layers of the model for each target concept. Then, the multiple LoRAs are integrated by a loss function to enable erasing multiple concepts.

CPE also works on the linear projections of *keys* and *values* in the cross-attention layers. It inserts a customized modular, named residual attention gate (ResAG), into each CA layer within U-net. ResAG is trained to make the projection output of $\mathbf{E}_{\text{tar}}$ similar to the output of $\mathbf{E}_{\text{sur}}$. Formally, the erasing objective is:

$$\min_{R_{\text{tar}}} \ell_{\text{era}} = \mathbb{E}_{(\mathbf{E}_{\text{tar}}, \mathbf{E}_{\text{sur}})} \left\| \left(\mathbf{W}\mathbf{E}_{\text{tar}} + R_{\text{tar}}\left(\mathbf{E}_{\text{tar}}\right)\right) - \left(\mathbf{W}\mathbf{E}_{\text{sur}} - \eta \mathbf{W}\left(\mathbf{E}_{\text{tar}} - \mathbf{E}_{\text{sur}}\right)\right)\right\|^2, \tag{18}$$

where R_{tar} is the ResAG for target concept. To prevent undesirable degradation on remaining concepts, CPE adopts a regularization term to minimize the deviation induced by ResAG on retain concepts:

$$\mathcal{L}_{\text{att}} = \mathbb{E}_{\mathbf{E}_{\text{retain}}} \left\| R_{\text{tar}}\left(\mathbf{E}_{\text{retain}}\right)\right\|_F. \tag{19}$$

Each ResAG is trained specifically to one target concept. Therefore, to erase multiple concepts, it is required to train multiple ResAGs.

B. Training details of SAE

We train an SAE using the output of the 8th transformer block of the text encoder `layers.8`, which we experimentally show has the best performance. We set $K = 64$ and $d_{\text{hid}} = 2^{19}$ following (Gao et al., 2024). We adopt Adam (Kinga et al., 2015) as the optimizer with the learning rate of $5e-5$ and a constant scheduler without warmup.

Following Gao et al. (2024), we set $\alpha = \frac{1}{32}$ and $K_{\text{aux}} = 256$. We train the SAE while simultaneously generating training samples with the text encoder, which does not require additional storage space to save samples. The batch size of prompts input into the text encoder is 50, which results in about 1000 samples to train SAE each time.

We train SAE on a single H100. For celebrity and artistic style erasure, where we train an SAE using celebrity and artist styles and the captions of COCO-30K, the training time is 56 minutes. For nudity erasure, we train an SAE using the captions of COCO-30 and 10K prompts from DiffusionDB (Wang et al., 2022) with an NSFW score larger than 0.8, the training time is 40 minutes.

C. Efficiency study

The structure of SAE used in `ItD` is very simple, with only two linear layers, two bias layers, and the TopK activation function. While it contains a large number of parameters due to the large hidden size, it is efficient during inference as it mainly requires two matrix multiplication operations.

We report the inference time of SAE along with its time ratio relative to the entire image generation process in Table 5. The inference time is measured per prompt, which consists of 77 tokens (the maximum length allowed in SD v1.4 and SD v2.1). The time required to generate a single image is computed by repeating the process 10 times and taking the average. The number of inference steps is set to 50. The results indicate that SAE is highly efficient, accounting for less than 1% of the total image generation time.

We note that unlike other module-based approaches that require increased inference time as more concepts are erased (Lyu et al., 2024; Lee et al., 2025), the inference time of `ItD` is independent of the number of concepts being erased.

	SD v1.4	SD v2.1
Inference Time (1000 prompts)	5.05s	6.37s
Time Ratio (per prompt)	0.22%	0.13%

Table 5. **Efficiency study**. The inference time of SAE along with its time ratio relative to the entire image generation process.

D. Implementation details

D.1. Celebrity Erasure

We select 50 celebrities from 200 celebrities provide in MACE (Lu et al., 2024) as target concepts to erase. The celebrities can be accurately generated by Stable Diffusion v1.4 (Rombach et al., 2022), which have over 99% accuracy of the GIPHY Celebrity Detector (GCD) (Nick Hasty & Korduban, 2025). The 50 target celebrities are listed in Table 6. We also select 100 celebrities as remaining concepts to preserve, as listed in Table 7. To generate their images, we used 5 prompt templates with 5 random seeds (1-5). The prompt templates are distinct for celebrities and artistic styles. We used 0 as a seed generating 5 images from a prompt for characters. The prompt templates are listed in Table 9.

To select features specific to the target celebrities, we adopt the remaining 100 celebrities as well as 1000 captions from COCO-30K as the retain set.

In the case of celebrities erasure, we set the following negative prompts to improve image quality: *"bad anatomy, watermark, extra digit, signature, worst quality, jpeg artifacts, normal quality, low quality, long neck, lowres, error, blurry, missing fingers, fewer digits, missing arms, text, cropped, humpbacked, bad hands, username"*

D.2. Artist Style Erasure

We select 100 artist styles as target concepts to erase, and 100 artist styles as remaining concepts to preserve. The 100 target artistic styles are listed in Table 8 and the remaining concepts are listed in Table 10. To generate their images, we used 5 prompt templates with 5 random seeds (1-5). The prompt templates are listed in Table 9, which are different from celebrity erasure.

To select features specific to the target artist styles, we adopt the remaining 100 artist styles as well as 1000 captions from COCO-30K as the retain set.

Table 6. List of target celebrities. We adopt the same 50 celebrities following (Lee et al., 2025). The selected celebrities have over 99% accuracy by the GIPHY Celebrity Detector (GCD) (Nick Hasty & Korduban, 2025).

