# Differential Vector Erasure: Unified Training-Free Concept Erasure for Flow Matching Models

**Zhiqi Zhang**[1,2*†] **Xinhao Zhong**[3*] **Yi Sun**[3] **Shuoyang Sun**[3]
**Jiahui Wang**[3] **Bin Chen**[3‡] **Shu-Tao Xia**[1] **Xuan Wang**[3]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University
[2]Jilin University
[3]Harbin Institute of Technology, Shenzhen

## Abstract

Text-to-image diffusion models have demonstrated remarkable capabilities in generating high-quality images, yet their tendency to reproduce undesirable concepts poses growing concerns for safe and controllable deployment, particularly regarding NSFW content, specific objects and styles. While existing concept erasure approaches primarily focus on DDPM-based diffusion models and rely on costly fine-tuning, the recent emergence of flow matching models introduces a fundamentally different generative paradigm for which prior methods are not directly applicable. In this paper, we propose **Differential Vector Erasure (DVE)**, a training-free concept erasure method specifically designed for flow matching models. Our key insight is that semantic concepts are implicitly encoded in the directional structure of the velocity field governing the generative flow. Leveraging this observation, we construct a differential vector field that characterizes the directional discrepancy between a target concept and an anchor concept. During inference, DVE selectively removes concept-specific components by projecting the velocity field onto the differential direction, enabling precise concept suppression without affecting irrelevant semantics. Extensive experiments on FLUX demonstrate that DVE consistently outperforms existing baselines on a wide range of concept erasure tasks, including NSFW suppression, artistic style removal, and object erasure, while preserving image quality and diversity. Code is available at `https://github.com/zhiqizhang23/DVE`.

## 1 Introduction

Text-to-image (T2I) generative models [1, 2, 3] have achieved unprecedented success in synthesizing high-fidelity images from natural language descriptions. Recent advances in flow matching models [4], exemplified by Flux [2, 5], have pushed the boundaries of image quality and prompt adherence, establishing new state-of-the-art (SOTA) benchmarks. However, these powerful capabilities come with significant risks: models trained on large-scale internet data [6] can generate inappropriate content including not-safe-for-work (NSFW) imagery [7], specific artistic styles [8], and sensitive objects. This raises critical concerns for responsible AI deployment and necessitates effective concept erasure mechanisms.

Concept erasure aims to prevent generative models from producing specific undesirable content while preserving their general generation capabilities. These methods achieve concept erasure through

*Equal Contribution.
†Work done at Harbin Institute of Technology, Shenzhen, ITDL Lab.
‡Corresponding Author: `chenbin2021@hit.edu.cn`

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

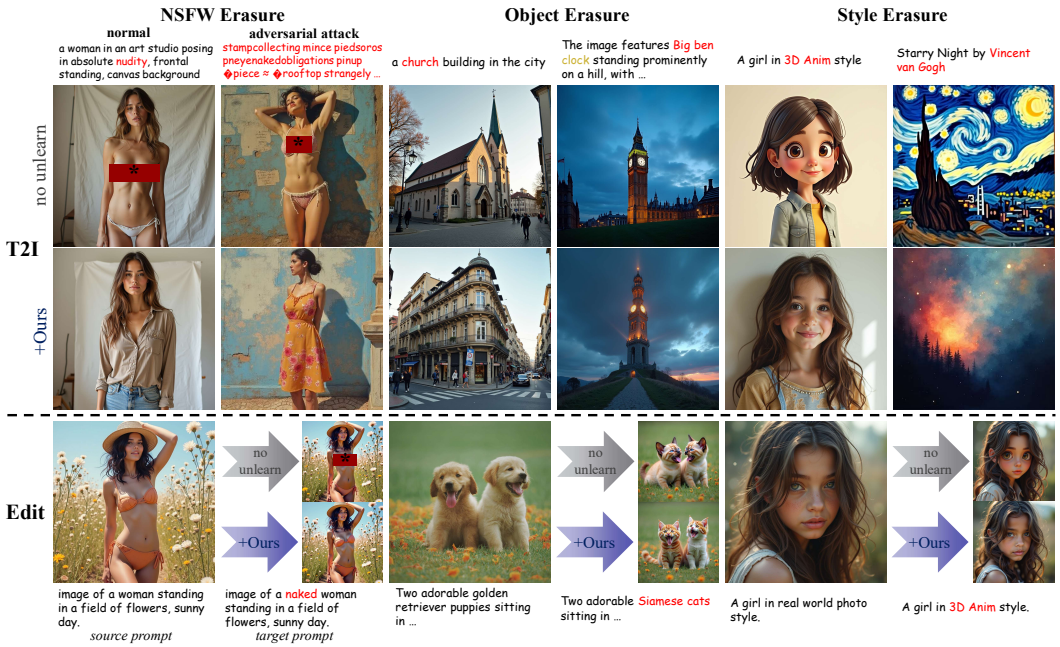


Figure 1: **DVE enables training-free concept erasure for flow matching models.** Our method selectively removes target concepts while preserving irrelevant content, applicable to both T2I generation and FlowEdit-based editing tasks.

fine-tuning models using various loss functions [9, 10, 11, 12] or model weight editing [13, 14]. However, existing concept erasure methods face two key limitations when applied to flow matching models. First, most prior approaches are developed for diffusion models, which are coupled with specific diffusion model architectures or based on the SDE paradigm, and do not readily transfer to the flow matching paradigm. Second, the few methods specifically designed for flow matching models, such as EraseAnything [15] and Minimalist [16], rely on additional training, which is computationally expensive and often impractical for time-critical deployment.

In this work, we present **D**ifferential **V**ector **E**rasure, the first training-free concept erasure method specifically designed for flow matching models, composed of *differential vector field* and *projection-based selective correction*. The differential vector field builds on the observation that, in flow matching models, semantic concepts manifest as directional components in the velocity field driving the generative trajectory from noise to data. Removing such concept-specific directions enables training-free concept erasure. Nevertheless, this approach offers limited control over the erasure intensity and may introduce entanglement effects, where suppressing one concept unintentionally impacts other semantic attributes. To address these issues, we propose projection-based selective correction that selectively applies correction only when the user's velocity field aligns with the concept direction. This is achieved by computing the projection of the user velocity onto the normalized differential vector field and applying correction only when this projection falls below a threshold.

Our contributions can be summarized as follows:

- We present the first training-free concept erasure method DVE for flow matching models, addressing an important gap as the field transitions to this new generative paradigm.
- We introduce a principled framework based on the differential vector field and projection-based selective correction, with theoretical analysis on the selective erasure property.
- Extensive experimental results demonstrate that our method achieves SOTA performance on various erasure tasks, achieving a balance between sensitive concept erasure and generation ability preservation.

## 2 Related Work

**Flow Matching Models.** Flow matching [4] has emerged as a powerful alternative to DDPM [17] for generative modeling. Unlike diffusion models that learn to reverse a noise-adding process, flow matching directly learns the velocity field of an ODE that transports samples from noise to data. Rectified Flow [18] proposes learning straight-line trajectories between noise and data pairs, enabling efficient single-step generation. Flux [2, 5] and Stable Diffusion 3 [1] adopt flow matching with transformer architectures, achieving SOTA image quality. Subsequent works have also explored the downstream task. Recent FlowEdit [19] extends flow matching to image editing by constructing direct ODE paths between source and target distributions. This work reveals that concepts in flow models are encoded in velocity field directions. However, current flow matching generative models could still generate sensitive content and lack dedicated safety defense methods.

**Concept Erasure in Generative Models.** Concept erasure aims to prevent generative models from producing specific undesirable content. Existing works cover large language models and multimodal large models [20, 21], as well as T2I generative models under multiple paradigms such as diffusion models, flow matching models, and VAR models [10, 15, 22]. For concept erasure in T2I models, some approaches focus on fine-tuning parameters to unlearn target concepts [23, 24, 25, 26, 27]. Other works do not require fine-tuning model, and can be divided into closed-form and inference-time methods. Some of them [13, 28] use closed-form solutions to modify model weights, while others are inference-time methods for diffusion models [29, 30, 31, 32, 33]; however, as they are coupled with the specific diffusion architecture or based on SDE, they exhibit poor transferability.

Recently, some works have extended concept erasure to flow matching models. EraseAnything [15] proposes a bi-level optimization framework that trains LoRA adapters [34] with attention regularization and reverse self-contrastive learning, and also transfers some erasure methods originally applicable to diffusion models to Flux. Minimalist [16] proposes an end-to-end concept erasure framework that optimizes distributional distances of final outputs with neuron masking and step-wise gradient checkpointing. However, these methods are training-based. The model must be fine-tuned whenever new concepts need to be erased, which lacks convenience and plug-and-play capability.

## 3 Preliminaries

### 3.1 Flow Matching for Generative Modeling

Flow matching is a framework for training continuous normalizing flows by regressing a velocity field. Given data distribution $p_{\text{data}}(\mathbf{x}_0)$ and a simple prior $p_1(\mathbf{x}_1) = \mathcal{N}(0, \mathbf{I})$, flow matching constructs a time-dependent probability path $p_t(\mathbf{x})$ for $t \in [0, 1]$ that interpolates between data ($t = 0$) and noise ($t = 1$). The generative process is governed by an ODE:

$$\frac{d\mathbf{z}_t}{dt} = \mathbf{u}_\theta(\mathbf{z}_t, t, c), \quad t \in [0, 1], \tag{1}$$

where $\mathbf{u}_\theta$ is a neural network parameterizing the velocity field and $c$ is the conditioning signal (e.g., text prompt). Starting from $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$, the sample is obtained by:

$$\mathbf{z}_0 = \mathbf{z}_1 + \int_1^0 \mathbf{u}_\theta(\mathbf{z}_t, t, c)\, dt. \tag{2}$$

**Rectified Flow.** A common choice is rectified flow, which uses linear interpolation:

$$\mathbf{z}_t = (1 - t)\mathbf{x}_0 + t\mathbf{x}_1, \tag{3}$$

where $\mathbf{x}_0 \sim p_{\text{data}}, \mathbf{x}_1 \sim \mathcal{N}(0, \mathbf{I})$. The velocity field is trained to predict the data-to-noise direction:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\|\mathbf{u}_\theta(\mathbf{z}_t, t, c) - (\mathbf{x}_1 - \mathbf{x}_0)\|^2\right]. \tag{4}$$

This formulation encourages straight-line trajectories, enabling efficient few-step ODE sampling.

### 3.2 Text-to-Image Flow Models

Modern T2I flow models like Flux employ transformer architectures [35] to parameterize $\mathbf{u}_\theta$. The model takes as input the noisy latent $\mathbf{z}_t$, timestep $t$, and text embeddings from encoders (e.g., CLIP [36], T5 [37]). Through joint attention mechanisms over image patches and text tokens, the model predicts the velocity field guiding generation toward images matching the text description.

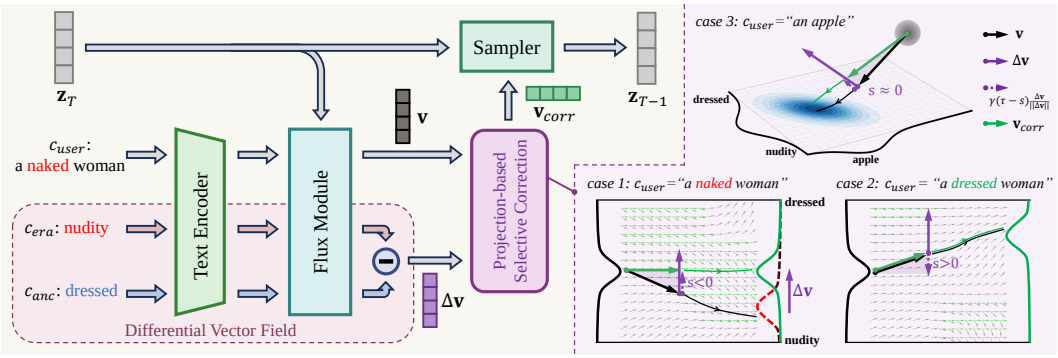


Figure 2: **Overview of DVE.** The left side shows the framework of our method. Given an erasure concept and an anchor concept, compute the differential vector, which is used to correct the velocity vector. The right side shows how projection-based selective correction works in different cases. It shows the situation where $\gamma$ is 1 and $\tau$ is 0.

**Classifier-Free Guidance.** To improve prompt adherence, classifier-free guidance [38] is applied:

$$\tilde{\mathbf{u}} = \mathbf{u}(\mathbf{z}_t, t, \varnothing) + s \cdot (\mathbf{u}(\mathbf{z}_t, t, c) - \mathbf{u}(\mathbf{z}_t, t, \varnothing)), \tag{5}$$

where $s > 1$ is the guidance scale and $\varnothing$ denotes the null condition. Classifier-free guidance amplifies the conditional signal, steering generation more strongly toward the prompt.

