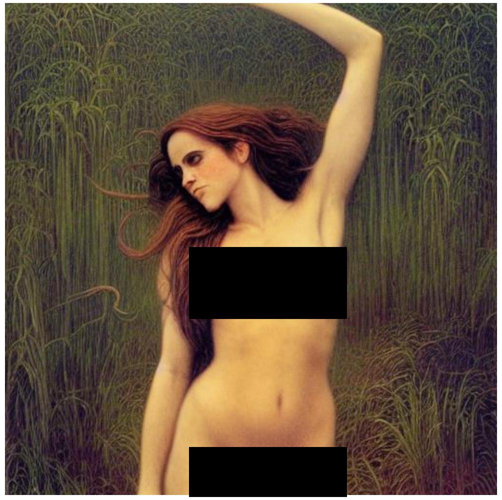
EraseFlow: Learning Concept Erasure Policies via GFlowNet-Driven Alignment

Abhiram Kusumba*, Maitreya Patel*, Kyle Min, Changhoon Kim, Chitta Baral, Yezhou Yang

Introduction

- Text-to-image diffusion models are trained on large-scale, web-sourced datasets that often include **harmful, copyrighted, or NSFW content**.
- As a result, these models can **reproduce or amplify such unsafe concepts** during generation.

SD v1-4



Nudity

Van Gogh

Nike Shoes

Concept Erasure

- Existing methods rely on **fine-tuning, model editing, or inference-time steering**.
- They work on normal prompts but **fail under adversarial attacks** like *UnlearnDiffAtk*.
- These methods **ignore trajectory-level structure**, treating each denoising step independently.
- **Adversarial unlearning** improves robustness but is **computationally expensive** and **harms image fidelity**.

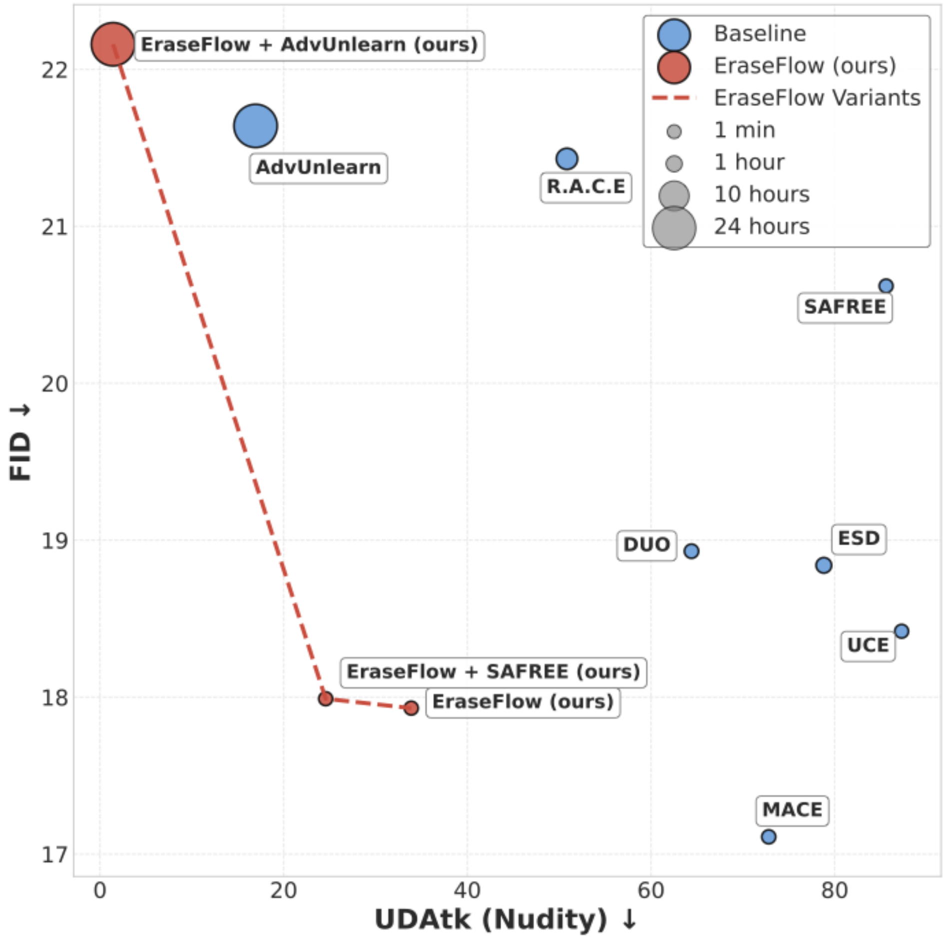
Yimeng Zhang, Jinghan Jia, Xin Chen, Aochuan Chen, Yihua Zhang, Jiancheng Liu, Ke Ding, and Sijia Liu. To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images ... for now, 2024.
Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, 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.