# of Celebrities to be erased	Celebrity
50	'Adam Driver', 'Adriana Lima', 'Amber Heard', 'Amy Adams', 'Andrew Garfield', 'Angelina Jolie', 'Anjelica Huston', 'Anna Faris', 'Anna Kendrick', 'Anne Hathaway', 'Arnold Schwarzenegger', 'Barack Obama', 'Beth Behrs', 'Bill Clinton', 'Bob Dylan', 'Bob Marley', 'Bradley Cooper', 'Bruce Willis', 'Bryan Cranston', 'Cameron Diaz', 'Channing Tatum', 'Charlie Sheen', 'Charlize Theron', 'Chris Evans', 'Chris Hemsworth','Chris Pine', 'Chuck Norris', 'Courteney Cox', 'Demi Lovato', 'Drake', 'Drew Barrymore', 'Dwayne Johnson', 'Ed Sheeran', 'Elon Musk', 'Elvis Presley', 'Emma Stone', 'Frida Kahlo', 'George Clooney', 'Glenn Close', 'Gwyneth Paltrow', 'Harrison Ford', 'Hillary Clinton', 'Hugh Jackman', 'Idris Elba', 'Jake Gyllenhaal', 'James Franco', 'Jared Leto', 'Jason Momoa', 'Jennifer Aniston', 'Jennifer Lawrence'

Table 7. List of celebrities to preserve. We adopt the same 100 celebrities following (Lee et al., 2025). The selected celebrities have over 99% accuracy by the GIPHY Celebrity Detector (GCD) (Nick Hasty & Korduban, 2025).

# of Celebrities to be preserve	Celebrity
100	'Aaron Paul', 'Alec Baldwin', 'Amanda Seyfried', 'Amy Poehler', 'Amy Schumer', 'Amy Winehouse', 'Andy Samberg', 'Aretha Franklin', 'Avril Lavigne', 'Aziz Ansari', 'Barry Manilow', 'Ben Affleck', 'Ben Stiller', 'Benicio Del Toro', 'Bette Midler', 'Betty White', 'Bill Murray', 'Bill Nye', 'Britney Spears', 'Brittany Snow', 'Bruce Lee', 'Burt Reynolds', 'Charles Manson', 'Christie Brinkley', 'Christina Hendricks', 'Clint Eastwood', 'Countess Vaughn', 'Dane Dehaan', 'Dakota Johnson', 'David Bowie', 'David Tennant', 'Denise Richards', 'Doris Day', 'Dr Dre', 'Elizabeth Taylor', 'Emma Roberts', 'Fred Rogers', 'George Bush', 'Gal Gadot', 'George Takei', 'Gillian Anderson', 'Gordon Ramsey', 'Halle Berry', 'Harry Dean Stanton', 'Harry Styles', 'Hayley Atwell', 'Heath Ledger', 'Henry Cavill', 'Jackie Chan', 'Jada Pinkett Smith', 'James Garner', 'Jason Statham', 'Jeff Bridges', 'Jennifer Connelly', 'Jensen Ackles', 'Jim Morrison', 'Jimmy Carter', 'Joan Rivers', 'John Lennon', 'Jon Hamm', 'Judy Garland', 'Julianne Moore', 'Justin Bieber', 'Kaley Cuoco', 'Kate Upton', 'Keanu Reeves', 'Kim Jong Un', 'Kirsten Dunst', 'Kristen Stewart', 'Krysten Ritter', 'Lana Del Rey', 'Leslie Jones', 'Lily Collins', 'Lindsay Lohan', 'Liv Tyler', 'Lizzy Caplan', 'Maggie Gyllenhaal', 'Matt Damon', 'Matt Smith', 'Matthew Mcconaughey', 'Maya Angelou', 'Megan Fox', 'Mel Gibson', 'Melanie Griffith', 'Michael Cera', 'Michael Ealy', 'Natalie Portman', 'Neil Degrasse Tyson', 'Niall Horan', 'Patrick Stewart', 'Paul Rudd', 'Paul Wesley', 'Pierce Brosnan', 'Prince', 'Queen Elizabeth', 'Rachel Dratch', 'Rachel Mcadams', 'Reba Mcentire', 'Robert De Niro'

E. Ablation studies for `ItD`

E.1. The Effect of strength λ

In this section, we investigate the impact of the strength parameter λ on SAE's effectiveness as a classifier for distinguishing between target and remaining concepts, as well as its ability to erase unwanted knowledge during text encoder inference.

We vary λ from -8 to 0 and conduct experiments on celebrity and artistic style erasure tasks. Specifically, we measure SAE's accuracy in correctly identifying remaining concepts, conditioned on successfully identifying all target concepts. This is analogous to the true negative rate (TN) under the condition that the false negative rate (FN) is zero. A higher value indicates greater effectiveness, ensuring that SAE would not misclassify normal concepts as target concepts. The results, presented in Figure 6, demonstrate that our approach is robust to the choice of τ. When $\tau < -1$, SAE performs well across the remaining 100 celebrities, 100 artistic styles, and COCO-30K. Moreover, when $\tau < -2$, SAE correctly classifies over

Table 8. List of target artist styles. We adopt the same 100 artist styles following (Lee et al., 2025). All artistic styles in these images were successfully generated using SD v1.4.