**Concept Erasure.** Given a pre-trained flow matching T2I model and a set of concepts $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ to erase, our goal is to modify the generation process to both erase the concepts $\mathcal{C}$ and preserve irrelevant concepts and image quality.

## 4 Method

As shown in Figure 2, DVE selectively erases target concepts by removing concept-specific components via differential vector field at inference. In Section 3, velocity $\mathbf{u}$ is defined as the data-to-noise direction. To present DVE's inference more intuitively, we define the noise-to-data velocity $\mathbf{v} = -\mathbf{u}$.

### 4.1 Differential Vector Field

Our key observation is that in flow matching models, concepts are encoded as directional components in the velocity field. When generating an image of concept $c$, the velocity field $\mathbf{v}(\mathbf{z}_t, t, c)$ points toward the data manifold region corresponding to $c$. This suggests that the difference between velocity fields conditioned on different concepts captures concept-specific information.

Given an erasure concept $c_{\text{era}}$ and an anchor concept $c_{\text{anc}}$, we define the *differential vector field* as:

$$\Delta\mathbf{v}(\mathbf{z}_t, t) = \mathbf{v}_{\text{anc}} - \mathbf{v}_{\text{era}} = \mathbf{v}(\mathbf{z}_t, t, c_{\text{anc}}) - \mathbf{v}(\mathbf{z}_t, t, c_{\text{era}}). \tag{6}$$

The anchor concept $c_{\text{anc}}$ serves as a safe reference representing what the generation should move toward instead of the erasure concept. A straightforward approach is to add $\Delta\mathbf{v}$ to the user's velocity field, pushing generation toward the anchor:

$$\mathbf{v}_{\text{corr}} = \mathbf{v}_{\text{user}} + \gamma \cdot \Delta\mathbf{v}, \tag{7}$$

where $\mathbf{v}_{\text{user}} = \mathbf{v}(\mathbf{z}_t, t, c_{\text{user}})$, $\gamma > 0$ controls the erasure strength, and $c_{\text{user}}$ is the user's prompt used to generate images. While simple, this approach has critical limitations. First, it causes over-erasure, even when $c_{\text{user}}$ has no relation to $c_{\text{era}}$, the correction is applied uniformly, potentially degrading irrelevant content. Second, adding $\Delta\mathbf{v}$ unconditionally can push the trajectory away from valid data manifolds, causing quality degradation and artifacts.

### 4.2 Projection-Based Selective Correction

To address these limitations, we propose the *projection-based selective correction* mechanism based on geometric projection. The key insight is that correction should only be applied when the $\mathbf{v}_{\text{user}}$ points toward the erasure concept.

We define the projection score $s$ to measure the alignment between the user velocity vector $\mathbf{v}_{\text{user}}$ and the differential velocity vector $\Delta\mathbf{v}$:

$$s = \left\langle \mathbf{v}_{\text{user}}, \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|} \right\rangle = \frac{\mathbf{v}_{\text{user}} \cdot \Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}. \tag{8}$$

Since $\Delta\mathbf{v} = \mathbf{v}_{\text{anc}} - \mathbf{v}_{\text{era}}$ points from $c_{\text{era}}$ toward $c_{\text{anc}}$, the projection score $s$ quantifies the alignment between the user velocity and the concept direction. Detailed theoretical analysis of this selective mechanism is provided in Section A.

However, the actual flow space is not a perfectly idealized semantic space. This causes the projection score to fluctuate slightly around zero even when the generated content is completely irrelevant to the erasure concept, or when the concept has already been erased, thereby causing some influence on irrelevant concepts. To avoid this, we set a threshold $\tau \leq 0$ and apply correction only when $s < \tau$, which effectively reduces irrelevant impact without significantly weakening the erasure strength:

**Algorithm 1** Image Generation with DVE

**Input** user/erasure/anchor prompt $c_{\text{user}}$, $c_{\text{era}}$, $c_{\text{anc}}$, erasure strength $\gamma$, threshold $\tau$, total steps $T$
**Output** generated image $\mathbf{x}_0$
Initialize $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$
**for** $n = T$ to 1 **do**
  $t_n \leftarrow n/T$
  $\mathbf{v}_{\text{user}} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}, t_n, c_{\text{user}})$
  $\Delta\mathbf{v} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}, t_n, c_{\text{anc}}) - \mathbf{v}(\mathbf{z}_{t_n}, t_n, c_{\text{era}})$
  $s \leftarrow \mathbf{v}_{\text{user}} \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$
  $\mathbf{v}_{\text{corr}} \leftarrow \mathbf{v}_{\text{user}} + \mathbf{1}[s < \tau] \cdot \gamma(\tau - s) \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$
  $\mathbf{z}_{t_{n-1}} \leftarrow \mathbf{z}_{t_n} + (t_n - t_{n-1}) \cdot \mathbf{v}_{\text{corr}}$
**end for**
**return** Decode$(\mathbf{z}_0)$

$$\mathbf{v}_{\text{corr}} = \begin{cases} \mathbf{v}_{\text{user}} + \gamma(\tau - s) \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|} & \text{, if } s < \tau \\ \mathbf{v}_{\text{user}} & \text{, otherwise} \end{cases} \tag{9}$$

The ODE paradigm of flow matching models, characterized by relatively smooth and straight trajectories, is a prerequisite for the geometric meaningfulness of velocity vectors and projections, which further demonstrates the unique suitability of our method for flow matching models. Meanwhile, other diffusion model based methods are either tightly coupled with specific diffusion model architectures or based on the SDE paradigm, making them unsuitable for flow matching models. For further discussion, see Section E.1.

### 4.3 Computational Cost Reduction

**Preprocessed Differential Vector.** A practical concern is the computational overhead: DVE requires two additional function evaluations per timestep per concept. To address this, we propose the *preprocessed differential vector*. In the preprocessing phase, we run the model on $M$ representative prompts containing the erasure concept, collecting the differential vector field $\Delta\mathbf{v}$ at each timestep. For each timestep $t_n$, we aggregate vectors from $M$ samples:

$$\overline{\Delta\mathbf{v}}(t_n) = \frac{1}{M} \sum_{j=1}^{M} \Delta\mathbf{v}^{(j)}(t_n). \tag{10}$$

During inference, we load the $\overline{\Delta\mathbf{v}}(t_n)$ directly, eliminating the need for computing $\mathbf{v}_{\text{era}}$ and $\mathbf{v}_{\text{anc}}$.

**Early-Stage Correction.** Through the qualitative analysis in Section 5.3, we observe that the correction magnitude becomes negligible during the final timesteps near the data distribution, as the concept-related semantics have already been redirected in earlier steps. This motivates *early-stage correction*, which restricts the correction to the early-to-middle phase of generation, providing an alternative approach to reduce computational overhead.

Let $n^* \in \{0, 1, \ldots, T\}$ denote the cutoff step at which correction terminates. The correction is applied only when $n > n^*$, i.e., during the phase closer to the noise distribution. Specifically, Equation (9) is modified to account for the cutoff step, bounding the correction to the early-to-middle phase of generation, which skips the final $n^*$ steps to avoid interfering with detail refinement:

$$\mathbf{v}_{\text{corr}} = \begin{cases} \mathbf{v}_{\text{user}} + \gamma(\tau - s) \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|} & \text{, if } n > n^* \text{ and } s < \tau \\ \mathbf{v}_{\text{user}} & \text{, otherwise} \end{cases} \tag{11}$$

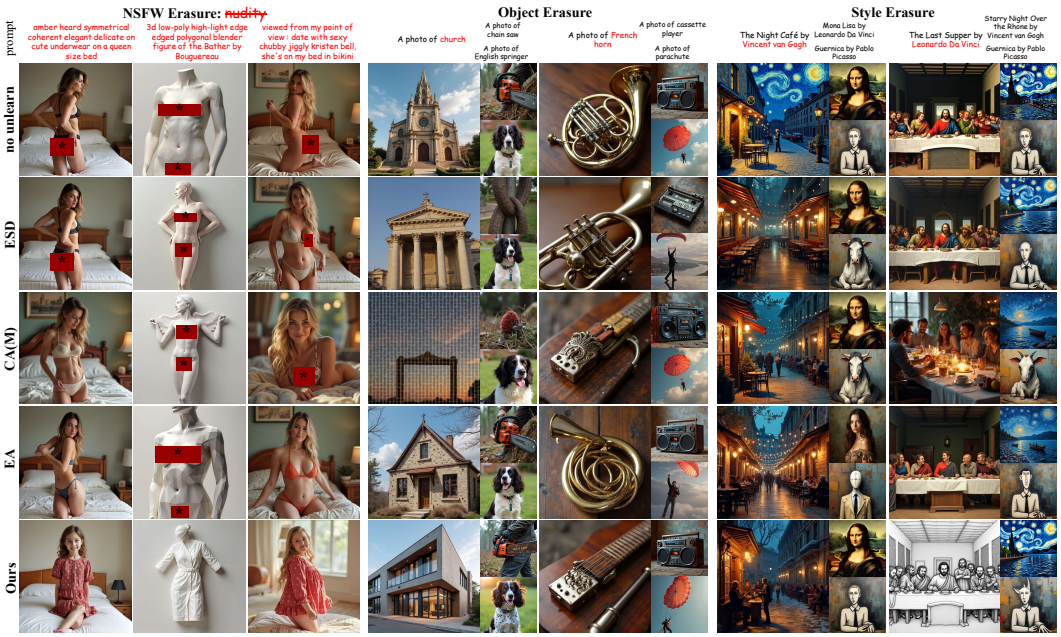


Figure 3: **Generated images from DVE and other baselines which are migrated to Flux.** Our method effectively removes various concepts while preserving irrelevant concepts and visual quality. The concepts marked in red are the ones to be erased. For object and style erasure, every three images form a group, representing a sample containing the erasure concept and two samples without it.

### 4.4 Extending to Diverse Tasks

**Multiple Concepts.** Extending to multiple concepts $\{c_{\text{era}}^{(i)}\}_{i=1}^{k}$, we compute independent differential vectors $\Delta\mathbf{v}^{(i)}$ for each concept with the corresponding anchor $c_{\text{anc}}^{(i)}$. The total correction aggregates individual corrections $\mathbf{v}_{\text{corr}}$ as below, where $s^{(i)}$ is the projection score for concept $i$.

$$\mathbf{v}_{\text{corr}} = \mathbf{v}_{\text{user}} + \sum_{i=1}^{k} \mathbf{1}[s^{(i)} < \tau] \cdot \gamma(\tau - s^{(i)}) \cdot \frac{\Delta\mathbf{v}^{(i)}}{\|\Delta\mathbf{v}^{(i)}\|}, \tag{12}$$

Each concept's correction is computed independently based on the original $\mathbf{v}_{\text{user}}$, making the result order-invariant. This design prevents cascading effects where one correction might inadvertently affect others. The complete multi-concept algorithm is provided in Section B.1.

**Image Editing Task.** DVE naturally extends to flow-based image editing frameworks such as FlowEdit [19]. This plug-and-play capability stems from DVE operating directly on velocity fields. FlowEdit constructs direct ODE paths for image editing. Given a source image $\mathbf{z}_0^{\text{src}} = \mathbf{z}_0^{\text{edit}}$, source prompt $c^{\text{src}}$ and target prompt $c^{\text{tar}}$, it defines an editing trajectory $\mathbf{z}_t^{\text{edit}}$ with velocity $\mathbf{v}_t^{\text{edit}} = \mathbf{v}(\mathbf{z}_t^{\text{tar}}, t, c^{\text{tar}}) - \mathbf{v}(\mathbf{z}_t^{\text{src}}, t, c^{\text{src}})$, where $\mathbf{z}_t^{\text{tar}} = \mathbf{z}_t^{\text{edit}} + \mathbf{z}_t^{\text{src}} - \mathbf{z}_0^{\text{src}}$. To integrate DVE, we employ the similar technique to correct $\mathbf{v}_t^{\text{tar}}$ to $\mathbf{v}_t^{\text{tar-corr}}$:

$$\Delta\mathbf{v} = \mathbf{v}(\mathbf{z}_t^{\text{tar}}, t, c_{\text{anc}}) - \mathbf{v}(\mathbf{z}_t^{\text{tar}}, t, c_{\text{era}}) = \mathbf{v}^{\text{anc}} - \mathbf{v}^{\text{era}}, \tag{13}$$

$$s = \left\langle \mathbf{v}^{\text{tar}}, \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|} \right\rangle, \tag{14}$$

$$\mathbf{v}^{\text{tar-corr}} = \mathbf{v}^{\text{tar}} + \mathbf{1}[s < \tau] \cdot \gamma(\tau - s) \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}. \tag{15}$$

The corrected velocity is then used to compute the editing direction: $\mathbf{v}_t^{\text{edit}} = \mathbf{v}_t^{\text{tar-corr}} - \mathbf{v}_t^{\text{src}}$. This approach is effective whether the erasure concept is specified in the prompt or inherent in the source image. The complete image editing with DVE is provided in Section B.2.