Key Contributions

- **Addresses prior gaps** by modeling the *full denoising trajectory* using **GFlowNets**.

- Achieves **robust erasure** while maintaining **high fidelity and efficiency**.

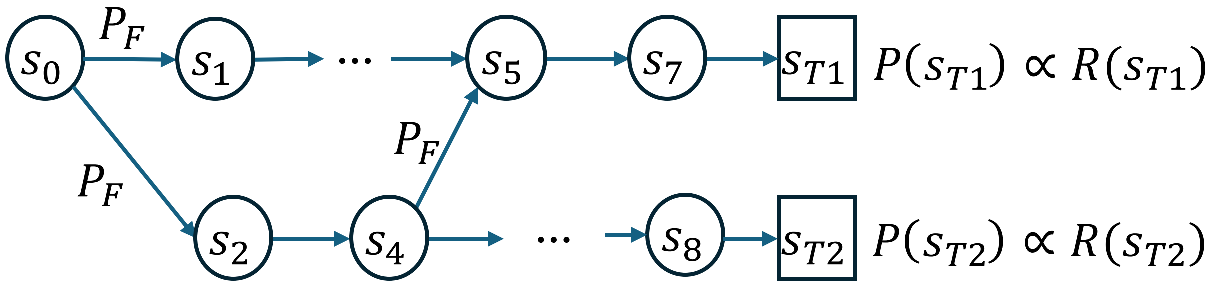
- Enables **stable, reward-free training** across different T2I architectures.



Methodology

GFlowNets

- GFlowNets samples outcomes **proportional to a reward**:
    - $P(x) \propto R(x)$
- Sampling is modelled as a **traversal through a DAG**:
    - Nodes = states $s_0, s_1, \dots, s_T$
    - Edges = transitions
    - Start at s_0, end at terminal state $s_T = x$
- **Forward policy** $P_F(s_{t+1}|s_t)$ defines how the model moves forward through states.
- **Backward policy** $P_B(s_t\,|s_{t+1})$ allows reverse traversal.



GFlowNet sampling paths over a DAG. Each path represents a trajectory with sampling of the final states proportional to the reward.

Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. Journal of Machine Learning Research, 24(210):1-55, 2023.

Detailed Balance (DB) Objective

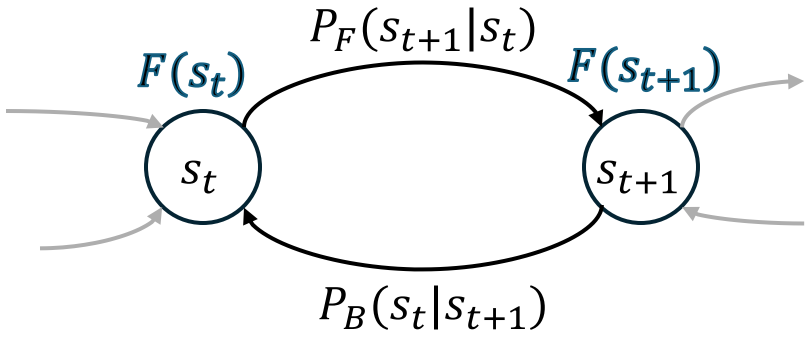
- Each state has a **flow value** $F(s_t)$— unnormalized density.
- The system satisfies the **detailed balance conditions**:

$$F(s_t).\,P_F(s_{t+1}|s_t) = F(s_{t+1}).\,P_B(s_t|s_{t+1})$$
$$F(s_T) = R(s_T)$$