# of Artist Styles to be erased	Artist Style
100	'Brent Heighton', 'Brett Weston', 'Brett Whiteley', 'Brian Bolland', 'Brian Despain', 'Brian Froud', 'Brian K. Vaughan', 'Brian Kesinger', 'Brian Mashburn', 'Brian Oldham', 'Brian Stelfreeze', 'Brian Sum', 'Briana Mora', 'Brice Marden', 'Bridget Bate Tichenor', 'Briton Riviere', 'Brooke Didonato', 'Brooke Shaden', 'Brothers Grimm', 'Brothers Hildebrandt', 'Bruce Munro', 'Bruce Nauman', 'Bruce Pennington', 'Bruce Timm', 'Bruno Catalano', 'Bruno Munari', 'Bruno Walpoth', 'Bryan Hitch', 'Butcher Billy', 'C. R. W. Nevinson', 'Cagnaccio Di San Pietro', 'Camille Corot', 'Camille Pissarro', 'Camille Walala', 'Canaletto', 'Candido Portinari', 'Carel Willink', 'Carl Barks', 'Carl Gustav Carus', 'Carl Holsoe', 'Carl Larsson', 'Carl Spitzweg', 'Carlo Crivelli', 'Carlos Schwabe', 'Carmen Saldana', 'Carne Griffiths', 'Casey Weldon', 'Caspar David Friedrich', 'Cassius Marcellus Coolidge', 'Catrin WelzStein', 'Cedric Peyravernay', 'Chad Knight', 'Chantal Joffe', 'Charles Addams', 'Charles Angrand', 'Charles Blackman', 'Charles Camoin', 'Charles Dana Gibson', 'Charles E. Burchfield', 'Charles Gwathmey', 'Charles Le Brun', 'Charles Liu', 'Charles Schridde', 'Charles Schulz', 'Charles Spencelayh', 'Charles Vess', 'Charles-Francois Daubigny', 'Charlie Bowater', 'Charline Von Heyl', 'Cha "im Soutine', 'Chen Zhen', 'Chesley Bonestell', 'Chiharu Shiota', 'Ching Yeh', 'Chip Zdarsky', 'Chris Claremont', 'Chris Cunningham', 'Chris Foss', 'Chris Leib', 'Chris Moore', 'Chris Ofili', 'Chris Saunders', 'Chris Turnham', 'Chris Uminga', 'Chris Van Allsburg', 'Chris Ware', 'Christian Dimitrov', 'Christian Grajewski', 'Christophe Vacher', 'Christopher Balaskas', 'Christopher Jin Baron', 'Chuck Close', 'Cicely Mary Barker', 'Cindy Sherman', 'Clara Miller Burd', 'Clara Peeters', 'Clarence Holbrook Carter', 'Claude Cahun', 'Claude Monet', 'Clemens Ascher'

Table 9. Prompt templates used for celebrity image generation. For each prompt, we generate 5 images with seed ranges from 1-5. Additionally, for celebrities, we set the following negative prompts to improve image quality: *"bad anatomy, watermark, extra digit, signature, worst quality, jpeg artifacts, normal quality, low quality, long neck, lowres, error, blurry, missing fingers, fewer digits, missing arms, text, cropped, humpbacked, bad hands, username"*

Domain	Seed	Prompt
Celebrity	1-5	*A portrait of {celebrity name}* *A sketch of {celebrity name}* *An oil painting of {celebrity name}* *{celebrity name} in an official photo* *An image capturing {celebrity name} at a public event*
Artist Style	1-5	*Image in the style of {artist name}* *Art inspired by {artist name}* *Painting in the style of { artist name }* *A reproduction of art by {artist name }* *A famous artwork by {artist name}*

95% of prompts in DiffusionDB-10K as normal concepts.

To demonstrate the effectiveness of λ in erasing target knowledge, we present generated images of target concepts with different λ. Figure 7 show the results experiment on sd v1.4 (Rombach et al., 2022) and and Figure 8 show the results experiment on sd v2.1 (Face & CompVis, 2023b). The results show that $\lambda = -2$ is sufficient to erase knowledge about celebrity and artist style concepts. The knowledge about "nudity" can be erased when $\lambda = -4$.

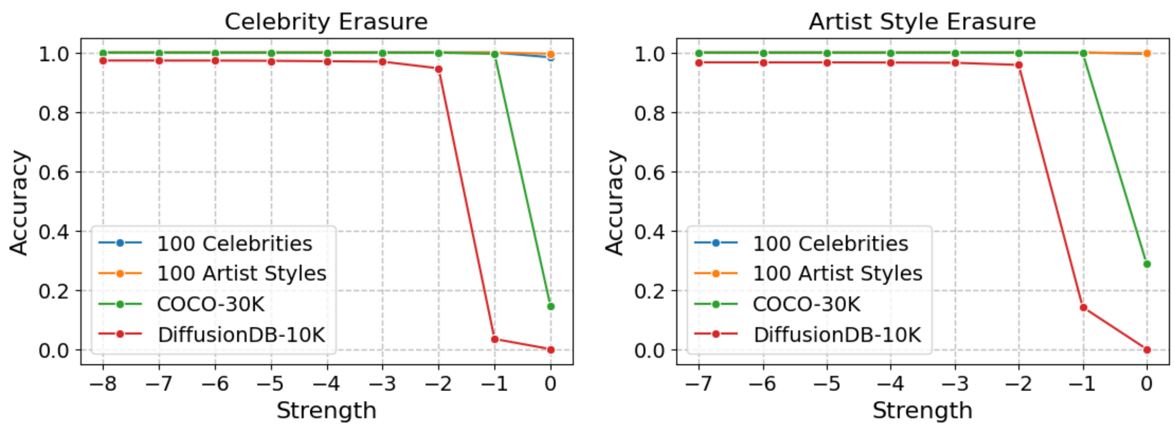


Figure 6. **Ablation study on the effect of strength λ in SAE classification for distinguishing target and remaining concepts.** The threshold τ is set to ensure zero false negatives; therefore, we primarily report accuracy on the remaining concepts. A higher accuracy indicates that SAE would not misclassify normal concepts as target concepts.

