FLOWERASE-RL: RETHINKING CONCEPT ERASURE AS REWARD OPTIMIZATION IN FLOW MATCHING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in flow matching models have significantly improved text-to-image generation quality, but also introduce growing safety risks due to the generation of harmful or undesirable content. Existing concept erasure methods are either inference-time interventions with limited effectiveness or rely on supervised fine-tuning (SFT), which requires precisely aligned data and struggles with scalability and multi-concept settings. In this paper, we propose *FlowErase-RL*, the first GRPO-based framework for concept erasure in flow matching models. We reformulate concept erasure as a reward optimization problem and introduce a **dynamic dual-path reward mechanism** that jointly optimizes (i) a Concept Erasure (CE) reward to suppress target concepts and (ii) a Non-target Space (NS) reward to preserve generative fidelity. The two reward paths are adaptively balanced during training via a performance-driven switching strategy, enabling stable optimization without explicit supervision. Extensive experiments on nudity, object, and artistic style erasure demonstrate that our method achieves state-of-the-art erasure performance while maintaining strong image quality and semantic alignment. Moreover, it exhibits robust resistance to adversarial attacks and scales effectively to multi-concept scenarios. Our results establish a new paradigm for safe and controllable generation in flow matching models.

1 INTRODUCTION

In recent years, attributable to the uncurated nature of Internet-sourced training data Milmo (2023), the proliferation of inappropriate content generated by text-to-image (T2I) models Dhariwal & Nichol (2021); Ho & Salimans (2022); Ho et al. (2020); Nichol et al. (2021); Saharia et al. (2022) has become a pressing concern. While the capability to synthesize high-quality images from textual prompts has advanced significantly, from the advent of DALL-E 2 Ramesh et al. (2022) and Stable Diffusion (SD) Rombach et al. (2022) to the beefed-up Flux Labs (2024), the risk of producing harmful outputs Jiang et al. (2023); Roose (2022); Setty (2023) necessitates effective countermeasures. Complete data removal followed by model retraining Nichol et al. (2021); Schramowski et al. (2023) represents a direct approach, yet its exorbitant computational overhead, inefficiency, and susceptibility to performance degradation O'Connor (2022) render it impractical. In contrast, concept erasure (CE) offers a compelling alternative, enabling the suppression of specific target concepts through lightweight interventions while safeguarding the model's generative proficiency.

Existing CE methods have been developed primarily for the SD paradigm, which relies on the DDPM Ho et al. (2020) and DDIM Song et al. (2020) sampling procedures coupled with the U-Net backbone. In contrast, the Flux series departs substantially from this design. Flux adopts flow matching Lipman et al. (2022) as its core generative mechanism and employs a transformer-based architecture instead of the U-Net structure. Furthermore, Flux integrates an additional text encoder, Google T5, and utilizes Rotary Position Embedding (RoPE) for both pixel and textual representations. These architectural distinctions render prior CE techniques ineffective when applied to the Flux framework. Existing approaches, whether training-free methods Wang et al. (2025); Chavhan et al. (2024); Gandikota et al. (2024) that merely intervene at inference time or training-based methods Chen et al. (2025); Lu et al. (2024); Cywi'nski & Deja (2025); Kim et al. (2024); Li et al. (2024); Zhong et al. (2025) tailored to the SD pipeline, fail to generalize to this modern architecture. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

inability to transfer concept erasure techniques from SD to Flux thus constitutes a critical gap that this work seeks to bridge.

To address this issue, we propose the first concept erasure method based on GRPO Shao et al. (2024); Liu et al. (2025) called *FlowErase-RL*. By formulating target concepts as a CE reward function, FlowErase-RL achieves effective erasure performance. To further enhance erasure performance while preserving generation capability, we design a dynamic dual-path reward mechanism. The verifiable CE Reward guides the model to move away from the target concept during image synthesis, while the Non-target Space (NS) Reward encourages the model to generate high-quality images for prompts unrelated to the target concept. During training, the CE reward employs prompt pairs with or without target concepts and the NS reward takes a retain set as input. The two reward paths are automatically switched per epoch with their strength and weighting dynamically adjusted based on the observed erasure performance. Extensive experiments on multiple target concepts demonstrate that this dual-objective formulation simultaneously coordinates safety and utility within a unified reward structure, allowing FlowErase-RL to achieve superior erasure effectiveness while maintaining high generative quality. It outperforms existing state-of-the-art methods on flow matching models and also shows excellent erasure capability on other T2I model architectures. In summary, we make the following contributions:

- We propose the first concept erasure method based on GRPO for flow matching models, achieving effective erasure while maintaining excellent generative capabilities.
- We introduce a dynamic dual-path reward mechanism that adaptively switches reward functions during training, striking an effective balance between efficient concept erasure and the preservation of generative capabilities.
- Extensive experimental results based on flow matching models demonstrate that our method achieves or approaches SOTA performance on major erasure tasks.