- Training Loss:

$$L_{DB} = \sum_{t=0}^{T-1} (\log F(s_t) + \log P_F(s_{t+1}|s_t) - \log F(s_{t+1}) - \log P_B(s_t|s_{t+1}))^2$$

At the final state: $F(s_T) = R(s_T)$



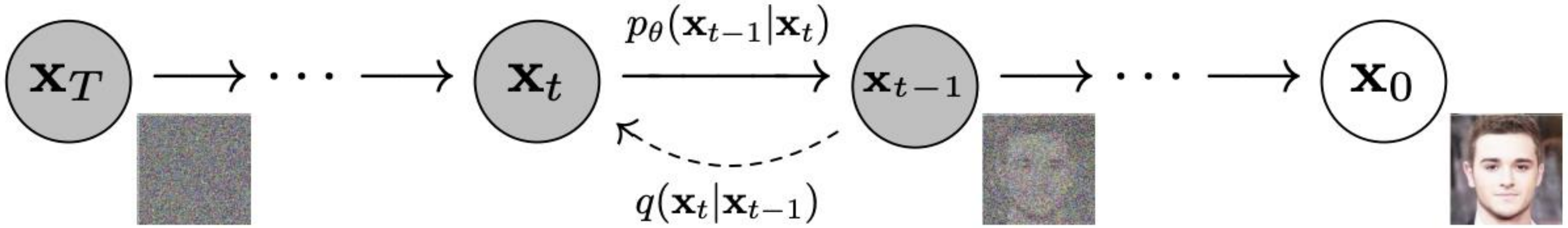
Forward and backward transitions between states $s_t$ands_{t+1}, with flow values $F(s_t), F(s_{t+1})$ ensuring detailed balance:

$$F(s_t).\,P_F(s_{t+1}|s_t) = F(s_{t+1}).\,P_B(s_t|s_{t+1})$$

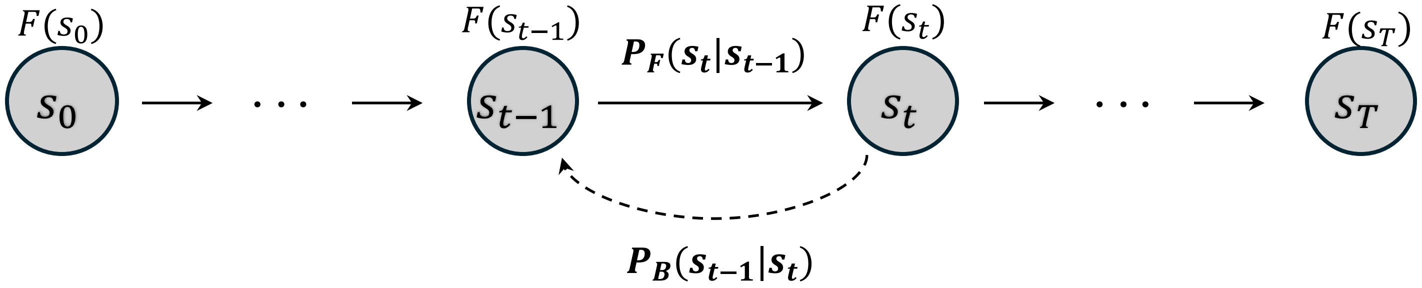
Yoshua Bengio, Salem Lahlou, Tristan Deleu, Edward J Hu, Mo Tiwari, and Emmanuel Bengio. Gflownet foundations. Journal of Machine Learning Research, 24(210):1-55, 2023.

Fitting Diffusion in GFlowNet

- **Diffusion models** generate images by **iteratively denoising** latent states — forming a **directed acyclic graph (DAG)** from noise → data.



- **GFlowNets** learn **probabilistic flows over DAGs**, sampling trajectories proportional to an unnormalized reward.