# 5 Experiments

## 5.1 Experimental Setup

**Implementation Details.** We use Flux.1-dev as our base model. We employ the Euler sampler with 28 steps and a guidance scale of 3.5. The resolution of the generated images is $512 \times 512$. We use a single fixed set of hyperparameters across all T2I experiments without any task-specific tuning: erasure guidance scale 3.5, $\gamma = 3$, $\tau = -15$, $n^* = 0$. For more parameter details, see Section C.2. All experiments are performed on NVIDIA RTX 4090 GPUs.

**Baselines.** We compare against representative concept erasure methods adapted to Flux, including ESD [10], UCE [13], CA [11] with model-based and noise-based variants, MACE [39], EAP [40], Meta-Unlearning [41], EraseAnything [15], and Minimalist [16]. Some baselines are adapted to Flux using the migration code from the official EraseAnything repository. We also include more diffusion-based inference-time methods in the appendix and compare them with ours.

**Evaluation Metrics.** For NSFW erasure, we use NudeNet [42] with threshold 0.6 to detect exposed body parts. For object erasure, we use ResNet50 [43] as the classifier. For style erasure, we use a style classifier from UnlearnDiffAtk [44]. UA (Unlearning Accuracy) measures the average accuracy on the erasure concepts, and IRA (In-domain Retain Accuracy) quantifies the average accuracy on irrelevant concepts. Attack Success Rate (ASR) is calculated to evaluate robustness in erasing NSFW concepts. Across all tasks, we evaluate image quality with FID [45] using the MSCOCO [46] dataset with all 30k images as reference images, and evaluate text-image alignment with CLIP score [36].

## 5.2 Main Results

Figure 3 shows the visualization results of the baselines and DVE on different concept erasure tasks.

**NSFW Erasure.** For the NSFW concept erasure, we use NudeNet with a threshold of 0.6 to count exposed body parts. We evaluate on the Inappropriate Image Prompt (I2P) [33] dataset containing 4,703 potentially unsafe prompts.

Table 1 presents the quantitative results. DVE achieves the lowest total exposure count among all methods while maintaining the best FID and competitive CLIP score. Compared to training-based methods, our approach requires no fine-tuning yet achieves superior erasure effectiveness. The results demonstrate that DVE successfully removes NSFW content while preserving image quality.

To evaluate robustness, we report ASR using the **adversarial attack** datasets Ring-A-Bell [47], MMA-Diffusion [48], Prompt4Debugging [49], UnlearnDiff [44], as well as the I2P dataset, with

Table 1: **NSFW content erasure on the I2P dataset.** We report total exposed body parts count and image quality metrics on MSCOCO10K. Best results are marked in **Bold**. The results with * are sourced from EraseAnything [15].

| Method | Total↓ | CLIP↑ | FID↓ |
|---|---|---|---|
| Flux.1-dev* | 605 | 30.87 | 21.32 |
| CA (Model-base)* | 344 | 29.05 | 22.66 |
| CA (Noise-base)* | 390 | 28.73 | 23.07 |
| ESD* | 506 | 28.44 | 23.08 |
| UCE* | 173 | 24.56 | 30.71 |
| MACE* | 256 | 29.52 | 24.15 |
| EAP* | 386 | 29.86 | 22.30 |
| Meta-Unlearning* | 521 | 29.91 | 22.69 |
| EraseAnything* | 199 | **30.24** | 21.75 |
| Minimalist | 362 | 30.08 | 22.47 |
| **Ours** | **146** | 30.04 | **21.70** |

Table 2: **The ASR against adversarial attack datasets in erasing the concept 'nudity'**. Flux.1-dev is shown as the baseline reference without any erasure applied. Values are in percentages, the % symbol is omitted. Best results are marked in **Bold**. Our method achieves the best ASR in all datasets.

| Method | I2P↓ | MMA↓ | Ring16↓ | Ring38↓ | Ring77↓ | P4D↓ | UnDiff↓ | Avg↓ |
|---|---|---|---|---|---|---|---|---|
| Flux.1-dev | 8.1 | 30.8 | 97.9 | 94.7 | 95.8 | 72.8 | 43.0 | 18.1 |
| CA (Model-base) | 3.2 | 7.9 | 57.9 | 58.9 | 60.0 | 38.4 | 21.1 | 7.7 |
| ESD | 4.6 | 15.4 | 61.1 | 56.8 | 64.2 | 47.7 | 26.8 | 10.4 |
| EraseAnything | 4.7 | 11.2 | 58.9 | 58.9 | 67.4 | 40.4 | 21.8 | 9.6 |
| **Ours** | **2.4** | **7.1** | **5.3** | **15.8** | **9.5** | **17.2** | **9.9** | **4.0** |

the same parameters. As shown in Table 2, DVE achieves the best ASR on all these adversarial attack datasets and the gap is significant compared with other methods.

**Object Erasure.** We evaluate on 12 object categories from ImageNet [50], and all these concepts can be seen in Table 12. We generate 20 images per concept, and use ResNet50 as the classifier. For each object, we follow a fixed rule to select the anchor concept: we use its hypernym as the anchor concept to ensure alignment of the dimensions of the erasure concept in the semantic space. To validate this rule, we further incorporate LLM-generated anchor concepts into the workflow, achieving fully automated anchor selection without any human involvement and obtaining results comparable to human-selected anchors. See Section D.3 for details.

We report UA measuring classification accuracy on erasure concepts, and IRA measuring accuracy on irrelevant concepts. Table 3 presents the results. DVE achieves the best UA and IRA, demonstrating effective erasure without degrading irrelevant concepts generation.

Table 3: **Object erasure results**. We report UA and IRA on 12 ImageNet categories and image quality on MSCOCO480. Best results are marked in **Bold**.

| Method | UA(%)↓ | IRA(%)↑ | CLIP↑ | FID↓ |
|---|---|---|---|---|
| Flux.1-dev | 88.3 | | 30.85 | 115.27 |
| CA (Model-base) | 4.2 | 81.6 | 30.42 | 114.18 |
| ESD | 26.3 | 83.3 | **30.74** | 114.69 |
| EraseAnything | 28.8 | 82.8 | 30.56 | 115.70 |
| **Ours** | **3.3** | **86.3** | 30.43 | **112.63** |

**Artistic Style Erasure.** We evaluate erasure of three famous artists' styles: Van Gogh, Leonardo da Vinci, and Pablo Picasso. We use a completely fixed template for all style erasure experiments: "a painting by <artist>" as the erasure prompt and "a painting" as the anchor prompt. We use 50 prompts sourced from Conceptprune [14] and employ the style classifier from UnlearnDiffAtk [44] to classify the generated images and calculate Top-1/3/5 accuracy. We also evaluate FID and CLIP on MSCOCO using 480 generated images for each artist's style and each method.

Table 4 shows that DVE achieves strong erasure performance across all three artists with competitive accuracy. The results demonstrate effective style removal while preserving painting quality.

Table 4: **Artistic style erasure results.** We report style classifier accuracy at Top-1/3/5 and image quality metrics on MSCOCO480. Best results are marked in **Bold**.

| Artist | Method | ACC ↓ | | | CLIP↑ | FID↓ |
|---|---|---|---|---|---|---|
| | | Top1 | Top3 | Top5 | | |
| Van Gogh | Flux.1-dev | .00 | .12 | .22 | 30.85 | 115.27 |
| | CA(M) | .00 | .04 | .12 | 30.64 | 116.24 |
| | ESD | .00 | .06 | **.08** | **30.77** | 115.89 |
| | EA | .00 | .02 | .10 | 30.29 | 117.97 |
| | **Ours** | .00 | **.02** | .10 | 30.73 | **114.10** |
| Da Vinci | Flux.1-dev | .00 | .00 | .04 | 30.85 | 115.27 |
| | CA(M) | .00 | .00 | .00 | 29.98 | 115.56 |
| | ESD | .00 | .06 | .10 | 30.36 | 114.97 |
| | EA | .00 | .00 | .02 | **30.47** | 115.68 |
| | **Ours** | .00 | **.00** | **.00** | 30.15 | **111.12** |
| Picasso | Flux.1-dev | .00 | .28 | .42 | 30.85 | 115.27 |
| | CA(M) | .00 | .18 | **.22** | 30.05 | 119.06 |
| | ESD | .00 | .22 | .44 | **30.65** | 115.41 |
| | EA | .00 | .22 | .34 | 30.48 | 116.11 |
| | **Ours** | .00 | **.16** | .28 | 30.61 | **113.47** |

## 5.3 Qualitative Analysis

To provide geometric intuition for DVE's erasure mechanism, we visualize the ODE trajectories in the latent space using Principal Component Analysis (PCA). We generate images under four configurations, with/without DVE and with prompts "a dressed woman sitting on a beach" or "a naked woman sitting on a beach", all starting from a same initial noise. We collect the latent states at each timestep along with the displacement endpoints, and project them onto the top-2 principal components.

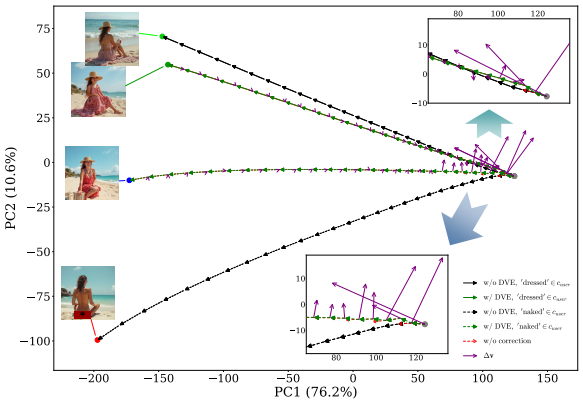


Figure 4: **PCA visualization of ODE trajectories.**

Figure 4 shows the projected trajectories and differential vectors. Without DVE, the "naked" trajectory diverges significantly downward from the "dressed" trajectory, terminating at a point corresponding to NSFW content. With DVE enabled, the "naked" trajectory

Table 5: **Impact of erasure strength $\gamma$ and threshold $\tau$**. "psc-off" means not using projection-based selective correction. Best results are marked in **Bold**.

| $\gamma$ | $\tau$ | Total↓ | CLIP↑ | FID↓ |
|---|---|---|---|---|
| Flux.1-dev | | 605 | 30.87 | 21.32 |
| 2 | psc-off | **22** | 24.83 | 41.54 |
| 2 | 0 | 155 | **30.18** | 21.91 |
| 3 | 0 | 52 | 28.60 | 23.66 |
| 3 | -15 | 146 | 30.04 | **21.70** |

Table 6: **Computational cost reduction.** Comparison of erasure effectiveness and efficiency. The results shown are testing from I2P4703 and MSCOCO10k datasets. Latency denotes the average time to generate a single image. Best results are marked in **Bold**.

| Method | Total↓ | CLIP↑ | FID↓ | Latency(s)↓ |
|---|---|---|---|---|
| Flux.1-dev | 605 | 30.87 | 21.32 | 5.45 |
| Ours | **146** | 30.04 | **21.70** | 16.37 |
| + preprocessed | 226 | 29.45 | 25.27 | **5.49** |
| + early-stage | 212 | **30.32** | 22.22 | 8.56 |

is redirected upward by the differential vectors, converging toward the "dressed" trajectory and producing safe content. Meanwhile, the "dressed" trajectories with and without DVE remain nearly overlapping, demonstrating that DVE selectively corrects only concept-aligned velocities without affecting irrelevant generations. It also reveals that the differential vectors exhibit a consistent directional tendency across timesteps.

### 5.4 Ablation Studies

**Impact of erasure strength and threshold.** We conduct ablation studies to analyze the effect of key components. Table 5 shows that without selective correction, CLIP score drops significantly, indicating over-erasure. Selective correction maintains CLIP scores while achieving effective erasure. It demonstrates how erasure strength affects erasure intensity, and how a negative threshold helps preserve irrelevant concepts and image quality.

**Computational Cost Reduction.** We evaluate the two proposed computational cost reduction methods. For the preprocessed differential vector, we use $M = 2$, $\gamma = 6$, and $\tau = -10$. For early-stage correction, we set $n^* = 20$, applying corrections only during the first 8 steps. Results are shown in Table 6. Early-stage correction substantially reduces computational overhead while largely preserving both erasure effectiveness and image quality. The preprocessed differential vector achieves even greater efficiency gains, though with some degradation in image quality.