2 Related Works

2.1 Flow Matching

Flow matching Lipman et al. (2022); Labs (2024) has become a powerful alternative to DDPM Ho et al. (2020) for generative modeling. While diffusion models 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 Liu et al. (2022) proposes learning straight-line trajectories between noise and data pairs, enabling efficient single-step generation. Flux Labs et al. (2025); Labs (2024) and Stable Diffusion 3 Esser et al. (2024) adopt flow matching with transformer architectures, achieving high image quality. Subsequent works have explored downstream tasks as well. Despite being efficient and high-quality, current flow matching generative models could still produce sensitive content and lack safety defense methods.

2.2 Concept Erasure Methods in T2I Models

Existing CE methods Rando et al. (2022); OpenAI (2023); Nichol et al. (2021) can be categorized into two groups: training-free and training-based methods. Training-based methods typically involve fine-tuning pre-trained models to remove specific concepts while training-free methods often relay on causal intervention and masks to modify model behavior without fine-tuning during inference.

training-free methods Unified Concept Editing (UCE) Gandikota et al. (2024) introduces closed-form solutions to edit the cross attention weights, changing the key and value matrices of specific text embeddings containing target concept, while retaining the matrix weights of unrelated concepts. Adaptive Value Decomposer (AdaVD) Wang et al. (2025) disentangles target semantics from the cross-attention layers at each denoising timestep, dynamically controls the erasure intensity through an adaptive token-wise shift mechanism. ActErase Sun et al. (2026) try to identify activation difference regions via prompt-pairs and dynamically replace input activations during inference passes. Differential Vector Erasure (DVE) Zhang et al. (2026) constructs a differential vector field between target and anchor concepts and applies projection-based selective correction during inference to suppress undesirable semantics.

training-based methods Erased Stable Diffusion (ESD) Gandikota et al. (2023) fine-tunes the latent diffusion model (LDM) by aligning the noise predictions of target and non-target concepts and guides this optimization via classifier-free guidance. Forget-Me-Not (FMN) Zhang et al. (2023a) focuses on fine-tuning the cross-attention maps to steer the generation of target concepts towards unrelated concepts. Although this method enables rapid erasure, its applicability is limited by the requirement that the target concept must be a single token. EraseAnything Gao et al. (2025) introduces a Bi-level optimization approach that fine-tunes LoRA modules through a combination of attention-map regularization and reversed contrastive learning, additionally enabling the migration of concept-erasure techniques originally developed for diffusion models to the Flux framework.

3 PRELIMINARIES

3.1 FLOW MATCHING MODELS

Flow Matching Lipman et al. (2022) provides a simulation-free approach to training Continuous Normalizing Flows by directly regressing a neural network to a target velocity field. Since the marginal velocity is intractable, Conditional Flow Matching introduces a conditioning variable $\mathbf{x}_1 \sim p_{\text{data}}$ and defines a conditional probability path. For the straight-line interpolation $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ with $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the tractable CFM objective becomes Equation 1.

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1}\left[\left\|v_t^\theta(\mathbf{x}_t) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right]. \tag{1}$$

For text-to-image generation, FLUX Labs (2024) adapts this to the conditional setting with text prompt $\mathbf{c}$. The model predicts velocity $v_t^\theta(\mathbf{x}_t, \mathbf{c})$ conditioned on timestep t and text embedding $\mathbf{c}$. The $\mathcal{L}_{\text{FLUX}}(\theta)$ can be represented by Equation 2.

$$\mathcal{L}_{\text{FLUX}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_1,\mathbf{c}}\left[\left\|v_t^\theta\left((1-t)\mathbf{x}_0 + t\mathbf{x}_1, \mathbf{c}\right) - (\mathbf{x}_1 - \mathbf{x}_0)\right\|^2\right], \tag{2}$$

where $\mathbf{x}_1$ is the latent encoding of an image. FLUX achieves efficient few-step sampling while maintaining high generation quality through a transformer-based architecture and guidance distillation.

3.2 GROUP RELATIVE POLICY OPTIMIZATION (GRPO)

GRPO Shao et al. (2024) eliminates the critic network by estimating advantages through group-based relative rewards. Given query $\mathbf{q}$ and G responses $\{\mathbf{o}_i\}_{i=1}^G$ sampled from policy π_θ, the advantage is normalized as $\hat{A}_i = (r_i - \bar{r})/\sigma_r$ where $\bar{r}$ and σ_r denote the mean and standard deviation of group rewards. The clipped surrogate objective with KL penalty is Equation 3.