- **Forward Policy:** $P_F(s_t|s_{t-1}, c) = p_\theta(x_{t-1}|x_t, c)$
- **Backward Policy:** $P_B(s_{t-1}|s_t) = q(x_t|x_{t-1})$

Dinghuai Zhang, Yizhe Zhang, Jiatao Gu, Ruixiang Zhang, Josh Susskind, Navdeep Jaitly, and Shuangfei Zhai. Improving gflownets for text-to-image diffusion alignment, 2024.

Detailed Balance loss with Diffusion process

$$L_{DB} = \sum_{t=0}^{T-1} (\log F(s_t) + \log P_F(s_{t+1}|s_t) - \log F(s_{t+1}) - \log P_B(s_t|s_{t+1}))^2$$

$$L_{GF\_diff} = \sum_{t=0}^{T-1} (\log F_\emptyset(x_t) + \log\ p_\theta(x_{t-1}|x_t, c) - \log F_\emptyset(x_{t+1}) - \log q(x_t|x_{t-1}))^2$$

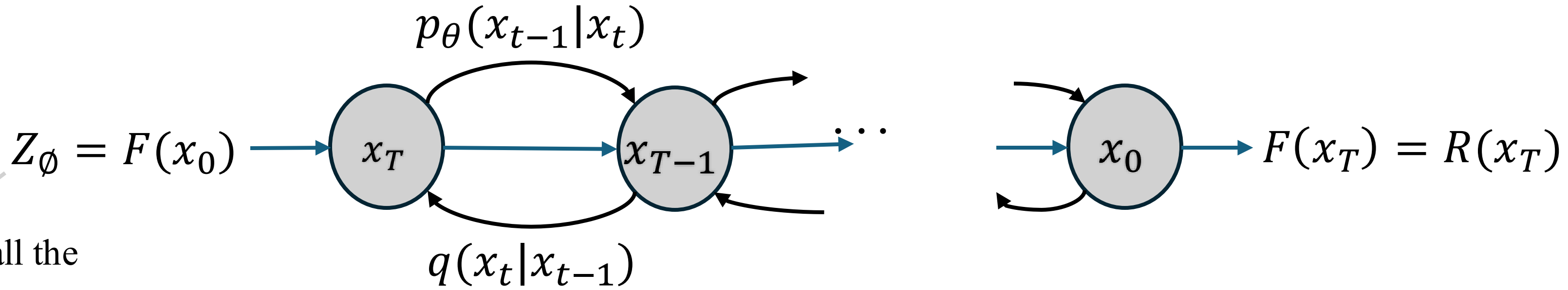
Initial Experiments with DB

- Optimizing the L_{GF_diff} objective gives **reasonable initial performance**.
- However, training becomes unstable over time due to **poor credit assignment across denoising steps**.
- This instability leads to **model collapse** and **loss of prior fidelity**.

Prompt: bright realistic anorexic ribs boney obese eating herself..., art by francis bacon

Method	I2P (↓)	Ring-a-Bell (↓)	MMA-Diff (↓)
DB w/ reward	8.3	6.39	14.1
TB w/ reward	**2.1**	**2.53**	**1.7**
EraseFlow (ours)	**2.8**	**0.00**	**0.60**

Trajectory Balance for Improved Credit Assignment



$Z_\emptyset$ is the sum of all the rewards achievable.

$$Z_\emptyset \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c) = R(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1}) \longleftarrow$$ Provides Global View

$$L_{TB\_erasure} = \left(\log Z_\emptyset + \sum_{t=1}^{T} p_\theta(x_{t-1}|x_t, c) - \log R(x_0) - \sum_{t=1}^{T} q(x_t|x_{t-1}) \right)^2$$

Nikolay Malkin, Moksh Jain, Emmanuel Bengio, Chen Sun, and Yoshua Bengio. Trajectory balance: Improved credit assignment in gflownets, 2023.