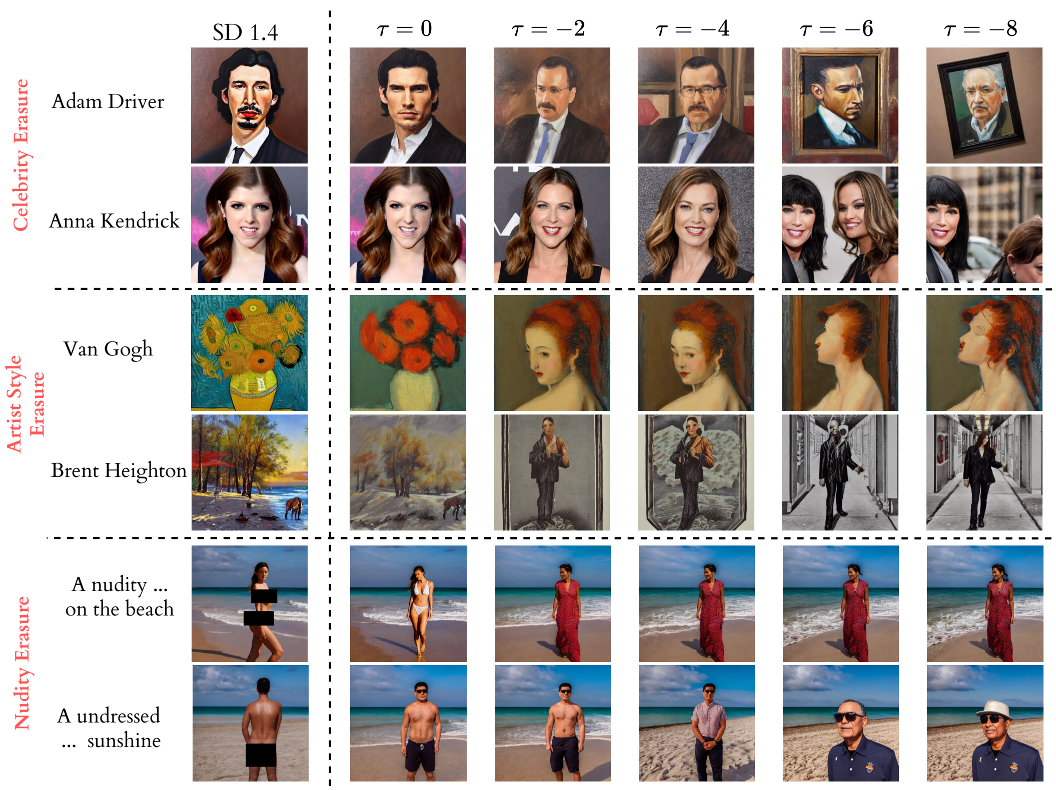


Figure 7. **Ablation study on the effect of strength λ in erasing unwanted knowledge.** Experiments on SD v1.4 (Rombach et al., 2022).

Table 10. List of artist styles to preserve. We adopt the same 100 artist styles following (Lee et al., 2025). All artistic styles in these images were successfully generated using SD v1.4.

# of Artist styles to preserve	Artist Style
100	‘A.J.Casson’, ‘Aaron Douglas’, ‘Aaron Horkey’, ‘Aaron Jasinski’, ‘Aaron Siskind’, ‘Abbott Fuller Graves’, ’Abbott Handerson Thayer’, ’Abdel Hadi Al Gazzar’, ’Abed Abdi’, ’Abigail Larson’, ’Abraham Mintchine’, ’Abraham Pether’, ’Abram Efimovich Arkhipov’, ’Adam Elsheimer’, ’Adam Hughes’, ’Adam Martinakis’, ’Adam Paquette’, ’Adi Granov’, ’Adolf Hiremy-Hirschl’, ’Adolph Got- ’tlieb’, ’Adolph Menzel’, ’Adonna Khare’, ’Adriaen van Ostade’, ’Adriaen van Outrecht’, ’Adrian Donoghue’, ’Adrian Ghenie’, ’Adrian Paul Allinson’, ’Adrian Smith’, ’Adrian Tomine’, ’Adrianus Eversen’, ’Afarin Sajedi’, ’Affandi’, ’Aggi Erguna’, ’Agnes Cecile’, ’Agnes Lawrence Pelton’, ’Agnes Martin’, ’Agostino Arrivabene’, ’Agostino Tassi’, ’Ai Weiwei’, ’Ai Yazawa’, ’Akihiko Yoshida’, ’Akira Toriyama’, ’Akos Major’, ’Akseli Gallen-Kallela’, ’Al Capp’, ’Al Feldstein’, ’Al Williamson’, ’Alain Laboile’, ’Alan Bean’, ’Alan Davis’, ’Alan Kenny’, ’Alan Lee’, ’Alan Moore’, ’Alan Parry’, ’Alan Schaller’, ’Alasdair McLellan’, ’Alastair Magnaldo’, ’Alayna Lemmer’, ’Albert Benois’, ’Albert Bierstadt’, ’Albert Bloch’, ’Albert Dubois-Pillet’, ’Albert Eckhout’, ’Albert Edelfelt’, ’Albert Gleizes’, ’Albert Goodwin’, ’Albert Joseph Moore’, ’Albert Koetsier’, ’Albert Kotin’, ’Albert Lynch’, ’Albert Marquet’, ’Albert Pinkham Ryder’, ’Albert Robida’, ’Albert Servaes’, ’Albert Tucker’, ’Albert Watson’, ’Alberto Biasi’, ’Alberto Burri’, ’Alberto Giacometti’, ’Alberto Magnelli’, ’Alberto Seveso’, ’Alberto Sughi’, ’Alberto Vargas’, ’Albrecht Anker’, ’Albrecht Durer’, ’Alec Soth’, ’Alejandro Burdisio’, ’Alejandro Jodorowsky’, ’Aleksey Savrasov’, ’Aleksi Briclot’, ’Alena Aenami’, ’Alessandro Allori’, ’Alessandro Barbucci’, ’Alessandro Gottardo’, ’Alessio Albi’, ’Alex Alemany’, ’Alex Andreev’ ’Alex Colville’, ’Alex Figini’, ’Alex Garant’

E.2. Selection of Residual Stream to perform SAE.