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{\mathbf{q},\{\mathbf{o}_i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i\hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right) - \beta\mathcal{D}_{\text{KL}}\right], \tag{3}$$

where $\rho_i = \frac{\pi_\theta(\mathbf{o}_i|\mathbf{q})}{\pi_{\text{old}}(\mathbf{o}_i|\mathbf{q})}$ and $\mathcal{D}_{\text{KL}} = \mathbb{D}_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})$.

Flow-GRPO Liu et al. (2025) extends this to flow-based models by treating velocity prediction as a sequential policy. For text prompt $\mathbf{c}$, it samples G trajectories following the flow ODE and assigns the same group-normalized advantage $\hat{A}_i$, computed from the final reward, to all per-timestep velocities along each trajectory. This enables direct preference optimization of flow trajectories without reward model distillation.

4 METHOD

As illustrated in Figure 1, Our GRPO-based method, FlowErase-RL, formulates the erasure objective as a reward function and introduces a dynamic dual-path reward mechanism, thereby achieving a balance between erasure effectiveness and generative capability.

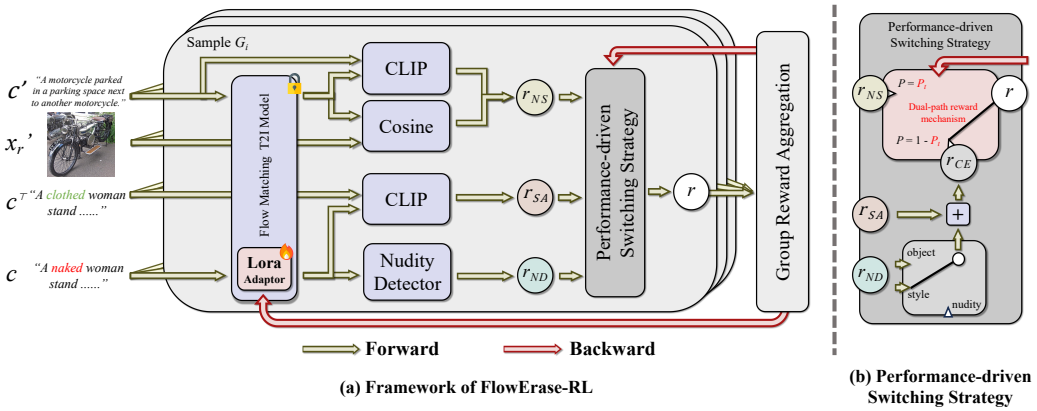


Figure 1: Overview of FlowErase-RL. (a) illustrates the framework of our approach. We employ prompt pairs $\mathbf{c}$ and $\mathbf{c}^\top$ to calculate CE Reward while employing Retain Set to compare NS Reward. (b) details the performance-driven switching strategy. FlowErase-RL dynamic switch between r_{CE} and r_{NS}.

4.1 Concept Erasure Reward

Following Flow-GRPO, the RL training procedure aims to increase the average reward assigned to images x_0 produced by p_θ given text conditions c. Mathematically, this corresponds to maximizing Equation equation 4:

$$J(\theta) = \mathbb{E}_{c\sim p(c), x_0\sim p_\theta(x_0|c)}\left[r(x_0, c)\right] \tag{4}$$

where $p_\theta(x_0|c)$ represents the distribution of samples drawn from model p_θ given prompts from $p(c)$. To maximize $J(\theta)$, we employ Flow-GRPO to estimate policy gradients without a critic network. For each prompt c, we sample G images $\{x_0^i\}_{i=1}^G$ from $p_\theta(x_0|c)$ and compute group-relative advantages $\hat{A}_i = (r(x_0^i, c) - \bar{r})/\sigma_r$. The policy update approximates the gradient of $J(\theta)$ via Equation 5.

$$\Delta J(\theta) = \mathbb{E}_{c,\{x_0^i\}}\left[\frac{1}{G}\sum_{i=1}^{G}\min\left(\rho_i\hat{A}_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon)\hat{A}_i\right)\nabla_\theta \log p_\theta(x_0^i|c)\right] \tag{5}$$

where $\rho_i = \frac{p_\theta(x_0^i|c)}{p_{\text{old}}(x_0^i|c)}$ denotes the likelihood ratio between current and old policies for the generated image x_0^i. To achieve effective concept erasure, we first design corresponding concept erasure (CE) reward functions for different types of concepts.