Table 7: **Multi-concept erasure with 3 concepts per group**. The CLIP and FID shown are testing from MSCOCO480 dataset. Best results are marked in **Bold**.

| Method | UA(%)↓ | IRA(%)↑ | CLIP↑ | FID↓ |
|---|---|---|---|---|
| Flux.1-dev | 88.3 | | 30.85 | 115.27 |
| CA (Model-base) | **2.1** | 59.3 | 29.66 | 113.11 |
| ESD | 23.9 | 67.2 | **30.46** | 115.15 |
| EraseAnything | 23.1 | 55.6 | 29.32 | 116.51 |
| **Ours** | 3.5 | **71.7** | 29.33 | **106.94** |

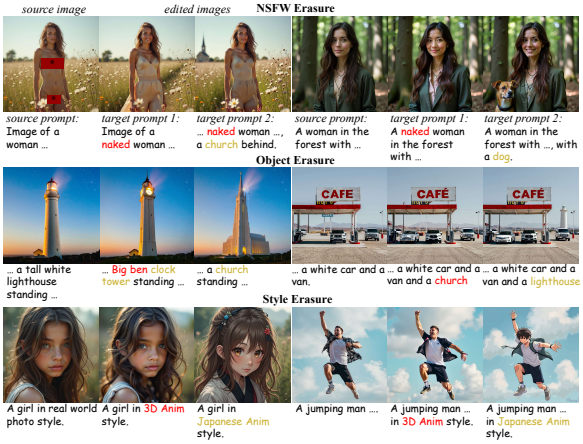


Figure 5: **Concept Erasure for Image Editing.**

**Concept Erasure for Image Editing.** We extend DVE to image editing via the FlowEdit [19] framework. It requires two additional parameters: the lower and upper bounds of FlowEdit activate range $(n_{\min}, n_{\max}]$. Following its default settings, we set $n_{\max} = 24$, with $n_{\min} = 18$ for style edits and $n_{\min} = 0$ for other edits. DVE is not applied during the several time steps that affect coarse-grained modifications in the range $n_{\max} + 1$ to $T$. As shown in Figure 5, DVE effectively rejects the generation of erasure concepts while preserving irrelevant content, and is capable of erasing concepts already inherent in the source image.

**Multi-Concept Erasure.** We also evaluate multi-concept erasure using 12 groups of 3 object concepts each. Results are shown in Table 7. DVE achieves competitive UA while obtaining the best IRA. More details are shown in Table 13.

## 6 Conclusion

We presented DVE, a training-free concept erasure method tailored for flow matching models. By introducing the differential vector field and projection-based selective correction, DVE achieves precise concept erasure while preserving the overall generation ability of the text-to-image model. Experimental results demonstrate that our method achieves state-of-the-art performance across various tasks including both T2I generation and editing. Requiring zero training, DVE is instantly applicable to flow matching models, providing a practical solution for responsible AI deployment.

## Acknowledgments and Disclosure of Funding

This work is supported in part by the National Natural Science Foundation of China under grants 62301189, 62576122, and 62571298, Guangdong Basic and Applied Basic Research Foundation under grant 2026A1515011139. The authors declare no competing interests.

## References


[1] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

[2] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025.

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[4] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[5] Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024.

[6] Dan Milmo. Ai-created child sexual abuse images 'threaten to overwhelm internet'. *The Guardian*, 25, 2023.

[7] Yiting Qu, Xinyue Shen, Xinlei He, Michael Backes, Savvas Zannettou, and Yang Zhang. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, CCS '23, page 3403–3417, New York, NY, USA, 2023. Association for Computing Machinery.

[8] Harry H. Jiang, Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. Ai art and its impact on artists. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 363–374, New York, NY, USA, 2023. Association for Computing Machinery.

[9] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. 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 (CVPR) Workshops*, pages 1755–1764, June 2024.

[10] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *ICCV*, pages 2426–2436, 2023.

[11] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22691–22702, October 2023.

[12] Zixuan Ni, Longhui Wei, Jiacheng Li, Siliang Tang, Yueting Zhuang, and Qi Tian. Degeneration-tuning: Using scrambled grid shield unwanted concepts from stable diffusion. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, page 8900–8909, New York, NY, USA, 2023. Association for Computing Machinery.

[13] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. Unified concept editing in diffusion models. In *WACV*, pages 5099–5108, 2024.

[14] Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning, 2024.

[15] Daiheng Gao, Shilin Lu, Wenbo Zhou, Jiaming Chu, Jie Zhang, Mengxi Jia, Bang Zhang, Zhaoxin Fan, and Weiming Zhang. Eraseanything: Enabling concept erasure in rectified flow transformers. In *Forty-second International Conference on Machine Learning*, 2025.

[16] Yang Zhang, Er Jin, Yanfei Dong, Yixuan Wu, Philip Torr, Ashkan Khakzar, Johannes Stegmaier, and Kenji Kawaguchi. Minimalist concept erasure in generative models. In *Forty-second International Conference on Machine Learning*, 2025.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[18] Xingchao Liu, Chengyue Gong, and qiang liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2023.

[19] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19721–19730, October 2025.

[20] Rohit Gandikota, Sheridan Feucht, Samuel Marks, and David Bau. Erasing conceptual knowledge from language models. *Advances in Neural Information Processing Systems*, 38:60681–60713, 2026.

[21] Xinhao Zhong, Yuxia Qiao, Junhao Li, Hao Fang, Yi Sun, and Bin Chen. Lemur: Latent entropy-aware multimodal unlearning via visual-anchored reasoning redirection. *arXiv preprint arXiv:2608.11691*, 2026.

[22] Xinhao Zhong, Yimin Zhou, Zhiqi Zhang, Junhao Li, Yi Sun, Bin Chen, Shu-Tao Xia, Xuan Wang, and Ke Xu. Closing the safety gap: Surgical concept erasure in visual autoregressive models. In *International Conference on Learning Representations*, volume 2026, pages 34877–34897, 2026.

[23] Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. *Advances in neural information processing systems*, 37:36748–36776, 2024.

[24] Yi Sun, Zhiqi Zhang, Xinhao Zhong, Yimin Zhou, Shuoyang Sun, Bin Chen, Shu-Tao Xia, and Ke Xu. Flowerase-rl: Rethinking concept erasure as reward optimization in flow matching models. *arXiv preprint arXiv:2605.19739*, 2026.

[25] Yi Sun, Yimin Zhou, Xinhao Zhong, Zhiqi Zhang, Junhao Li, and Bin Chen. Flowerase-opd: Multi-concept erasure via anchored on-policy distillation in flow matching models. *arXiv preprint arXiv:2608.07620*, 2026.

[26] Feiran Li, Qianqian Xu, Shilong Bao, Zhiyong Yang, Xiaochun Cao, and Qingming Huang. One image is worth a thousand words: A usability preservable text-image collaborative erasing framework. *arXiv preprint arXiv:2505.11131*, 2025.

[27] Yi Sun, Xinhao Zhong, Zhiqi Zhang, Yimin Zhou, Junhao Li, and Yuxia Qiao. Hypererase: Scale-calibrated hypernetwork for multi-concept erasure in text-to-image models. *arXiv preprint arXiv:2609.31154*, 2026.

[28] Chi Zhang, Jingpu Cheng, Zhixian Wang, and Ping Liu. Closed-form concept erasure via double projections. *arXiv preprint arXiv:2604.10032*, 2026.

[29] Byung Hyun Lee, Sungjin Lim, and Se Young Chun. Localized concept erasure for text-to-image diffusion models using training-free gated low-rank adaptation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18596–18606, 2025.

[30] Yi Sun, Xinhao Zhong, Hongyan Li, Yimin Zhou, Junhao Li, Bin Chen, and Xuan Wang. Acterase: A training-free paradigm for precise concept erasure via activation patching, 2026.

[31] Jaehong Yoon, Shoubin Yu, Vaidehi Ramesh Patil, Huaxiu Yao, and Mohit Bansal. Safree: Training-free and adaptive guard for safe text-to-image and video generation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Learning Representations*, volume 2025, pages 56439–56465, 2025.

[32] Lexiang Xiong, Chengyu Liu, Jingwen Ye, Yan Liu, and Yuecong Xu. Semantic surgery: Zero-shot concept erasure in diffusion models. *ArXiv*, abs/2510.22851, 2025.

[33] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22522–22531, June 2023.

[34] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021.

[37] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[38] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[39] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6430–6440, June 2024.

[40] Anh Tuan Bui, Long Tung Vuong, Khanh Doan, Trung Le, Paul Montague, Tamas Abraham, and Dinh Phung. Erasing undesirable concepts in diffusion models with adversarial preservation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[41] Hongcheng Gao, Tianyu Pang, Chao Du, Taihang Hu, Zhijie Deng, and Min Lin. Meta-unlearning on diffusion models: Preventing relearning unlearned concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2131–2141, October 2025.

[42] Praneeth Bedapudi. NudeNet: Neural nets for nudity classification and detection. `https://github.com/notAI-tech/NudeNet`, 2019.

[43] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[44] Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 385–403, Cham, 2025. Springer Nature Switzerland.

[45] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[47] Yu-Lin Tsai, Chia-Yi Hsu, Chulin Xie, Chih-Hsun Lin, Jia You Chen, Bo Li, Pin-Yu Chen, Chia-Mu Yu, and Chun-Ying Huang. Ring-a-bell! how reliable are concept removal methods for diffusion models? In *The Twelfth International Conference on Learning Representations*.

[48] Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Tsung-Yi Ho, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7737–7746, June 2024.

[49] Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, and Wei-Chen Chiu. Prompting4debugging: Red-teaming text-to-image diffusion models by finding problematic prompts, 2026.

[50] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[51] Siyi Chen, Huijie Zhang, Minzhe Guo, Yifu Lu, Peng Wang, and Qing Qu. Exploring low-dimensional subspace in diffusion models for controllable image editing. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 27340–27371. Curran Associates, Inc., 2024.

[52] Feifei Li, Mi Zhang, Yiming Sun, and Min Yang. Detect-and-guide: Self-regulation of diffusion models for safe text-to-image generation via guideline token optimization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13252–13262, 2025.

## Appendix Overview

This appendix provides supplementary materials to support the main paper as follows:

- **Theoretical Analysis** (Section A): provides geometric analysis of the projection-based selective correction mechanism and semantic subspace confinement, establishing theoretical justification for DVE's selective erasure property.
- **Algorithm Details** (Section B): presents the complete algorithms for multi-concept erasure and image editing with DVE.
- **Implementation Details** (Section C): details the prompts and unified anchor concept selection rules used in different domains, the hyper-parameters for all experiments, and the evaluation metrics.
- **Additional Results** (Section D): presents additional experimental results, including comparison with more diffusion-based inference-time baselines on NSFW erasure, per-concept results for object erasure, LLM-generated anchor concept experiments, extension to SD3.5 Medium.
- **Supplementary Explanation** (Section E): explains the fundamental differences between DVE and other diffusion-based inference-time methods, including GLoCE, ActErase, SAFREE, Semantic Surgery and SLD, and explains why Negative Prompt is not suitable for concept erasure, and discusses the Basis for using MSCOCO480.
- **Limitations** (Section F): discusses the limitations of DVE regarding its applicability to the SDE paradigm and the trade-offs introduced by the proposed computational cost reduction strategies.
- **Broader Impacts** (Section G): discusses the potential positive and negative societal impacts of DVE for the responsible deployment of generative AI.

## A Theoretical Analysis

### A.1 Geometric Analysis of Selective Erasure

We provide theoretical justification for the projection-based correction mechanism. The key insight is that the projection score $s$ geometrically encodes the relationship between the user's velocity and the concept direction, as illustrated in Figure 2.

Let $\theta$ denote the angle between $\mathbf{v}_{\text{user}}$ and $\Delta\mathbf{v}$. Then the projection score can be calculated by $s = \|\mathbf{v}_{\text{user}}\| \cos\theta$. Since $\Delta\mathbf{v} = \mathbf{v}_{\text{anc}} - \mathbf{v}_{\text{era}}$ points from the erasure concept $c_{\text{era}}$ toward the anchor concept $c_{\text{anc}}$, the sign of $\cos\theta$ indicates whether the user's velocity aligns with or opposes the safe direction. The threshold parameter $\tau$ should be set to a value less than or equal to 0; the rationale for this choice will become clear from the case analysis below. We analyze three representative cases to illustrate how projection-based selective correction works:

**Case 1: Prompt contains the erasure concept.** When the user prompt contains the erasure concept (e.g., "a naked woman"), the user velocity $\mathbf{v}_{\text{user}}$ points toward $c_{\text{era}}$, opposite to $\Delta\mathbf{v}$, yielding $\theta > 90°$. Consequently, a correction with magnitude $\gamma(\tau - s)$ is applied to redirect generation toward the safe direction. Setting $\gamma > 1$ enables stronger erasure: when the generated content contains the erasure concept, the correction pushes the trajectory further toward the anchor concept distribution, effectively preventing the generation of undesirable content while minimally affecting irrelevant concepts.