The text encoder consists of 12 transformer blocks (`layers.0-11`) connected sequentially. Each residual stream of these layers can be used to train SAE. However, since different layers may capture different types of knowledge (Jin et al., 2024), applying SAE to different residual streams results in varying performance.

To identify the most approximate layer for concept erasing, we conduct experiments on different residual streams. The training setup follows the details provided in Appendix B, with the only variation being the choice of residual stream. We evaluate the task of erasing 50 celebrities and 100 artist styles separately. The remaining concepts consist of 100 other celebrities and 100 other artistic styles, COCO-30K, and DiffusionDB-10K. For efficiency, similar to Figure 6, we use accuracy as a metric to assess the impact of different layers. This metric quantifies the proportion of correctly identified remaining concepts when SAE is used as a classifier.

The results are summarized in Table 9. For remaining celebrities, artistic styles, and COCO-30K, the accuracy is approximately 100%, demonstrating that SAE, as a classifier, effectively identifies remaining concepts regardless of the training layer. Even for DiffusionDB-10K, which exhibits the lowest performance, the accuracy remains at least 96%.

We also provide quantitative results to better understand the effects of erasing at different layers. Our experiments cover three domains: celebrities, artistic styles, and nudity content. The erasure strength τ is set to -6 for celebrities and nudity and -2 for artistic styles. As shown in Figure 11, all layers can be used to erase target concepts. However, certain layers (e.g., `layers.2-4`) still generate images with exposed chests when attempting to erase the concept of ”nudity”.

For the celebrity domain, we use the prompts *”An oil painting of Adam Driver”* and *”Anna Kendrick in an official photo.”*. The results indicate that applying SAE to `layers.6-10` still allows the model to generate a person within the corresponding context (*”oil painting”* and *”official photo”*), suggesting that erasing at these layers preserves the structural integrity of the scene while removing the target identity. For artistic style erasure, applying SAE to `layers.6-10` also results in the generation of a photo of a person, even though the original images did not contain any people.

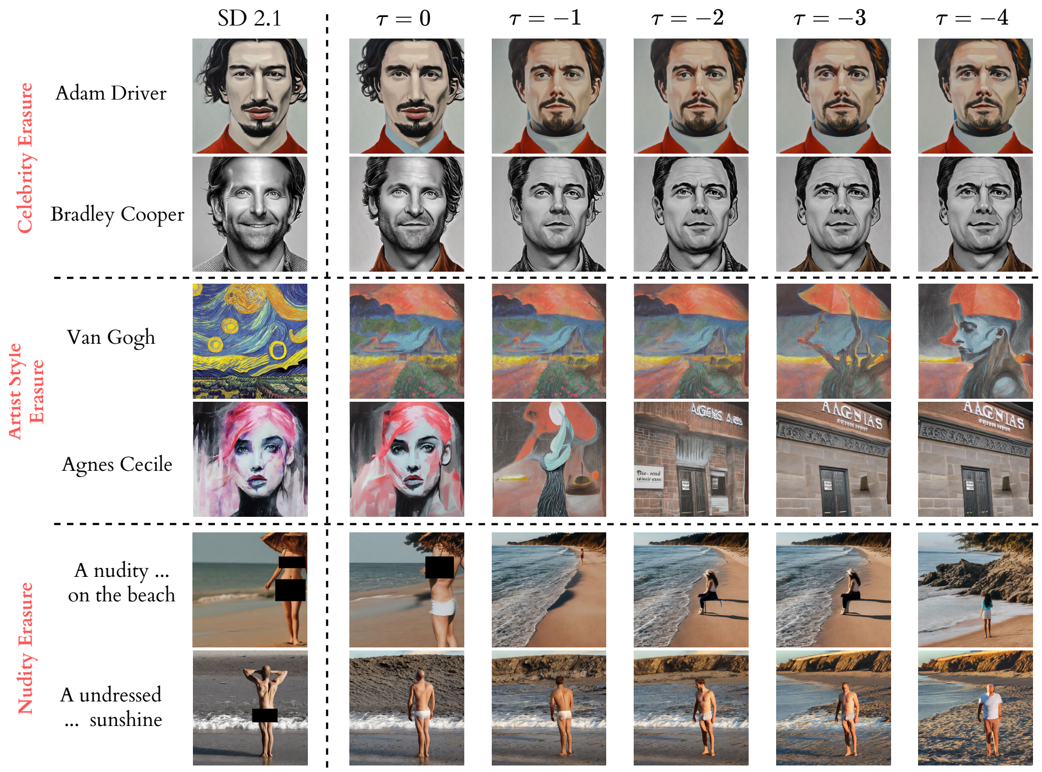


Figure 8. **Ablation study on the effect of strength λ in erasing unwanted knowledge.** Experiments on SD v2.1 (Face & CompVis, 2023b).

E.3. The effect of hidden size d_{hid}

We investigate the effect of SAE with different hidden sizes for concept erasing. We vary d_{hid} from 2^{15} to 2^{19}, which is about 43 to 683 times larger than the hidden dimension of SD v1-4 (Rombach et al., 2022). We present the feature density of the SAE as well as the reconstruction loss for target and remaining concepts in Figure 10. The results indicate that the log feature density is relatively higher for smaller hidden sizes compared to larger hidden sizes, suggesting that SAE with small hidden layers learns denser features. However, this increased feature density leads to poorer performance in distinguishing between target and remaining concepts, as shown in the histogram of reconstruction loss. This may be because larger hidden layers are capable of learning more fine-grained features, which in turn enhance the ability to differentiate between different concepts.