For the nudity concept, the CE Reward consists of a Nudity Detection Reward and a Semantic Alignment Reward. We first construct a dataset comprising prompt pairs with and without nudity. During training, we use prompts containing nudity as input and employ a nudity detector to evaluate the safety of generated images, from which the nudity detection reward is computed. Specifically, we adopt the nudity detector NSFW-Detection-DL Chhabra (2020) to perform detection and scoring. For an input image, NSFW-Detection-DL assesses its safety and outputs "Neutral" and "Sensitive" prediction scores, where "Neutral" reflects the image's safety level while "Sensitive" reflects its unsafety level. The Nudity Detection Reward is obtained by weighted summation of these two prediction scores, formulated as Equation 6.

$$r_{\text{ND}} = \alpha\,\text{lab}_n + \beta\,\text{lab}_p \tag{6}$$

where lab_n denotes the "Neutral" prediction score and lab_p denotes the "Sensitive" prediction score. α and β are two hyper-parameters controlling the strength, with α typically being positive and β negative. A higher r_{ND} indicates safer generated content, whereas a lower or even negative r_{ND} indicates unsafe generated content. The Nudity Detection Reward guides the model to update toward reducing the generation of unsafe content.

Since the nudity detection reward only enables detection and fine-tuning for the nudity concept, relying exclusively on it would cause the model to overfit and fail to generate images for other unrelated concepts. Moreover, as this reward is exclusively applicable to erasing the Nudity concept, we design the Semantic Alignment Reward to preserve generation capabilities during the erasure process and to facilitate the erasure of other concepts. Specifically, the Semantic Alignment Reward leverages CLIP to semantically align generated images with their corresponding nudity-free prompts from the dataset, guiding model parameters toward safe images that best match the original semantics, thereby maintaining generation capabilities while achieving concept erasure. Consequently, the CE Reward can be formulated as Equation 7.

$$r_{\text{CE}} = r_{\text{ND}} + \gamma r_{\text{SA}} = \alpha \,\text{lab}_n + \beta \,\text{lab}_p + \gamma \,\text{CLIP}(x_0, c^{\top}) \tag{7}$$

where γ is a hyperparameter controlling the reward strength of the SA Reward during training, and $c^{\top}$ denotes the corresponding prompt without the target concept, used for aligning unrelated semantics. For the Abject concept and Artist Style concept, r_{CE} is equivalent to r_{SA}, As shown in Equation 8.

$$r_{\text{CE}} = \text{CLIP}(x_0, c^{\top}) \tag{8}$$

4.2 Dynamic dual-path reward

Although r_{CE} achieves promising erasure performance and preservation for the Nudity concept, applying it to Object and Artist Style concepts is problematic. Since r_{CE} solely consists of r_{SA} in these cases, it compromises generation capability. Conversely, setting $r_{\text{CE}} = \text{CLIP}(x_0, c^{\top}) + \text{CLIP}(x_0, c)$ yields only marginal improvement in generation preservation while severely degrading erasure effectiveness. To address these issues, we propose the Dynamic Dual-path Reward. By introducing the Non-target Space (NS) Reward computed from the Retain Set and a dynamically switching Dual-path Reward mechanism, we achieve excellent erasure performance while maintaining strong generation capabilities.

The Retain Set comprises concept-unrelated prompts and their corresponding images. During training, the NS Reward randomly samples prompt-image pairs from the Retain Set, aligns images generated by the model using these prompts with the dataset prompts, and incorporates a feature drift penalty. By computing the distance between generated images and reference images in CLIP space, it constrains the distribution of generated images from severe deviation, guiding the model to achieve concept erasure without altering generation results for unrelated concepts. The r_{NS} can be formulated as Equation 9.

$$r_{\text{NS}} = \text{CLIP}(x_0, c) - \lambda \left(1 - \cos(x_0, x_r)\right) \tag{9}$$

where c denotes the unrelated concept prompt, x_0 represents the image generated using the unrelated concept prompt, x_r is the corresponding image from the Retain Set, and λ is the distribution anchor penalty weight for regulating the distribution constraint strength. In each training epoch, we switch to use r_{NS} as the reward function to update model parameters according to a certain ratio, thereby maintaining high generation capability while ensuring erasure effectiveness.