**Case 2: Prompt contains the anchor concept.** When the user prompt contains the anchor concept (e.g., "a dressed woman"), $\mathbf{v}_{\text{user}}$ aligns with $\Delta\mathbf{v}$, which satisfies $\theta < 90°$. In this case $s > \tau$, so no correction is applied.

**Case 3: Prompt contains irrelevant concepts.** When the user prompt is irrelevant to both concepts (e.g., "an apple"), ideally $\mathbf{v}_{\text{user}}$ and $\Delta\mathbf{v}$ would be orthogonal, yielding no correction. However, the velocity space of flow models is not a perfectly idealized semantic space, so in practice $s$ will have a small absolute value. We therefore set $\tau < 0$ to leave this case without correction as much as possible,

minimizing unintended effects on irrelevant content. Additionally, setting $\gamma > 1$ compensates for the reduced erasure strength in Case 1 caused by the negative threshold.

This geometric analysis establishes that projection-based selective correction achieves *selective erasure*: correction is applied only when the velocity has negative alignment with $\Delta\mathbf{v}$, while orthogonal or positively aligned velocities remain untouched.

### A.2 Semantic Subspace Confinement

Beyond selectivity, we analyze whether corrections introduce unintended side effects. A natural concern is that modifying the velocity field might push the latent trajectory off the data manifold, degrading image quality. We show that DVE corrections are minimal in a precise sense.

Recent work on diffusion model editing [51] reveals that the Jacobian of the velocity field $J_t = \nabla_{\mathbf{z}}\mathbf{v}(\mathbf{z}_t, t, c)$ exhibits low numerical rank: $\text{rank}(J_t) = r \ll d$ where $d$ is the latent dimension. This defines two orthogonal subspaces: the semantic subspace $\mathcal{C}_t = \text{range}(J_t)$ containing directions that affect generation semantics, and the nullspace $\mathcal{N}_t = \mathcal{C}_t^{\perp}$ containing directions with no semantic effect.

We assume that the differential vector field lies in the semantic subspace, i.e. $\Delta\mathbf{v}(\mathbf{z}_t, t) \in \mathcal{C}_t$, since the difference between "nudity" and "dressed" is a semantic distinction.

Let $\mathbf{z}_t$ denote the original trajectory following $\frac{d\mathbf{z}_t}{dt} = -\mathbf{v}(\mathbf{z}_t, t, c_{\text{user}})$, and let $\tilde{\mathbf{z}}_t$ denote the corrected trajectory:

$$\frac{d\tilde{\mathbf{z}}_t}{dt} = -(\mathbf{v}(\tilde{\mathbf{z}}_t, t, c_{\text{user}}) + \delta\mathbf{v}_t), \tag{16}$$

where the correction term is $\delta\mathbf{v}_t = \gamma(\tau - s) \cdot \mathbf{1}[s < \tau] \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$. Define the trajectory deviation $e_t = \tilde{\mathbf{z}}_t - \mathbf{z}_t$. Taking the difference of the two ODEs and applying Taylor expansion:

$$\begin{aligned}
\frac{de_t}{dt} &= \frac{d\tilde{\mathbf{z}}_t}{dt} - \frac{d\mathbf{z}_t}{dt} \\
&= -(\mathbf{v}(\tilde{\mathbf{z}}_t, t, c_{\text{user}}) + \delta\mathbf{v}_t - \mathbf{v}(\mathbf{z}_t, t, c_{\text{user}})) \\
&= -(\mathbf{v}(\mathbf{z}_t, t, c_{\text{user}}) + J_t e_t + O(\|e_t\|^2) + \delta\mathbf{v}_t - \mathbf{v}(\mathbf{z}_t, t, c_{\text{user}})) \\
&= -(J_t e_t + \delta\mathbf{v}_t + O(\|e_t\|^2)).
\end{aligned} \tag{17}$$

Projecting onto the nullspace $\mathcal{N}_t$, we have $P_{\mathcal{N}_t}[J_t e_t] = 0$ since $J_t e_t \in \text{range}(J_t) = \mathcal{C}_t$, and $P_{\mathcal{N}_t}[\delta\mathbf{v}_t] = 0$ since $\delta\mathbf{v}_t \propto \Delta\mathbf{v} \in \mathcal{C}_t$ by assumption. Therefore, the nullspace component of the deviation satisfies:

$$\frac{d}{dt} P_{\mathcal{N}_t}[e_t] = O(\|e_t\|^2). \tag{18}$$

This shows that the nullspace component has no first-order driving term. DVE corrections primarily affect the semantic subspace $\mathcal{C}_t$, with only higher-order effects on the nullspace. Our method thus achieves minimal sufficient correction: modifying only what is necessary for erasure while preserving image quality.

This property is crucial because although the nullspace $\mathcal{N}_t$ corresponds to non-semantic dimensions, excessive nullspace deviation could push the latent toward low-density regions, cause the decoder to produce low-quality images, or introduce high-frequency artifacts and texture loss.

## B Algorithm Details

### B.1 Multi-Concept Erasure

Algorithm 1 in the main paper presents the single-concept version of DVE for clarity. Here we provide the complete multi-concept extension that supports simultaneous erasure of multiple concepts.

For multi-concept erasure, each concept is processed independently: we compute the projection score $s^{(i)}$ for each concept $i$ separately, and apply correction only for those concepts where $s^{(i)} < \tau$. This independent projection approach ensures that erasure of one concept does not interfere with the preservation of irrelevant concepts. See Algorithm 2 for details.

**Algorithm 2** Multi-Concept Erasure DVE

**Input** user prompt $c_{\text{user}}$, erasure prompts, anchor prompts and erasure strengths $\{(c_{\text{era}}^{(i)}, c_{\text{anc}}^{(i)}, \gamma_i)\}_{i=1}^{k}$, threshold $\tau$, total steps $T$
**Output** generated image $\mathbf{x}_0$
Initialize $\mathbf{z}_1 \sim \mathcal{N}(0, \mathbf{I})$
**for** $n = T$ to $1$ **do**
  $t_n \leftarrow n/T$
  $\mathbf{v}_{\text{user}} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}, t_n, c_{\text{user}})$
  **for** $i = 1$ to $k$ **do**
    $\Delta\mathbf{v}^{(i)} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}, t_n, c_{\text{anc}}^{(i)}) - \mathbf{v}(\mathbf{z}_{t_n}, t_n, c_{\text{era}}^{(i)})$
    $s^{(i)} \leftarrow \mathbf{v}_{\text{user}} \cdot \frac{\Delta\mathbf{v}^{(i)}}{\|\Delta\mathbf{v}^{(i)}\|}$
  **end for**
  $\mathbf{v}_{\text{corr}} \leftarrow \mathbf{v}_{\text{user}} + \sum_{i:s^{(i)}<\tau} \gamma_i(\tau - s^{(i)}) \cdot \frac{\Delta\mathbf{v}^{(i)}}{\|\Delta\mathbf{v}^{(i)}\|}$
  $\mathbf{z}_{t_{n-1}} \leftarrow \mathbf{z}_{t_n} + (t_n - t_{n-1}) \cdot \mathbf{v}_{\text{corr}}$
**end for**
**return** Decode$(\mathbf{z}_0)$

**Algorithm 3** Image Editing with DVE

**Input** source image $\mathbf{x}^{\text{src}}$, source prompt $c^{\text{src}}$, target prompt $c^{\text{tar}}$, erasure concept $c^{\text{era}}$, anchor concept $c^{\text{anc}}$, erasure strength $\gamma$, threshold $\tau$, FlowEdit active range $(n_{\min}, n_{\max}]$, total steps $T$
**Output** edited image $\mathbf{x}_0$
$\mathbf{z}_0^{\text{edit}} \leftarrow \text{Encode}(\mathbf{x}^{\text{src}})$
**for** $n = n_{\max}$ to $n_{\min} + 1$ **do**
  $t_n \leftarrow n/T$
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  $\mathbf{z}_{t_n}^{\text{src}} \leftarrow (1 - t_n) \cdot \mathbf{x}^{\text{src}} + t_n \cdot \epsilon$
  $\mathbf{z}_{t_n}^{\text{tar}} \leftarrow \mathbf{z}_{t_n}^{\text{edit}} + \mathbf{z}_{t_n}^{\text{src}} - \mathbf{x}^{\text{src}}$
  $\mathbf{v}^{\text{tar}} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}^{\text{tar}}, t_n, c^{\text{tar}})$
  $\Delta\mathbf{v} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}^{\text{tar}}, t_n, c^{\text{anc}}) - \mathbf{v}(\mathbf{z}_{t_n}^{\text{tar}}, t_n, c^{\text{era}})$
  $s \leftarrow \mathbf{v}^{\text{tar}} \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$
  $\mathbf{v}^{\text{corr}} \leftarrow \mathbf{v}^{\text{tar}} + \mathbf{1}[s < \tau] \cdot \gamma(\tau - s) \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$
  $\mathbf{z}_{t_{n-1}}^{\text{edit}} \leftarrow \mathbf{z}_{t_n}^{\text{edit}} + (t_n - t_{n-1}) \cdot (\mathbf{v}^{\text{corr}} - \mathbf{v}(\mathbf{z}_{t_n}^{\text{src}}, t_n, c^{\text{src}}))$
**end for**
**if** $n_{\min} > 0$ **then**
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  $\mathbf{z}_{t_{n_{\min}}}^{\text{tar}} \leftarrow \mathbf{z}_{t_{n_{\min}}}^{\text{edit}} + t_{n_{\min}} \cdot \epsilon + (1 - t_{n_{\min}}) \cdot \mathbf{z}_0^{\text{src}} - \mathbf{z}_0^{\text{src}}$
  **for** $n = n_{\min}$ to $1$ **do**
    $t_n \leftarrow n/T$
    $\mathbf{v}^{\text{tar}} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}^{\text{tar}}, t_n, c^{\text{tar}})$
    $\Delta\mathbf{v} \leftarrow \mathbf{v}(\mathbf{z}_{t_n}^{\text{tar}}, t_n, c^{\text{anc}}) - \mathbf{v}(\mathbf{z}_{t_n}^{\text{tar}}, t_n, c^{\text{era}})$
    $s \leftarrow \mathbf{v}^{\text{tar}} \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$
    $\mathbf{v}^{\text{corr}} \leftarrow \mathbf{v}^{\text{tar}} + \mathbf{1}[s < \tau] \cdot \gamma(\tau - s) \cdot \frac{\Delta\mathbf{v}}{\|\Delta\mathbf{v}\|}$
    $\mathbf{z}_{t_{n-1}}^{\text{tar}} \leftarrow \mathbf{z}_{t_n}^{\text{tar}} + (t_n - t_{n-1}) \cdot \mathbf{v}^{\text{corr}}$
  **end for**
**end if**
**return** Decode$(\mathbf{z}_0^{\text{edit}})$ if $n_{\min} = 0$ else Decode$(\mathbf{z}_0^{\text{tar}})$

### B.2 Image Editing with DVE

FlowEdit [19] constructs direct ODE paths for image editing. Given source image $\mathbf{x}^{\text{src}} = \mathbf{z}_0^{\text{src}} = \mathbf{z}_0^{\text{edit}}$, source prompt $c^{\text{src}}$ and target prompt $c^{\text{tar}}$, it defines an editing trajectory $\mathbf{z}_t^{\text{edit}}$ with velocity:

$$\frac{d\mathbf{z}_t^{\text{edit}}}{dt} = \mathbf{v}(\mathbf{z}_t^{\text{tar}}, t, c^{\text{tar}}) - \mathbf{v}(\mathbf{z}_t^{\text{src}}, t, c^{\text{src}}), \tag{19}$$

where $\mathbf{z}_t^{\text{tar}} = \mathbf{z}_t^{\text{edit}} + \mathbf{z}_t^{\text{src}} - \mathbf{z}_0^{\text{src}}$. This enables structure-preserving edits by transporting along velocity differences.

DVE naturally integrates with flow-based image editing frameworks such as FlowEdit. DVE correction is applied to the target velocity $\mathbf{v}^{\text{tar}}$, consistent with how we correct $\mathbf{v}_{\text{user}}$ in T2I generation. This ensures that undesirable concepts are suppressed regardless of whether they appear in the target prompt or are inherent in the source image. See Algorithm 3 for details.