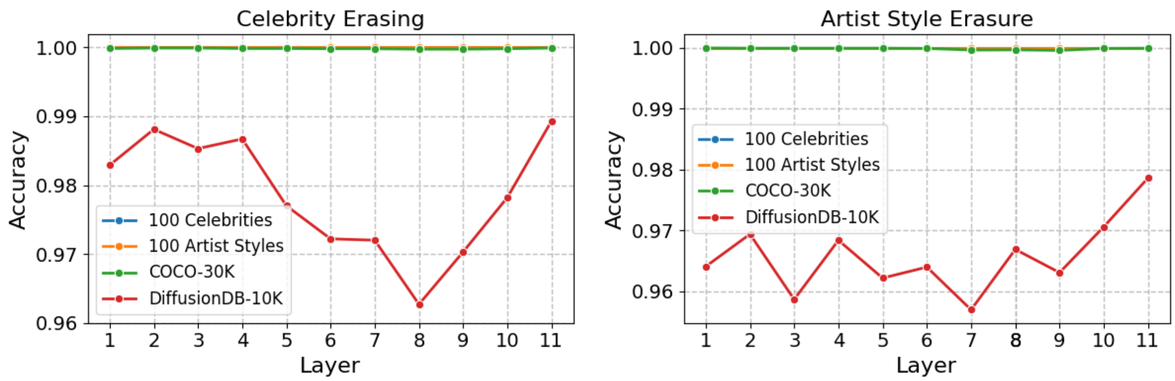
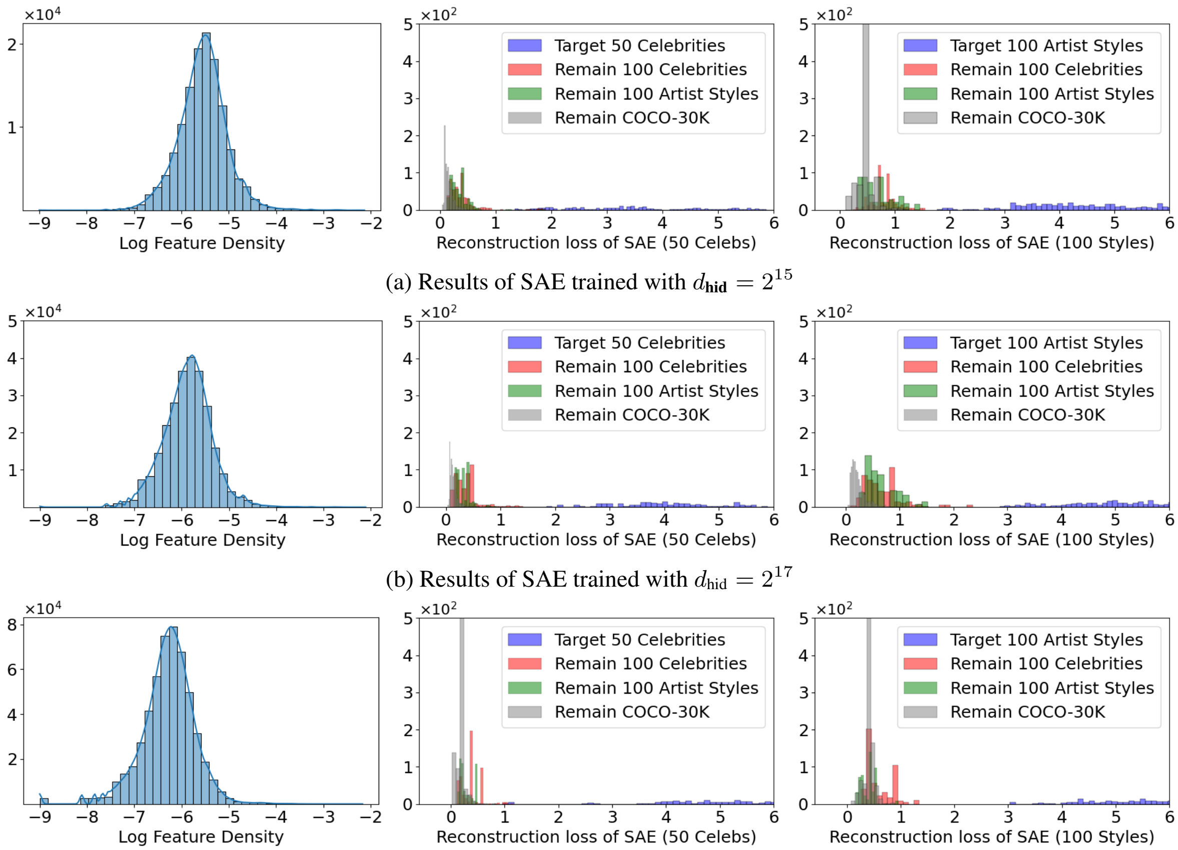


Figure 9. The log feature density is calculated as $\log(n/N + 10^{-9})$, where n represents the number of tokens that activate the feature, and N denotes the total number of tested tokens.



(a) Results of SAE trained with $d_{\textbf{hid}} = 2^{15}$

(b) Results of SAE trained with $d_{\text{hid}} = 2^{17}$

(c) Results of SAE trained with $d_{\textbf{hid}} = 2^{19}$

Figure 10. The log feature density and the reconstruction loss for SAEs trained with different hidden sizes.

Figure 11. Qualitative results of performing `ItD` on different transformer blocks.

F. Additional Qualitative results

F.1. Example 1 of celebrities erasure.