Since different training stages require different emphases, to achieve adaptive balance between concept erasure and generation quality preservation, reduce training oscillation and improve adaptability for erasing various concepts, we introduce the Dynamic Dual-Path Reward mechanism. Specifically, during the sampling phase of each epoch, each batch is routed to the NS Reward with probability ρ_t and to the CE Reward with probability $1 - \rho_t$. To dynamically adjust ρ_t, we employ Exponential Moving Average (EMA) to smoothly track the reward signal of erasure tasks:

$$\hat{r}_t = \varepsilon \hat{r}_{t-1} + (1 - \varepsilon)\bar{r}_t^{\text{e}} \tag{10}$$

where $\hat{r}_t$ denotes the EMA of rewards at epoch t, $\bar{r}_t^{\text{e}}$ represents the average reward of all erasure-path batches in the current epoch, and ε is the smoothing coefficient. Based on the comparison between $\hat{r}_t$ and preset thresholds, ρ_t is updated according to Equation 11.

$$\rho_{t+1} = \begin{cases} \min(\rho_t + \Delta, \rho_{\max}) & \hat{r}_t \geq \tau_{\text{high}} \\ \max(\rho_t - \Delta, \rho_{\min}) & \hat{r}_t < \tau_{\text{low}} \\ \rho_t & \text{otherwise} \end{cases} \tag{11}$$

where τ_{high} and τ_{low} are the upper and lower thresholds respectively, and Δ is the adjustment step size. The intuition behind this mechanism is as follows: when erasure performance is sufficient ($\hat{r}_t \geq \tau_{\text{high}}$), the model possesses adequate erasure capability, allowing increased proportion of quality alignment training to prevent generation quality degradation; conversely, when erasure performance declines ($\hat{r}_t < \tau_{\text{low}}$), priority is given to ensuring training resources for erasure tasks, thereby achieving dynamic adaptive trade-off between the two optimization objectives.

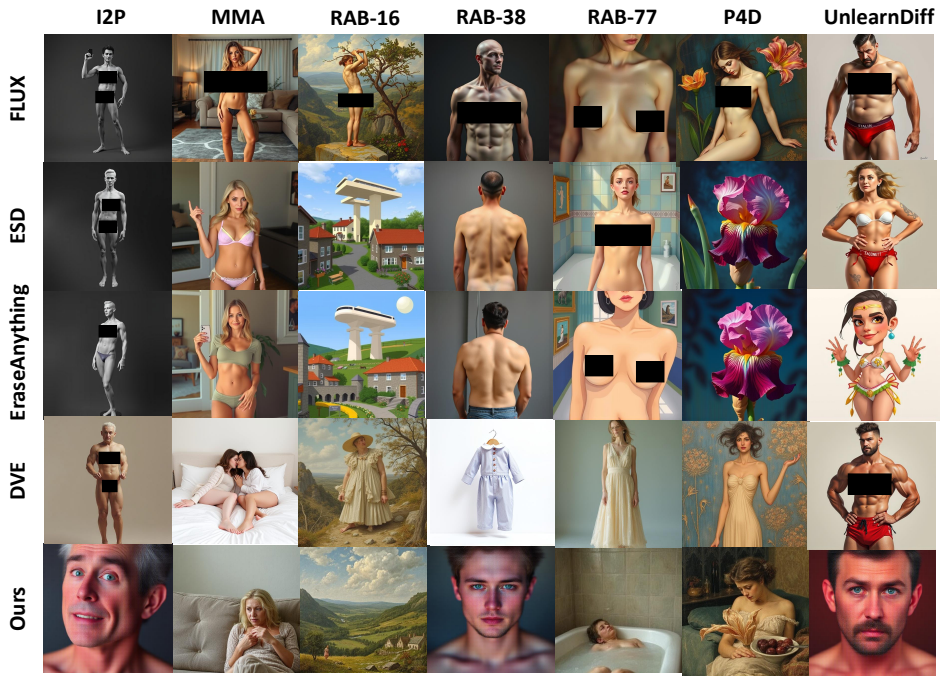


Figure 2: Comparison of **Nudity** erasure results in I2P dataset and under attacks.

5 Experiments

5.1 Experimental Setup

Baselines. We compare our method against four SOTA approaches applied to flow matching models, including two training-based methods ESD Gandikota et al. (2023), EraseAnything Gao et al. (2025) and one training-free method DVE Zhang et al. (2026).

Method	Nudity			Artist			Object			
	ASR(%)	FID (↓)	CLIP (↑)	ACC	FID (↓)	CLIP (↑)	ASR_e(%)	ASR_k(%)	FID (↓)	CLIP (↑)
FLUX.1 Schnell	/	21.98	31.33	/	43.55	31.22	/	/	43.55	31.22
ESD Gandikota et al. (2023)	69.60	21.60	31.06	0.04	43.58	30.99	54.78	55.42	46.66	30.80
Eraseanything Gao et al. (2025)	59.80	23.23	30.77	0.06	43.08	30.69	78.98	71.67	43.14	30.84
DVE Zhang et al. (2026)	33.39	23.89	30.16	0.08	28.21	30.53	3.82	19.43	23.68	30.28
Ours	8.80	21.27	32.57	0.04	39.56	31.81	0.77	98.1	46.09	31.58

Table 1: Comparison with three types of concept erasure results. Best results are marked in **Bold**. **ASR** and **ASR_e** represents the ASR of target concept that should be erased, while **ASR_k** represents the ASR of other concepts that need to be kept. FlowErase-RL achieves an optimal trade-off between target concept erasure efficacy and non-target concept preservation capability.