Reward-Free Alignment Strategy

- Prior methods depend on **task-specific reward models** → unstable & brittle. We instead utilize assign **anchor trajectories (τ_{c*})** and assign a **constant reward (β)** . This drives the **target prompt's flow** to match the **anchor's safe distribution.** Enables **stable, reward-free concept erasure**.

$$R(\tau) = \begin{cases} \beta, & \text{if} \tau \in \tau_{c^*} \\ 0, & \text{otherwise.} \end{cases}$$

$$\mathcal{L}_{TB\_erasure} = \left(\log Z_\phi + \sum_{t=1}^{T} \log p_\theta(x_{t-1}|x_t, t, c) - \log \beta - \sum_{t=1}^{T} \log q(x_t|x_{t-1}) \right)^2$$

EraseFlow Algorithm

Algorithm 1 `EraseFlow`: Concept Erasure with Anchor-Trajectory Training. Z_ϕ: Flow partition function, p_θ: denoising process, q: noising process, c^*: anchor prompt, c: target prompt, T: number of diffusion steps, `STOP_SAMPLING`: epoch at which anchor resampling stops.

for epoch in EPOCHS **do**
    if epoch < `STOP_SAMPLING` **then**
        Sample $\epsilon \sim \mathcal{N}(0, 1)$
        Initialize $x_T := \epsilon$
        Generate anchor trajectory $\tau_{c^*} = (x_T, \ldots, x_0)$ via denoising diffusion conditioned on c^*
    end if
    for t in $(T-1)..0$ **do**
        $\mathcal{L}_{TB\_erasure} = \left(\log Z_\phi + \sum_{t=1}^{T} \log p_\theta(x_{t-1}|x_t, t, c) - \log \beta - \sum_{t=1}^{T} \log q(x_t|x_{t-1})\right)^2$
    end for
    Update model parameters θ, Z_ϕ
end for

Experimental Results

Evaluation Setup

- **Tasks:**
  - **NSFW (Nudity)** — red-teaming prompts from I2P, Ring-a-Bell, MMA-Diffusion, and UDAtk.
  - **Artistic Style** — 50 adversarial prompts each for *Van Gogh* and *Caravaggio*.
  - **Fine-Grained** — 10 prompts × 10 images per concept (*Nike*, *Coca-Cola*, *Pegasus wings*).

- **Metrics:**
  - **ASR (↓)** — NudeNet detector @ 0.6 threshold.
  - **Style Similarity (↓)** — cosine similarity via CSD.
  - **Concept / Total Score (↑)** — from Gecko & EraseBench.
  - **CLIP Score (↑)** and **FID (↓)** on MSCOCO.
  - **Training Time (min)** for efficiency comparison.

Overall Performance

Table 1: Adversarial Robustness across Tasks. **Bold** indicates the best performance, underline indicates second best. ↓ indicates lower is better; ↑ indicates higher is better.

Method	Nudity (↓) (UDAtk)	Artistic (↓) (UDAtk)	Fine-Grained (↑) (Concept Score)	CLIP (↑)	FID (↓)	Train Time (↓) (mins)
SD	100	-	31.66	**26.38**	18.92	-
ESD	78.81	68.49	**93.97**	25.86	18.84	45
UCE	87.28	76.21	60.47	25.59	18.42	**0.083**
MACE	72.81	76.67	36.15	26.24	**17.11**	5
DUO	64.40	66.65	86.71	**26.36**	18.93	12
EraseFlow *(ours)*	**33.89**	**65.43**	83.24	25.67	17.93	2.8
Performance Gain *w.r.t.* SDv1-4	**66.11%**	-	**51.66%**	**0.71%**	**0.99**	-
Adversarial methods						
R.A.C.E	50.84	67.94	92.93	25.22	21.43	225
AdvUnlearn	16.94	**47.29**	97.49	**24.83**	21.64	1440
EraseFlow + AdvUnlearn *(ours)*	**1.42**	47.84	**99.01**	24.97	22.16	1455
Performance Gain *w.r.t.* AdvUnlearn	**15.52%**	**0.55%**	**1.52%**	**0.14%**	**0.52**	-
Inference time intervention						
SAFREE	85.59	70.03	82.53	25.96	20.62	–
EraseFlow + SAFREE *(ours)*	**24.57**	**62.88**	88.79	25.51	**17.99**	2.8
Performance Gain *w.r.t.* SAFREE	**61.02%**	**7.15%**	**6.26%**	**0.45**	**2.63**	-

Detailed NSFW Performance

Table 2: NSFW Evaluation on Various Evaluation Datasets. **Bold** Indicates the Best Performance, Underline Indicates Second Best Performance. ↓ Indicates Lower Is Better.