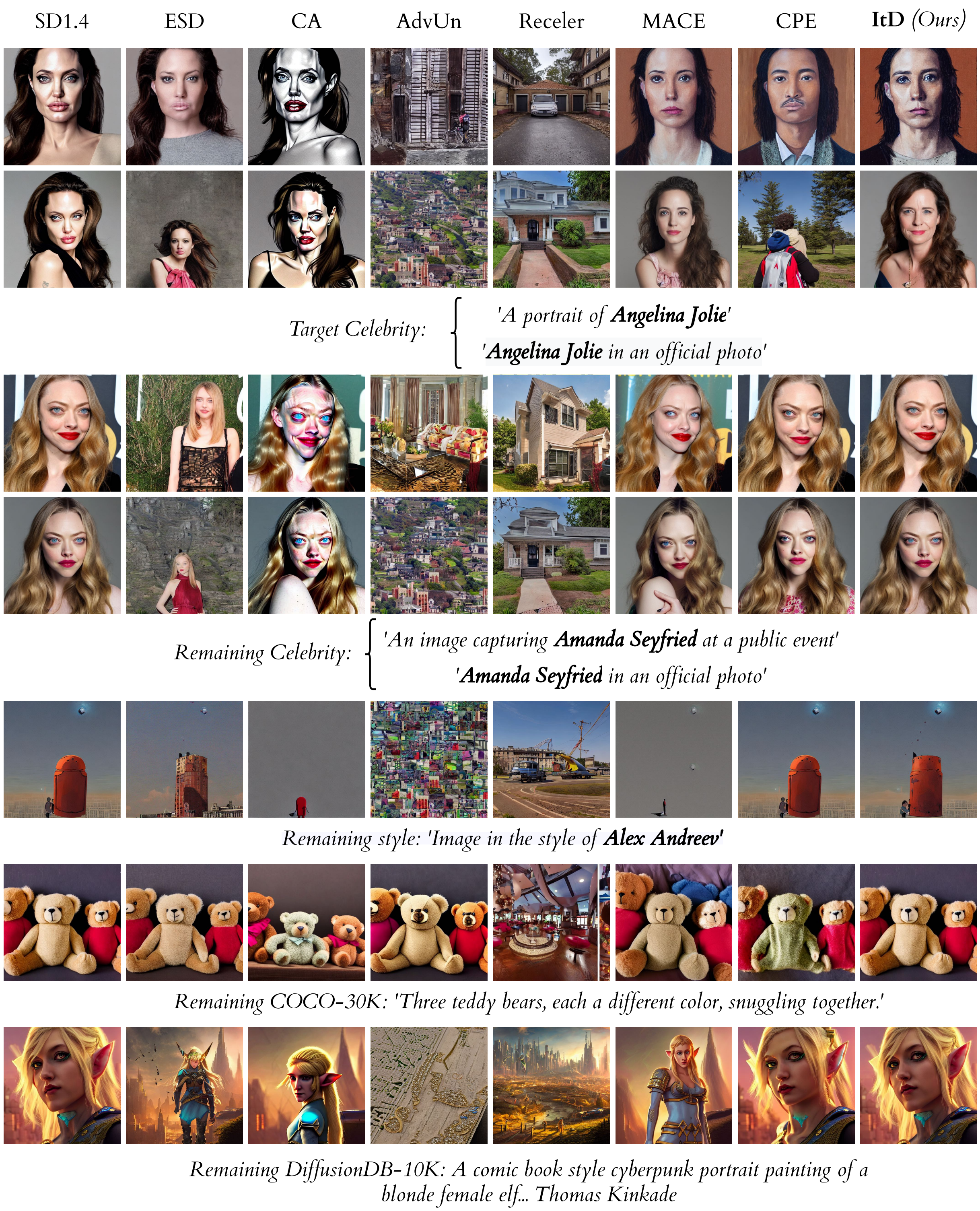


Figure 12. Qualitative comparison on celebrities erasure. The images on the same row are generated using the same seed.

F.2. Example 2 of celebrities erasure.