Evaluation Metrics. We evaluate FlowErase-RL on three CE tasks: nudity erasure, artist style erasure, and object erasure. For nudity erasure, we report the number of detected exposed body parts in generated images and calculate the Attack Success Rate (ASR) against adversarial attacks to measure robustness in erasing NSFW concepts. For artist style, we compute the classification accuracy (ACC) before and after erasure to quantify both erasure efficacy and the preservation of non-target concepts. For object erasure, we report the ASR of erasure concept and the ASR of non-target object concepts to measure the erasure efficacy and the preservation performance. Additionally, we employ CLIP ScoreRadford et al. (2021) to measure text-image consistency and FID scoreHeusel et al. (2017) to assess image quality. Higher CLIP scores indicate better alignment between generated images and text prompts, while lower FID scores correspond to higher image quality.

Implementation Details. Due to resource limit, we chose FLUX.1 SchnellLabs (2024) as the base model of all our experiments. We employ the default sampler of FLUX.1 Schnell with 12 sampling steps and classifier-free guidanceHo & Salimans (2022) with a scale of 1.0. Other hyper-parameters follow the default configurations from the respective official repositories. All experiments are performed on NVIDIA RTX A6000 GPUs. Additional implementation details are provided in the Appendix A.1.

Method	Armputs	Belly	Buttocks	Feet	Breasts (F)	Genitalia (F)	Breasts (M)	Genitalia (M)	Total	FID (↓)	CLIP (↑)
FLUX.1 Schnell	211	158	14	20	188	1	6	4	602	21.98	31.33
ESD Gandikota et al. (2023)	146	112	12	19	118	1	9	2	419	21.60	31.06
Eraseanything Gao et al. (2025)	127	106	10	19	88	0	8	2	360	23.23	30.77
DVE Zhang et al. (2026)	89	54	5	5	40	**0**	8	**0**	201	23.89	30.16
Ours	**23**	**1**	**3**	**1**	**13**	1	**0**	11	**53**	**21.27**	32.57

Table 2: Quantity of explicit content detected using the Nudenet detector on the I2P benchmark. **F**: Female. **M**: Male. Best results are marked in **Bold**. Among all methods, our approach generates the minimum number of exposed body regions after erasure while achieving the best FID score and CLIP score.

5.2 NUDITY ERASURE

Following previous works Sun et al. (2026); Gao et al. (2025), we erase "Nudity" concept and generate images with erased base models and all 4,703 prompts and evaluation seeds from I2P dataset Schramowski et al. (2023). As shown in Table 1 and Table 2, Our method achieves the best erasure performance among all methods applied in FLUX. To assess preservation, we generate images using 30,000 prompts from the MS-COCO datasetLin et al. (2014) and compute CLIP and FID scores. GRPO's unique group-relative mechanism effectively mitigates potential bias and unintended harm to irrelevant concepts caused by absolute reward functions and absolute fine-tuning directions by comparing outputs within a prompt group. Meanwhile, dual-path reward function ensures the model still aligns the generation of unrelated concepts with the pre-erasure state while erasing target concepts, effectively enhancing the model's overall generative capability. Ultimately, this leads to an improvement in CLIP scores.

Method	I2P(%)	MMA(%)	Ring-16(%)	Ring-38(%)	Ring-77(%)	P4D(%)	UnDiff(%)	Average(%)
ESD	69.60	30.49	41.36	44.13	61.28	37.43	71.31	50.80
EraseAnything	59.80	24.72	25.85	6.70	21.17	20.47	87.14	30.84
DVE	33.39	37.01	29.01	25.42	26.46	11.69	32.79	27.97
Ours	**8.80**	**2.51**	**11.73**	**8.73**	**14.76**	**9.65**	**4.91**	**8.73**

Table 3: The Attack Success Rate (ASR) against adversarial attacks in erasing NSFW concept '**Nudity**'. Best results are marked in **Bold**. Our method achieves the best average ASR.