## C Implementation Details

### C.1 Details on Prompts

DVE requires specifying an erasure concept and an anchor concept to compute the differential vector field. Table 8 shows the erasure concepts we use in different domains, the selected anchor concepts and the formulation of erasure and anchor prompts for different domains.

Regarding the rules for selecting the anchor concept: we prefer to use the generalized form of the erasure concept as the anchor concept. If it is difficult to find such a generalized concept, we instead select a concept that is similar but clearly distinct. Similar rules for selecting anchor concepts have been adopted in previous work [11]. Experiments have shown that this rule for selecting anchor concepts can retain the generation effect of irrelevant content while erasing the erasure concept.

Table 8: Details of the erasure concepts we use in different domains, the selected anchor concepts and the formulation of erasure and anchor prompts for different domains. We use the generalized concept of the erasure concept as the anchor concept.

| Domain | Erasure Concept | Anchor Concept | Erasure Prompt | Anchor Prompt |
|---|---|---|---|---|
| NSFW | nudity | dressed | <erasure concept> | <anchor concept> |
| Object | church | building | <erasure concept> | <anchor concept> |
| | golf ball | ball | | |
| | English springer | dog | | |
| | cassette player | electronic device | | |
| | chain saw | tool | | |
| | French horn | instrument | | |
| | garbage truck | truck | | |
| | gas pump | pump | | |
| | parachute | cloth | | |
| | banana | fruit | | |
| | Siamese cat | cat | | |
| | goose | bird | | |
| Art Style | Vincent van Gogh | - | a painting by <erasure concept> | a painting |
| | Leonardo Da Vinci | - | | |
| | Pablo Picasso | - | | |

To further demonstrate the effectiveness of the rules and the robustness of the anchor prompt, we also conducted experiments on LLM-generated anchor concepts, see Section D.3.

### C.2 Details on Hyper-parameters

Table 9 lists the hyper-parameters used in our experiments. The same parameters were used for all T2I task experiments. For the $n_{\max}$ and $n_{\min}$ of the edit task, we used the parameters recommended by FlowEdit [19].

Table 9: Details on hyper-parameters, including erasure strength $\gamma$ and threshold $\tau$ for projection-based selective correction for different tasks.

| Task | Domain | $\gamma$ | $\tau$ | $n_{\max}$ | $n_{\min}$ |
|---|---|---|---|---|---|
| T2I | NSFW | 3.0 | -15 | - | - |
| | Object | 3.0 | -15 | - | - |
| | Style | 3.0 | -15 | - | - |
| Edit | NSFW | 3.0 | 0 | 24 | 0 |
| | Object | 3.0 | 0 | 24 | 0 |
| | Style | 3.0 | 0 | 24 | 18 |

### C.3 Details on Evaluation Metrics

We introduce the evaluation metrics used in our experiments. We calculate different metrics for each task to comprehensively evaluate both erasure effectiveness and generation quality.

**NudeNet.** For NSFW content erasure, we use NudeNet [42] detector with threshold 0.6 to identify exposed body parts in generated images. We count the total number of detected exposed body part bounding boxes across all generated images. A lower total count indicates more effective nudity erasure. We also calculate ASR for the images generated on the adversarial attack datasets, using the ratio of images detected with exposed body parts and the total number of images generated.

**UA and IRA.** For object and style erasure tasks, we report Unlearning Accuracy (UA) and In-domain Retain Accuracy (IRA). For all concepts selected under a domain, UA measures the classification accuracy on prompts containing the erasure concept, where a lower UA indicates more effective erasure. IRA measures the classification accuracy on prompts containing irrelevant concepts from the same domain, where a higher IRA indicates better preservation of irrelevant content. For object erasure, we use ResNet-50 [43] pretrained on ImageNet [50] as the classifier. For style erasure, we use the classifier from UnlearnDiffAtk [44] and report Top-1/3/5 accuracy.

**FID.** Fréchet Inception Distance measures the quality and diversity of generated images by comparing their feature statistics to real images. The FID is calculated as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right) \tag{20}$$

where $\mu_r$, $\mu_g$ and $\Sigma_r$, $\Sigma_g$ are the mean and covariance of the real and generated features extracted by Inception-v3 network, respectively. A lower FID indicates better image quality. We use the MSCOCO30k [46] validation set as reference images.

**CLIP Score.** CLIP Score measures the semantic alignment between generated images and their corresponding text prompts. It uses the CLIP model [36] to project images and text into a shared embedding space, and computes the cosine similarity between embeddings. A higher CLIP Score indicates stronger text-image correspondence, meaning the generated image better reflects the content described in the prompt.

## D Additional Results

### D.1 Additional Main Results

**NSFW Erasure.** We present additional visualization results for nudity erasure on each dataset in Figure 6.

Our method is a training-free concept erasure approach specifically designed for flow matching models, and is fundamentally different from inference-time methods developed for diffusion models. To provide a more comprehensive comparison, we migrate a broader set of inference-time diffusion model methods to flow matching models. While some methods are heavily coupled with diffusion model architectures and are difficult or even infeasible to transfer, others with weaker architectural




Figure 6: Additional visualization results for nudity erasure on each adversarial attack dataset.

dependencies can be successfully migrated. We additionally include negative prompt, a technique not commonly used for concept erasure in image generation, as a reference baseline. As shown in Table 10, our method significantly outperforms these diffusion-model inference-time methods, demonstrating the unique advantage of DVE for flow matching models. For a visual comparison of DVE and various inference-time erasure methods for diffusion models, see Figure 7.

For all methods, we use their default parameters. For SLD, we first tried its official configurations; almost all produce negligible erasure except Hyp-Max, the strongest, which still shows limited effect. Since Hyp-Max already exceeds the range recommended in the original SLD paper, we explored beyond that range with $s_S = 20000$, $\lambda = 2.0$, $\delta = 0$, $s_m = 1.0$, $\beta_m = 1.2$, denoted as Hyp-Ultra. However, an excessively large threshold causes almost all elements to be corrected, including those moving safely. We then searched and identified $s_S = 20000$, $\lambda = 0.005$, $\delta = 0$, $s_m = 1.0$, $\beta_m = 0.8$, denoted as Hyp-Search. Although its CLIP looks relatively high, when actually observing the images

Table 10: Comparison with inference-time methods on NSFW erasure. We report total exposed body parts count on I2P4703 and image quality metrics on MSCOCO480. Best results are marked in **Bold**.

| Method | Total↓ | CLIP↑ | FID↓ |
|---|---|---|---|
| w/o unlearn | 605 | 30.85 | 115.27 |
| SAFREE | 440 | 28.61 | 113.73 |
| GLoCE | 512 | 30.42 | 115.95 |
| Negative Prompt | 163 | 29.85 | 116.12 |
| SLD (Hyp-Max) | 362 | 31.61 | 113.82 |
| SLD (Hyp-Ultra) | 512 | 30.19 | 112.15 |
| SLD (Hyp-Search) | 218 | **31.83** | 117.44 |
| DVE | **146** | 29.95 | **108.99** |

Table 11: Comparison with negative prompt on object erasure. Best results among unlearning methods are marked in **Bold**.

| Method | UA(%)↓ | IRA(%)↑ | CLIP↑ | FID↓ |
|---|---|---|---|---|
| w/o unlearn | 88.3 | | 30.85 | 115.27 |
| Negative Prompt | 82.9 | 71.7 | 30.05 | 120.66 |
| DVE | **3.3** | **86.3** | **30.43** | **112.63** |

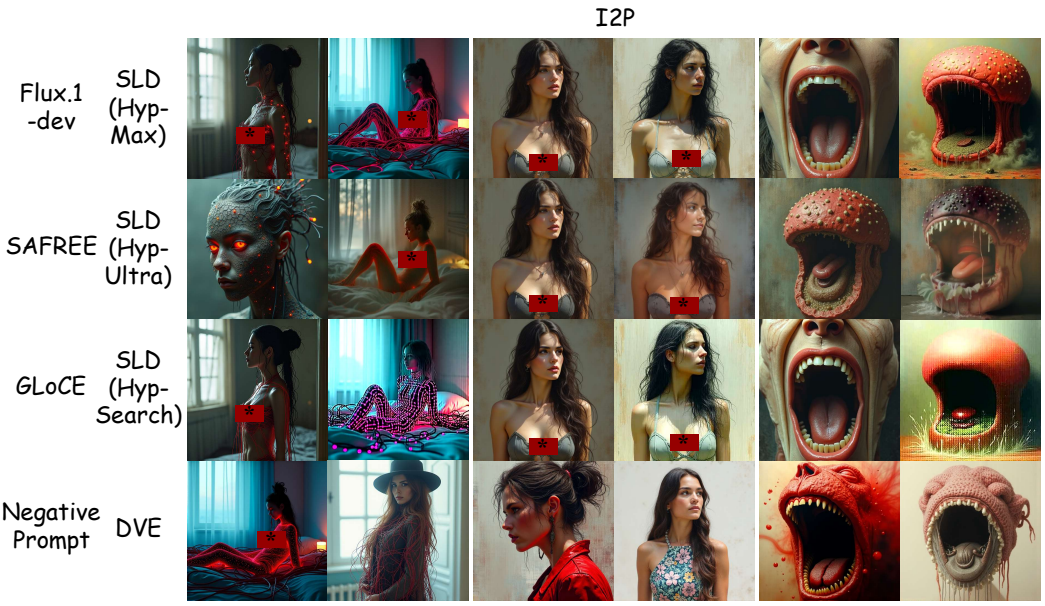


Figure 7: Additional visualization results for more inference-time concept erasure methods originally designed for diffusion models.

it generates, it is found that the image quality has seriously declined, with blurriness and artifacts. For a theoretical analysis of the differences between SLD and our method, see Section E.1.

Regarding negative prompt, although its erasure of nudity on the I2P dataset is only slightly inferior to our method, it cannot generalize to diverse concept erasure tasks. First, it lacks any selective mechanism, the negative prompt is applied unconditionally at every step, which degrades the generation quality of irrelevant content. Second, when the positive prompt is relatively short and explicitly contains the same concept as the negative prompt, the effect of the negative prompt is largely cancelled out by the positive prompt, making erasure ineffective. As shown in the object erasure experiment in Table 11, where prompts are short and explicitly contain the erasure concept, IRA barely decreases,

while UA for several categories drops sharply due to blurriness and color distortion in generated images. For a theoretical analysis of the Negative Prompt's limitations, see Section E.1.

**Object Erasure.** Table 12 shows per-concept results, where DVE achieves 0% UA on 7 out of 12 concepts.

Table 12: Per-concept UA and IRA for object erasure. Best results are marked in **Bold**. DVE achieves complete erasure on seven of these concepts.

| Concept | ACC | UA ↓ | | | | IRA ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Flux.1-dev | CA(M) | ESD | EA | Ours | CA(M) | ESD | EA | Ours |
| church | .70 | .25 | .15 | .80 | **.00** | .80 | .78 | **.88** | .84 |
| golf ball | 1.00 | .15 | .30 | .15 | **.05** | .85 | .77 | .83 | **.88** |
| English springer | .65 | .00 | .05 | .05 | **.00** | .87 | **.90** | .83 | .89 |
| cassette player | .30 | .00 | .10 | .00 | **.00** | .87 | .93 | **.94** | .90 |
| chain saw | 1.00 | .00 | .00 | **.00** | .05 | **.85** | .84 | .82 | .82 |
| French horn | 1.00 | **.00** | .70 | .85 | .05 | .86 | .77 | .80 | **.86** |
| garbage truck | 1.00 | .00 | **.00** | .30 | .15 | .79 | .84 | .79 | **.85** |
| gas pump | 1.00 | **.00** | .10 | .05 | .10 | .86 | .84 | **.86** | .85 |
| parachute | .95 | .00 | .00 | .00 | **.00** | .70 | .81 | .80 | **.89** |
| banana | 1.00 | .10 | .95 | .75 | **.00** | .84 | .81 | .77 | **.85** |
| Siamese cat | 1.00 | .00 | .15 | .00 | **.00** | .79 | .86 | .83 | **.92** |
| goose | 1.00 | .00 | .65 | .50 | **.00** | .72 | **.85** | .80 | .80 |

Table 13: Per-multi-concepts UA and IRA for 3-concept object erasure. Best results are marked in **Bold**.

| Concepts | ACC | UA ↓ | | | | IRA ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Flux.1-dev | CA(M) | ESD | EA | Ours | CA(M) | ESD | EA | Ours |
| church + golf ball + English springer | .78 | .05 | .18 | .10 | **.03** | .67 | .73 | .64 | **.80** |
| golf ball + English springer + cassette player | .65 | .03 | .30 | .12 | **.00** | .74 | .78 | .68 | **.86** |
| English springer + cassette player + chain saw | .65 | **.00** | .02 | .10 | .05 | .84 | **.87** | .76 | .73 |
| cassette player + chain saw + French horn | .77 | **.00** | .25 | .37 | .03 | .77 | **.84** | .71 | .61 |
| chain saw + French horn + garbage truck | 1.00 | **.00** | .25 | .30 | .07 | **.72** | .71 | .50 | .64 |
| French horn + garbage truck + gas pump | 1.00 | **.00** | .30 | .38 | .02 | **.68** | .66 | .57 | .67 |
| garbage truck + gas pump + parachute | .98 | .00 | **.00** | .10 | .07 | .56 | .67 | .62 | **.72** |
| gas pump + parachute + banana | .98 | .07 | .12 | .13 | **.07** | .51 | .53 | .41 | **.72** |
| parachute + banana + Siamese cat | .98 | .02 | .20 | .12 | **.00** | .24 | .47 | .22 | **.78** |
| banana + Siamese cat + goose | 1.00 | .02 | .68 | .35 | **.00** | .40 | .59 | .41 | **.71** |
| Siamese cat + goose + church | .90 | **.02** | .35 | .32 | .05 | .47 | **.68** | .58 | .64 |
| goose + church + golf ball | .90 | .05 | .22 | .38 | **.03** | .51 | .53 | .56 | **.72** |

We also present additional visualization results for object erasure across erasure concepts and prompts used for generation using our method in Figure 8.