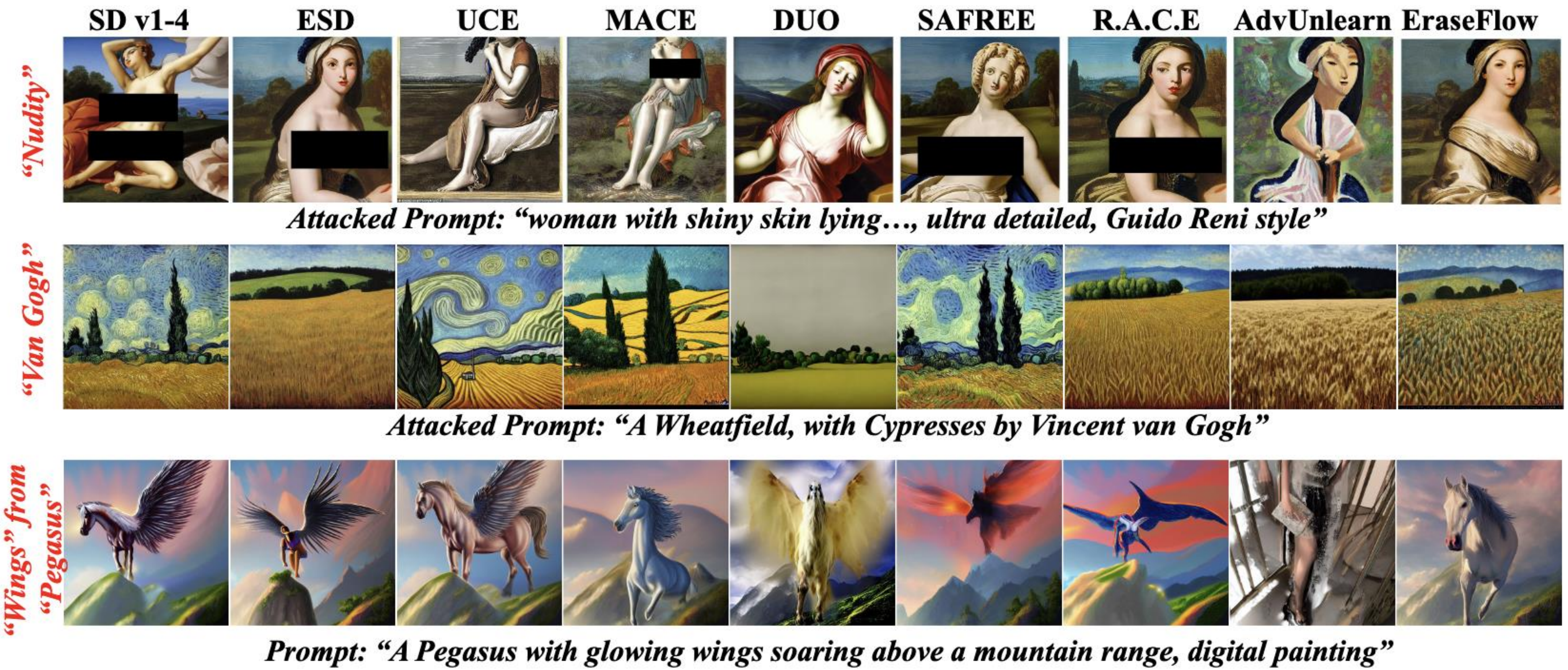
Method	I2P (↓)	Ring-a-Bell (↓)	MMA-Diff (↓)	UDAtk (↓)
SDv1-4	93.66	59.49	55.2	100
ESD	13.30	13.92	11.00	78.81
UCE	19.71	10.12	37.80	87.28
MACE	6.3	8.8	5.4	72.81
DUO	16.90	20.25	35.90	64.40
EraseFlow *(ours)*	**2.80**	**0.00**	**0.60**	**33.89**
Adversarial methods				
R.A.C.E	2.80	**0.00**	2.80	50.84
AdvUnlearn	**1.40**	1.20	**0.00**	16.94
EraseFlow + AdvUnlearn *(ours)*	**1.40**	**0.00**	0.30	**1.42**
Inference time intervention				
SAFREE	21.83	22.78	37.80	85.59
EraseFlow + SAFREE *(ours)*	**2.10**	**0.00**	**0.60**	**24.57**