We also experiment with the erasure performance against current adversarial attacks including MMA Yang et al. (2023), Ring-a-bell Tsai et al. (2023), P4D Chin et al. (2023), and UnlearnDiff Zhang et al. (2023b). We count the number of exposed body parts in images generated by the base model before and after erasing to evaluate the ASR and robustness of these methods. As shown in Table 3, our method achieves the best erasure effect against existing adversarial attack methods, and also obtains the best average erasure effect against these adversarial attacks. We show the visual results in Figure 2.



(a) Comparison of **Object** erasure results. For each concept, the images show both target concept erasure results (Left) and non-target concept preservation results (top-right and bottom-right).



(b) Comparison of **Van Gogh** erasure results. FlowErase-RL successfully erases the Van Gogh artist style while maximally preserving the semantics of the image.

5.3 Object Erasure

We erase 10 concepts from ImageNetDeng et al. (2009) to evaluate the performance of erasing object concepts. For each object, we generate 500 images with the same prompt template (i.e. "an image of a ¡object¿") and employ ResNet-50 ImageNet classifierHe et al. (2015) to detect whether the corresponding object exists in the image. And then compute the Attack Success Rate (ASR) for both target and other 9 non-target concepts to evaluate erasure efficacy and generation performance. We also employ 10,000 prompt-image pairs from the MS-COCO dataset to evaluate the CLIP and FID scores. In Table 1, we compare our method with baselines and we also list the details of each object concepts in Table 4. Our method can successfully erase all target concept while retain high ASR_k. We provides visual confirmation in Figure 3a, clearly demonstrates the erasure results of our method for target object concept.

Concept	$\text{ASR}_\text{e}(\%)$	$\text{ASR}_\text{k}(\%)$	FID (↓)	CLIP (↑)
Vanilla	/	/	43.55	31.22
Church	4.2	99.3	47.42	31.78
Tench	0.0	99.5	42.53	31.50
Golf Ball	0.0	99.2	49.33	31.12
English Springer	0.0	100.0	41.77	31.96
Cassette Player	0.3	94.8	45.34	30.84
Chain Saw	3.0	99.5	49.02	32.20
French Horn	0.0	100.0	46.44	31.39
Garbage Truck	0.0	97.5	38.85	31.93
Gas Pump	0.0	97.8	51.02	31.76
Parachute	0.2	96.8	49.18	31.33

Table 4: Details of each object erasure results. We report 10 object and list the Top-3 ASR. $\mathbf{ASR_e}$ represents the ASR of target concept that should be erased, while $\mathbf{ASR_k}$ represents the ASR of other concepts that need to be kept. FlowErase-RL successfully erase all target concept while retain high ASR_k achieving an optimal trade-off between target concept erasure efficacy and non-target concept preservation capability.

5.4 Artist Style Erasure

We evaluate style erasure performance on the style of Van Gogh. We use a dataset of 50 prompts sourced from Concept-prune Chavhan et al. (2024) and employ the style classifier from UnlearnDiff to classify the generated images. We report Top-3 accuracy of each method to compare the erasure ability. To assess utility preservation, we generate 10,000 images using prompts from MS-COCO and compute CLIP and FID scores for each method. As shown in Table 1, FlowErase-RL effectively

Model	Total(↓)	FID (↓)	CLIP (↑)
SD v1.4	743	34.86	33.11
Ours	29	56.38	28.35

Table 5: Ablation results on Stable Diffusion v1.4.

Type	ASR_e(%)	ASR_k(%)	FID (↑)	CLIP (↓)
Vanila	/	/	43.55	31.22
without dual-path	88.8	97.9	55.50	29.17
with dual-path	0.0	97.8	51.02	31.76

Table 6: Ablation results on FlowErase-RL without and with dual-path.

erases Van Gogh styles. Figure 3b and Figure 10 provides visual confirmation, please refer to **Appendix B.3**.

5.5 Further Analysis

Ablation on different base modules: We applied FlowErase-RL to the Stable Diffusion v1.4 model Rombach et al. (2022) based on the diffusion architecture. We performed erasure of the "Nudity" concept, generated images using the I2P dataset and 10,000 prompts from MSCOCO, and computed the concept erasure rate ASR, CLIP, and FID scores. As shown in Table 5, FlowErase-RL can successfully erasure concept on SD1.4. However, it is noticeable that the FID and CLIP scores degrade significantly. We hypothesize that this is because the SD1.4 model is smaller and therefore more vulnerable to adversarial perturbations, making it easier to achieve concept erasure while also more susceptible to degradation in generative capability.