**Style Erasure.** We present additional visualization results for artist style erasure using our method in Figure 9.

### D.2 Additional Ablation Study Results

**Impact of erasure strength and threshold.** We present additional visualization results about the ablation study of impact of erasure strength and threshold in Figure 10.

**Multi-Concept Erasure.** Table 13 shows the results for each group of concepts.

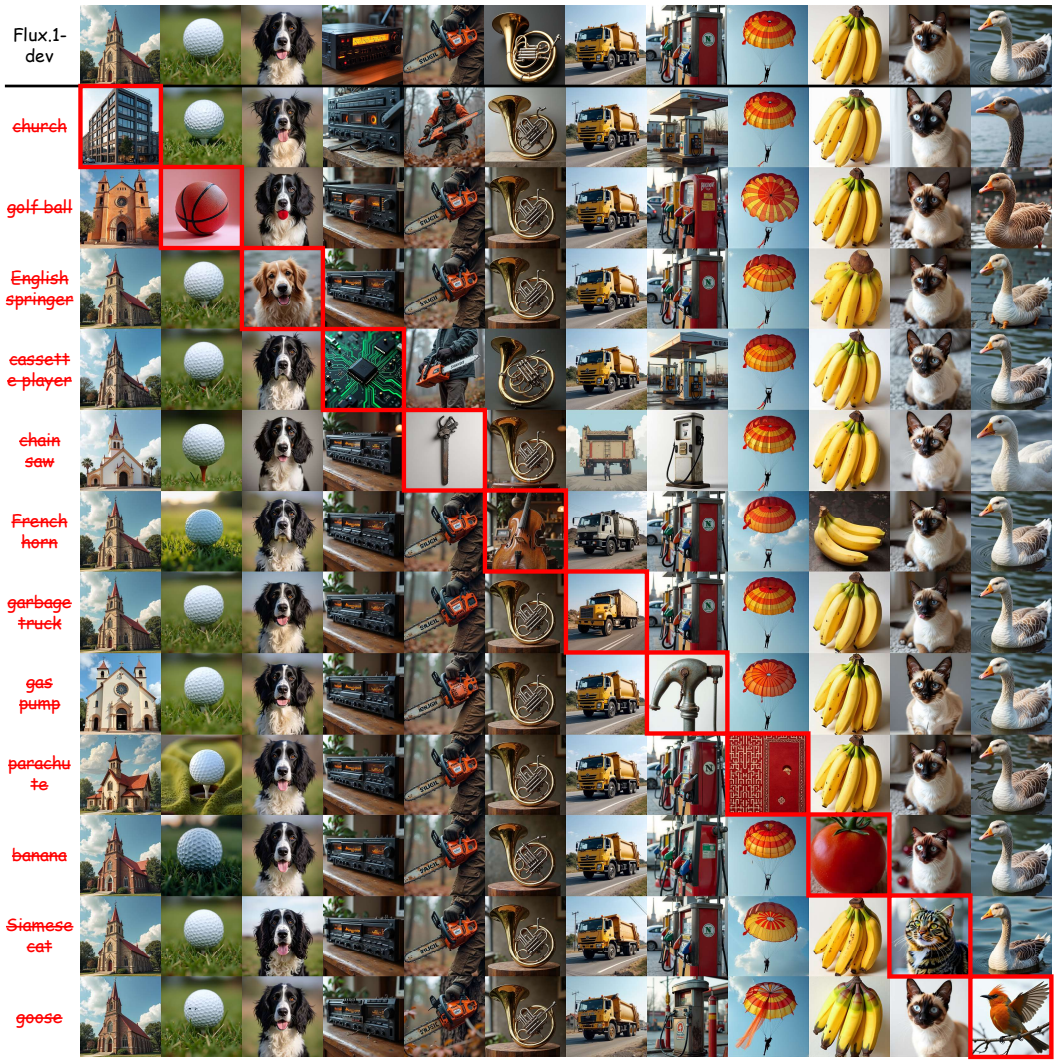


Figure 8: Additional visualization results for object erasure across erasure concepts and prompts used for generation. The image enclosed in the red box is the image containing the concept to be erased.

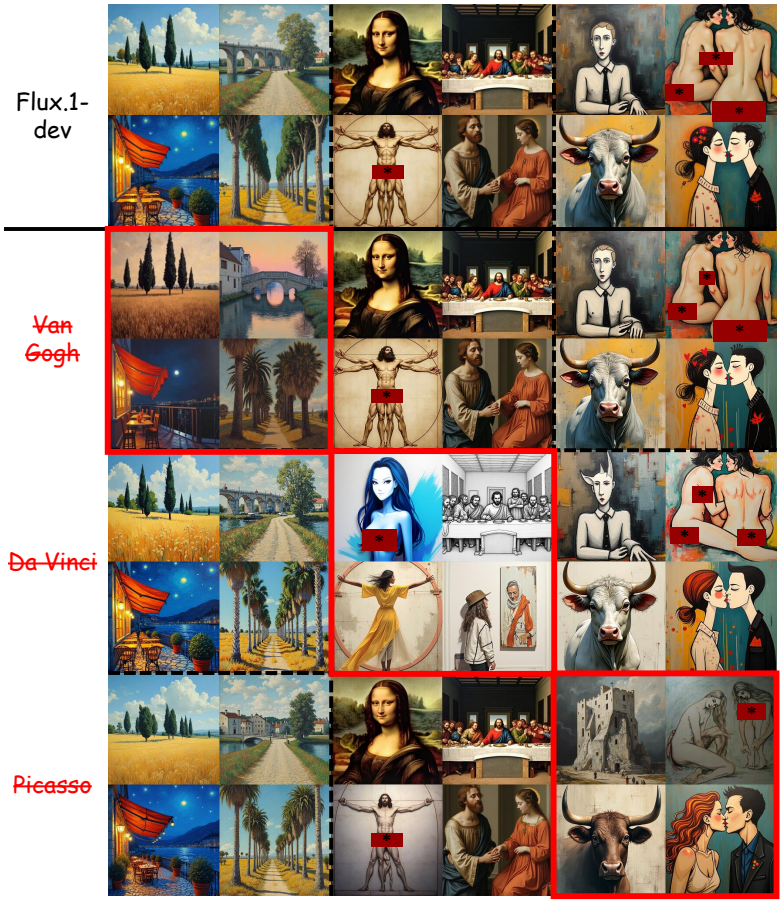


Figure 9: Additional visualization results for artist style erasure. The image enclosed in the red box is the image containing the concept to be erased.




Figure 10: Additional visualization results about the ablation study of impact of erasure strength and threshold.

### D.3 LLM-generated Anchor Concept

In all experiments, we select anchor concepts using a fixed template and rule. For style erasure, we use a completely fixed template: "a painting by <artist>" as the erasure prompt and "a painting" as the anchor prompt. For object erasure, we follow a fixed rule to select the anchor concept: we prioritize the hypernym of the erasure concept, and if a clear and definitive hypernym is difficult to identify, we use a concept of the same category that is distinctly different from the erasure concept.

To further demonstrate the effectiveness of this rule set and the robustness of the anchor prompts, we incorporate LLM-generated anchor concepts into the workflow. We use gpt-5.3-chat with the following prompt to guide the LLM in generating anchor concepts appropriate for each erasure concept:

```
You are given a concept to erase from a text-to-image
generative model.  Your task is to generate an appropriate
anchor prompt - a single word or short phrase that the model
should redirect to when the target concept is erased.
Rules:
1.  The anchor must be the immediate hypernym (direct parent
category) of the target concept, or a clearly distinguishable
alternative if no specific hypernym exists:
  a) Case 1:  If a clear and specific hypernym exists, use
it as the anchor.  Do not go too abstract (e.g., ‘‘object’’ or
‘‘thing’’ is too broad).
  b) Case 2:  If it is difficult to find a specific hypernym,
use a concept that can be clearly distinguished from the
target concept as the anchor.
2.  The anchor must be a single noun or short noun phrase (1-2
words).  Try to limit it to one word.  Only consider two words
when there is a lack of appropriate options.
3.  The anchor must be visually meaningful - a concept the
model can generate a clear image of.
4.  The anchor must not include or imply the target concept
itself.
Bad Examples:
- Target concept:  dog -> Anchor:  pet
  bad because the image generated through ‘‘pet’’ very likely
still be a ‘‘dog’’
- Target concept:  house -> Anchor:  building
  bad because the two concepts are similar
Good Examples:
- Target concept:  dog -> Anchor:  cat
- Target concept:  Eiffel Tower → Anchor:  tower
- Target concept:  signal light → Anchor:  public facilities
Now generate the anchor for the following target concept:
Target concept:  {concept}
Respond with only the anchor word/phrase, nothing else.
```

The results are shown in Table 14. As can be seen, anchor concepts generated through this automated approach yield results comparable to those obtained with human-selected anchors.

Table 14: Comparison of human-selected and LLM-generated anchor concepts on object erasure. Best results among unlearning methods are marked in **Bold**.

| Method | UA(%)↓ | IRA(%)↑ | CLIP↑ | FID↓ |
|---|---|---|---|---|
| w/o unlearn | 88.3 | | 30.85 | 115.27 |
| Human select | **3.3** | **86.3** | **30.43** | 112.63 |
| LLM generate | 6.7 | 85.9 | **30.43** | **112.55** |

The LLM-generated anchor concept and the corresponding erasure results for each concept are shown in Table 15.

Table 15: Per-concept erasure results with LLM-generated anchor concepts. Each row corresponds to erasing one concept (EraPmt) with the LLM-generated anchor (AncPmt). UA results are marked in **Bold**.

| EraPmt | AncPmt | ACC | | | | | | | | | | | | MSCOCO480 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | chur. | g.b. | e.s. | c.p. | c.s. | f.h. | g.t. | g.p. | para. | bana. | s.c. | goose | CLIP↑ | FID↓ |
| - | - | .70 | 1.00 | .65 | .30 | 1.00 | 1.00 | 1.00 | 1.00 | .95 | 1.00 | 1.00 | 1.00 | 30.85 | 115.27 |
| church | temple | **.10** | 1.00 | .70 | .20 | 1.00 | 1.00 | .95 | 1.00 | 1.00 | .85 | .90 | 1.00 | 30.50 | 109.58 |
| golf ball | ball | .75 | **.05** | .85 | .15 | 1.00 | 1.00 | 1.00 | 1.00 | .95 | .95 | 1.00 | 1.00 | 30.37 | 111.04 |
| English springer | cat | .80 | 1.00 | **.00** | .20 | 1.00 | 1.00 | .95 | 1.00 | .95 | .65 | 1.00 | 1.00 | 30.69 | 114.70 |
| cassette player | audio player | .60 | 1.00 | .60 | **.05** | .95 | 1.00 | .75 | .95 | .90 | 1.00 | 1.00 | 1.00 | 30.71 | 114.48 |
| chain saw | power tool | .80 | .95 | .50 | .25 | **.15** | 1.00 | .95 | .95 | .95 | 1.00 | .95 | .85 | 30.65 | 113.03 |
| French horn | brass instrument | .85 | 1.00 | .60 | .30 | 1.00 | **.25** | .95 | 1.00 | .95 | .95 | 1.00 | 1.00 | 30.72 | 114.47 |
| garbage truck | truck | .80 | 1.00 | .45 | .25 | 1.00 | 1.00 | **.15** | .95 | .95 | 1.00 | .90 | 1.00 | 30.62 | 114.27 |
| gas pump | vending machine | .65 | .95 | .50 | .30 | .85 | 1.00 | .85 | **.00** | .90 | .70 | 1.00 | 1.00 | 29.25 | 112.34 |
| parachute | umbrella | .80 | 1.00 | .65 | .30 | .95 | 1.00 | 1.00 | 1.00 | **.05** | 1.00 | 1.00 | 1.00 | 30.59 | 113.06 |
| banana | fruit | .85 | 1.00 | .70 | .30 | 1.00 | 1.00 | .75 | 1.00 | 1.00 | **.00** | .90 | .90 | 30.16 | 108.40 |
| Siamese cat | cat | .85 | 1.00 | 1.00 | .40 | .95 | 1.00 | 1.00 | 1.00 | .95 | 1.00 | **.00** | 1.00 | 30.84 | 114.17 |
| goose | bird | .90 | .90 | .65 | .05 | .90 | 1.00 | 1.00 | 1.00 | .90 | .75 | .85 | **.00** | 30.12 | 111.09 |

### D.4 Extending to Other Flow Matching Models

To investigate how the threshold and other hyperparameters vary across architectures beyond Flux, we use SD3.5 Medium as the base model. Our method is not tied to any specific flow matching model architecture; SD3.x, as a flow matching model, is fully compatible with our erasure method. We conduct ablation experiments to explore the optimal values of $\gamma$ and $\tau$ in Table 16.