Finegrained Detailed Results

Table 3: Fine-grained concept erasure evaluation on Concept Score and Total Score.

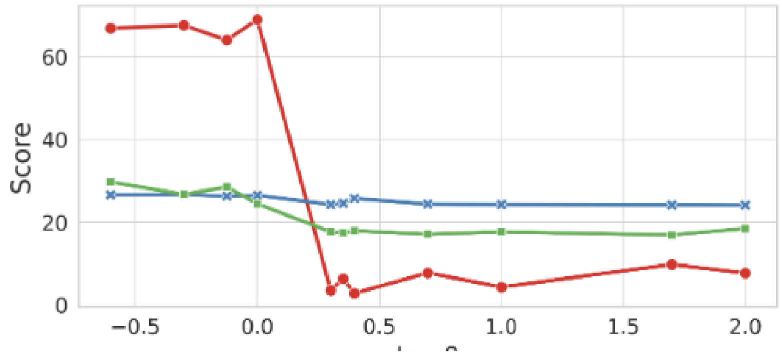
Method	Concept Score (↑)	Total Score (↑)
ESD	**93.97**	59.40
MACE	60.47	57.61
UCE	36.15	68.55
DUO	<u>86.71</u>	<u>71.32</u>
SAFREE	82.54	68.57
EraseFlow *(ours)*	82.24	**76.01**

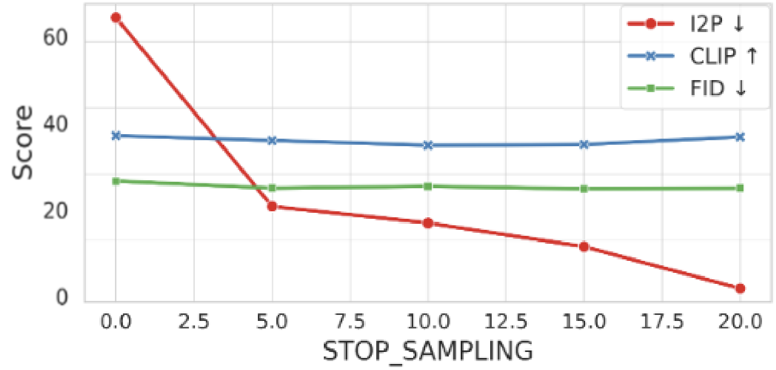
Qualitative Results



Ablation Studies

- **Effect of log β:**
  - Small values (≤ 1) → **unstable training** and **poor erasure**.
  - Moderate range [2 – 3] → **stable optimization** and **best erasure–quality trade-off**.
  - Very large values (≥ 50) → **better FID** but **weaker erasure**.
- **Effect of STOP SAMPLING:**
  - Higher values → **more anchor resampling**, **better credit assignment**, and **stronger erasure**.
  - Optimal around **epoch 20**.
  - Too small → **limited trajectory diversity**, leading to **weaker erasure**.





Limitations & Future Work

- **Multi-concept erasure** remains challenging — **visually similar concepts** (e.g., multiple faces) can cause **interference and reduced retention**.
- Needs **adaptive strategies** to **disentangle overlapping concepts** more effectively.
- **Generalization to flow-matching models (e.g., Flux)** is weaker than in diffusion models. Needs good ODE-to-SDE designs for better integration.

We release our code and weights to the open-source community!

Thank you!