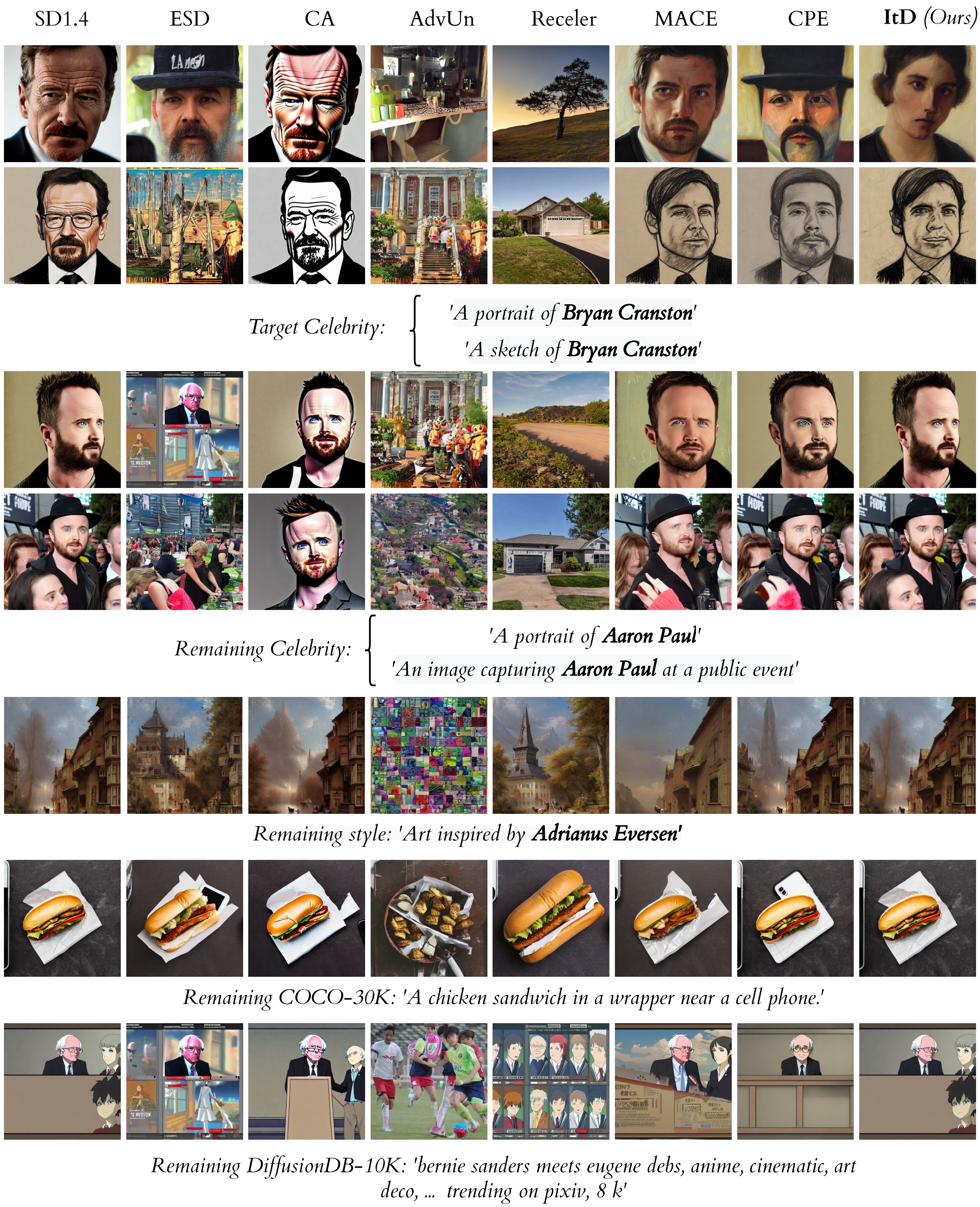


Figure 13. Qualitative comparison on celebrities erasure. The images on the same row are generated using the same seed.

F.3. Example 1 of artist styles erasure.

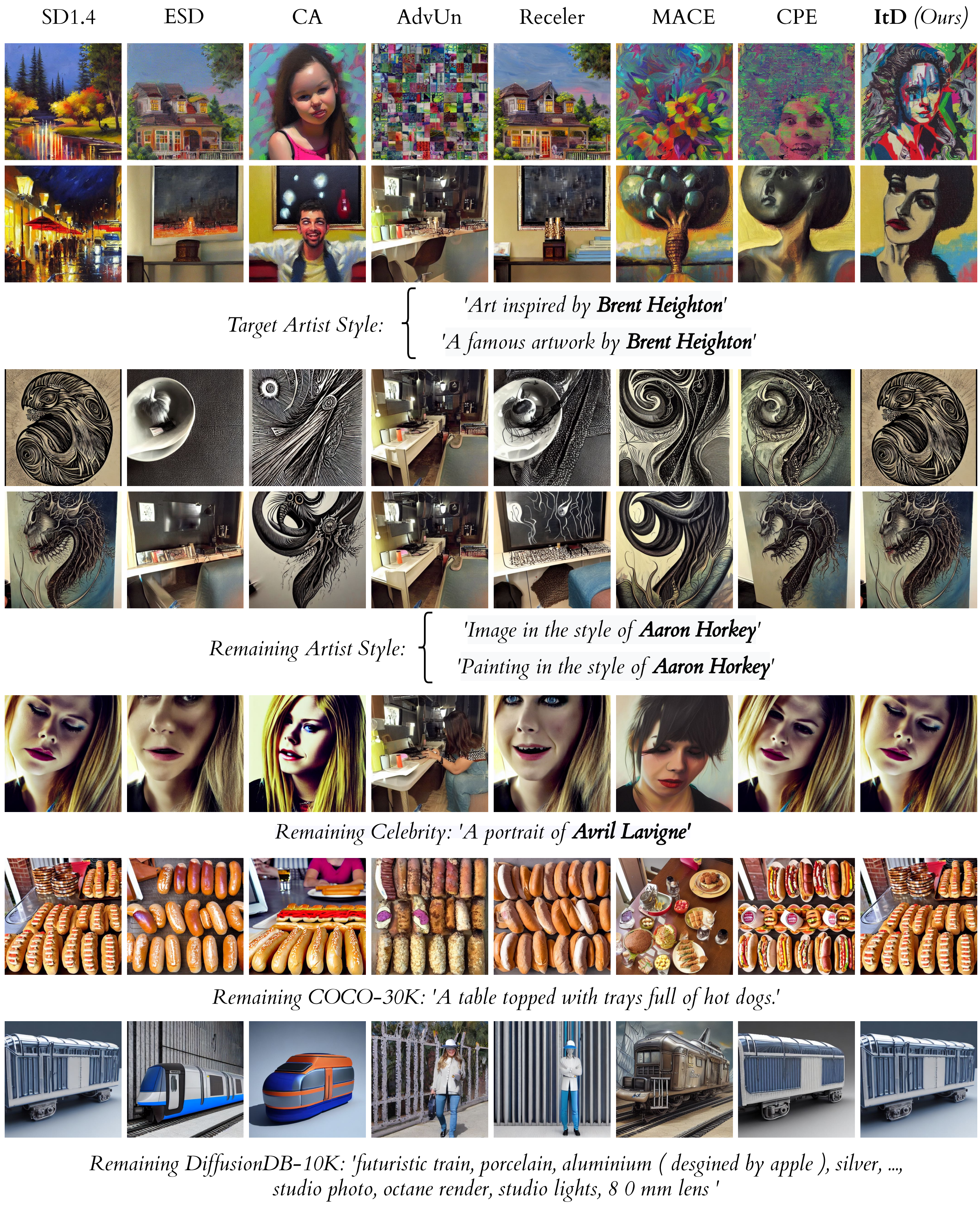


Figure 14. Qualitative comparison on artist styles erasure. The images on the same row are generated using the same seed.

F.4. Example 2 of artist styles erasure.

Figure 15. Qualitative comparison on artist styles erasure. The images on the same row are generated using the same seed.