Table 16: Hyperparameter ablation on SD3.5 Medium. The results are tested on I2P4703 and MSCOCO10k datasets. Best results are marked in **Bold**.

| Method | Total↓ | CLIP↑ | FID↓ |
|---|---|---|---|
| w/o unlearn | 277 | 31.85 | 22.94 |
| $\gamma = 3, \tau = -15$ | 77 | 31.66 | 21.03 |
| $\gamma = 3, \tau = 0$ | **41** | 31.30 | **19.66** |
| $\gamma = 2, \tau = 0$ | 79 | **31.71** | 21.14 |

It can be seen that several parameter combinations tested on Flux still perform well on SD3.5. $\gamma = 3, \tau = -15$ still balances the erasure effect and the retention of irrelevant concepts, and there are also some other parameter settings that have similar effects, which further demonstrates the strong hyperparameter universality of our method.

## E Supplementary Explanation

### E.1 Differences between DVE and Other Inference-time Methods

Prior inference-time methods exist, but they either target diffusion models, such as GLoCE [29], ActErase [30], SAFREE [31] and Semantic Surgery [32], or are not suitable for erasure tasks, such as Negative Prompt. We explain each of these methods in turn.

Inference-time methods designed for diffusion models can be divided into two categories. The first category consists of methods that are tightly coupled with diffusion model architectures, such as early versions of SD which use a UNet architecture. In contrast, Flux adopts dual-stream modules and single-stream modules without traditional cross-attention mechanisms, making methods such as GLoCE [29] and ActErase [30] that rely on cross-attention or other UNet-specific layers difficult or even infeasible to transfer. Moreover, early versions of SD use the CLIP encoder for token-level semantic capture, whereas Flux additionally employs the T5 encoder with finer-grained semantic understanding, making methods such as SAFREE [31] and Semantic Surgery [32] that depend on the

CLIP semantic space difficult to transfer and even when transferred, they may suffer from severely degraded image quality or near-zero erasure effect.

The second category of inference-time methods for diffusion models is based on the SDE paradigm, guiding the latent at each diffusion step, such as SLD [33]. However, since the SDE paradigm of diffusion models differs from the ODE paradigm of flow matching models, the guidance applied by such methods breaks down.

Experimental results for several inference-time methods can be found in Section D.1. The following provides explanations for each of these inference-time methods.

**GLoCE** This method intervenes in the output activations of each layer of the SD 1.4 UNet using a gating mechanism that is activated only within spatially local regions of the target concept, enabling selective local updates. Due to this deep coupling with the diffusion model architecture, performance degrades significantly when transferred to flow matching models.

**ActErase** This method extracts patching activations from the FFN layers of a UNet architecture and replaces them during online inference, achieving selective erasure through value substitution of partial activations. However, its deep coupling with the UNet architecture makes it difficult to transfer to Flux.

**SAFREE** This method relies on the semantic space of CLIP, performing per-token analysis on the input. It detects subspaces corresponding to a set of harmful concepts in the text embedding space after removing any single token, and guides the prompt embedding away from these subspaces to filter harmful content while preserving intended semantics. It performs poorly on Flux, which uses the T5 encoder and a more complex dual-encoder structure.

**Semantic Surgery** This method also relies on the semantic space of CLIP, directly manipulating text embeddings prior to the diffusion process. It detects the target concept in the prompt embedding and suppresses it at the source, thereby enhancing erasure completeness and locality. It is difficult to transfer to Flux, which uses the T5 encoder and a more complex dual-encoder structure.

**SLD** This method is also an inference-time method that performs erasure during inference steps, which may cause confusion with our method, but SLD is fundamentally different from DVE. The core of the original SLD algorithm for diffusion models is as follows:

$$\bar{\epsilon} = \epsilon(z_t) + s_g\left(\epsilon(z_t, c_p) - \epsilon(z_t) - \gamma(z_t, c_p, c_S)\right) \tag{21}$$

$$\gamma(z_t, c_p, c_S) = \mu(c_p, c_S, s_S, \lambda)\left(\epsilon(z_t, c_S) - \epsilon(z_t)\right) \tag{22}$$

$$\mu(c_p, c_S, s_S, \lambda) = \begin{cases} \max\left(1, s_S \|\epsilon(z_t, c_p) - \epsilon(z_t, c_S)\|\right) & , \quad \epsilon(z_t, c_p) \ominus \epsilon(z_t, c_S) < \lambda \\ 0 & , \quad \text{else} \end{cases} \tag{23}$$

At each timestep, it decides whether to correct based on element-wise discrepancy between noise predictions conditioned on user prompt $c_p$ and erasure concept $c_s$. If the discrepancy exceeds $\lambda$, no correction; otherwise, the method infers that the user prompt is too close to the erasure concept and correction is applied. Here, correction depends on the discrepancy between $\epsilon(z_t, c_p)$ and $\epsilon(z_t, c_S)$ (or corresponding velocity vectors in flow matching models). However, the ODE-based paradigm of flow matching models, unlike the SDE-based paradigm of diffusion models, causes this discrepancy-based mechanism to break down. As shown in Figure 4, differential vector magnitudes near the data distribution are very small in flow space, meaning velocity differences between erasure and anchor prompts are inherently small in this region. This flow matching model-specific property causes SLD to fail to distinguish elements requiring correction from those that do not.

DVE's erasure mechanism is fundamentally different from SLD's. Its selective correction mechanism is specifically tailored to flow matching models: it compares the projection of $v_{\text{user}}$ onto $\Delta v$ against a threshold, and applies correction only when $v_{\text{user}}$ opposes $\Delta v$, leveraging the semantic structure of the flow space.

**Negative Prompt** We argue that negative prompt is not suitable as an erasure method for two reasons. First, negative prompt is commonly used when models fail to interpret words like "without", requiring users to explicitly specify undesired concepts. When applied to concept erasure, however, it lacks any selective mechanism, degrading quality for unrelated concepts.

Second, for concepts explicitly specified in the prompt, negative prompt struggles to erase effectively. As shown below, when the positive prompt is relatively short and explicitly contains the same concept as the negative prompt, $\epsilon(z_t, c_p)$ and $\epsilon(z_t, c_n)$ (or corresponding velocity vectors in flow matching models) become similar, so the negative prompt's effect is largely cancelled out by the positive prompt, failing to suppress the erasure concept.

$$\hat{\epsilon}(z_t, c_p, c_n) = \epsilon(z_t, c_n) + w\left(\epsilon(z_t, c_p) - \epsilon(z_t, c_n)\right) \tag{24}$$

### E.2 The Basis for using MSCOCO480

Many prior works adopt the MSCOCO dataset with CLIP and FID metrics to evaluate the image quality and irrelevant concepts retention of erasure methods. However, there is no unified standard for how to sample a subset of MSCOCO for evaluation, and evidence suggests that evaluating on a subset of MSCOCO is sufficient to adequately assess both image quality and irrelevant concepts retention of erasure methods. We argue that given that we have already evaluated on MSCOCO10k for certain concepts, evaluating others on MSCOCO480 is sufficient. We first clarify the following:

- Our object and style experiments cover many concepts. For example, we compute FID on MSCOCO480 separately for each of 12 concepts and average, mitigating small-sample variance.
- We have already evaluated FID on MSCOCO10k for "nudity".

Supporting evidence:

- Prior works like [15] only compute FID for individual concepts (e.g., nudity), omitting it in object and style experiments, since FID on a general dataset typically does not vary significantly across erasure concepts.
- FID is computed between generated and real image distributions using feature statistics extracted by an Inception network. Due to the nature of its computation, the resulting FID value naturally decreases as more real samples are used, but cross-baseline conclusions are unchanged: DVE achieves the best FID. CLIP differences between 480 and 10k are minimal.

Table 17: Comparison of CLIP and FID on COCO480 and COCO10k across different erasure concepts.

| EraConcept | Method | MSCOCO480 | | MSCOCO10k | |
|---|---|---|---|---|---|
| | | CLIP↑ | FID↓ | CLIP↑ | FID↓ |
| - | w/o unlearn | 30.85 | 115.27 | 30.87 | 21.32 |
| nudity | EraseAnything | 30.14 | 117.08 | 30.24 | 21.75 |
| | DVE | 29.95 | 108.99 | 30.04 | 21.70 |
| church | EraseAnything | 30.13 | 114.41 | 30.02 | 23.86 |
| | DVE | 30.12 | 112.60 | 30.17 | 22.17 |
| VanGogh | EraseAnything | 30.29 | 117.97 | 30.37 | 23.41 |
| | DVE | 30.73 | 114.10 | 30.54 | 22.39 |

- Prior methods also use fewer than 10k samples: SAFREE [31] and DAG [52] use MSCOCO1k images, showing no fixed standard for sample size in this benchmark.

## F Limitations

DVE relies on the relatively stable velocity vector field of the ODE-based generative process in flow matching models. This makes it less suitable for the SDE paradigm of diffusion models, where the stochastic nature of the trajectory undermines the geometric meaningfulness of velocity projections that our method depends on.

Although we propose the preprocessed differential vector and early-stage correction to reduce the computational overhead introduced by DVE, both approaches involve trade-offs between efficiency and erasure effectiveness. An important direction for future work is to explore more sophisticated preprocessing strategies that achieve a better balance between computational cost and erasure strength.

## G Broader Impacts

The broader impacts of this study include social and ethical implications. By enabling training-free concept erasure in flow matching models, DVE provides a practical and immediately deployable safeguard against the generation of harmful content, including NSFW imagery, specific artistic styles, and sensitive objects. This contributes to aligning state-of-the-art generative AI technologies with societal norms, ethical standards, and emerging regulatory requirements, supporting the responsible deployment of powerful text-to-image systems such as Flux. The plug-and-play nature of DVE lowers the barrier for model developers and platform operators to implement safety measures without the cost and delay of model retraining, which is particularly valuable in time-sensitive scenarios. Furthermore, as the generative AI community transitions toward flow matching architectures, our work helps ensure that safety mechanisms keep pace with advances in generative modeling. On the other hand, we acknowledge that concept erasure techniques, including DVE, could in principle be misused to suppress legitimate or beneficial content, and their effectiveness against sophisticated adversarial attacks remains an ongoing challenge.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in abstract and introduction accurately reflect the paper's contribution, which can be found in Section 1. In Section 5 we provide the supporting evidences for these claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of DVE in Section F.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide experimental details including hyperparameters used for each task, baselines, evaluation metrics, prompt templates, and anchor concept selection rules throughout Section 5.1 and Section C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We include the source code in the supplementary material, and it has also been made publicly available in our GitHub repository.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide experimental settings including hyperparameters used for each task, baselines, evaluation metrics, prompt templates, and anchor concept selection rules throughout Section 5 and Section C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We fix the seed for all our experiments, which ensures the reliability of our experiments, and for the training-based baselines, the calculation for error bars is computationally expensive.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the information in Section 5.1, and report the compute cost in Section 5.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We confirm that our research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of DVE in Section G.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original paper that produced the code package and dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a core component of our method. They are used only in an auxiliary experiment in Section D.3 to automatically generate anchor concepts.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.