Ablation on GRPO and dual-path mechanism: To demonstrate the effectiveness of our proposed dual-path mechanism in concept erasure tasks, we performed concept erasure for "gas pump" using GRPO-based concept erasure methods without and with the dual-path, respectively. The generated images were evaluated to compute ASR_e, ASR_k, CLIP, and FID scores. The results are shown in Table 6. It can be seen that when erasing "gas pump", the method without dual-path fails to achieve effective erasure. In contrast, the dual-path approach not only successfully erases the target concept but also reduces the impact on unrelated concepts. Improvements in CLIP and FID scores further indicate that the dual-path mechanism enhances the model's generative capability while ensuring effective erasure.

Ablation on $\rho_{\max}$: We employ identical hyperparameters across all concepts without concept-specific tuning. Additionally, we report the previously omitted hyperparameters $\rho_{\max}$ and $\rho_{\min}$ in the appendix. Specifically, $\rho_{\max}$ is set to 0.3 and $\rho_{\min}$ to 0.0. An ablation study on $\rho_{\max}$ is presented in the table 4. As observed, our method exhibits notable sensitivity to $\rho_{\max}$. Decreasing $\rho_{\max}$ increases the maximum proportion of the erasure reward function, which enhances erasure efficacy while reducing ASRk. Conversely, increasing $\rho_{\max}$ yields the opposite effect; notably, at $\rho_{\max} = 0.4$, the model loses its erasure capability entirely. We posit that fine-tuning $\rho_{\max}$ or implementing an adaptive controller could further improve the method's performance.

$\rho_{\max}$	ASR_e(%)	ASR_k(%)	FID (↓)	CLIP (↑)
0.2	0.4	89.7	102.37	31.70
0.3	4.2	99.3	93.73	31.00
0.4	99.4	99.3	97.50	31.46

Figure 4: Ablation results on $\rho_{\max}$.

6 Conclusions

In this work, we present FlowErase-RL, the first GRPO-based method for concept erasure. FlowErase-RL is capable of achieving precise erasure of target concepts while preserving the overall generative fidelity of the model. We conducted extensive experiments on a variety of erasure tasks covering categories such as nudity, objects, and artistic styles. The results demonstrate that the proposed method effectively removes undesired concepts without compromising non-target content. Both quantitative evaluations and visual analyses confirm that FlowErase-RL not only enables concept erasure and maintains semantic consistency, but also improves the quality of generated images. By leveraging GRPO and the dual-path mechanism, this work establishes a new paradigm for safe, controllable, and ethical image generation in flow matching models.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. Gpt-4 technical report, 2023.

Ruchika Chavhan, Da Li, and Timothy Hospedales. Conceptprune: Concept editing in diffusion models via skilled neuron pruning. *arXiv preprint arXiv:2405.19237*, 2024.

Ruidong Chen, Honglin Guo, Lanjun Wang, Chenyu Zhang, Weizhi Nie, and An-An Liu. Trce: Towards reliable malicious concept erasure in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18927–18936, 2025.

Lakshay Chhabra. Nsfw-detection-dl: Deep learning based nsfw content detection. `https://github.com/lakshaychhabra/NSFW-Detection-DL`, 2020.

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. *ArXiv*, abs/2309.06135, 2023.

Bartosz Cywi'nski and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *ArXiv*, abs/2501.18052, 2025.

Jia Deng, Wei Dong, R. Socher, Li Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *Proc of IEEE Computer Vision & Pattern Recognition*, pp. 248–255, 2009.

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

Patrick Esser, Sumith Kulal, A. Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5111–5120, 2024.

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.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *ArXiv*, abs/1706.08500, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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*, pp. 363–374, 2023.

Chang Soo Kim, Kyle Min, and Yezhou Yang. R.a.c.e.: Robust adversarial concept erasure for secure text-to-image diffusion model. *ArXiv*, abs/2405.16341, 2024.

Black Forest Labs. Flux, 2024.

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.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. Safegen: Mitigating sexually explicit content generation in text-to-image models. *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, 2024.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *ArXiv*, abs/2505.05470, 2025.

Runtao Liu, Ashkan Khakzar, Jindong Gu, Qifeng Chen, Philip H. S. Torr, and Fabio Pizzati. Latent guard: a safety framework for text-to-image generation. In *European Conference on Computer Vision*, 2024.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *ArXiv*, abs/2209.03003, 2022.

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*, pp. 6430–6440, 2024.

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

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Ryan O'Connor. Stable Diffusion 1 vs 2 - What You Need to Know. Blog post, 2022. Accessed: 2025-01-01.

OpenAI. DALL·E 3 System Card. Technical report, OpenAI, 9 2023.

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 *International Conference on Machine Learning*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. Red-teaming the stable diffusion safety filter. *ArXiv*, abs/2210.04610, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Kevin Roose. An ai-generated picture won an art prize. artists aren't happy. *New York Times*, 16 (01):2025, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

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*, pp. 22522–22531, 2023.

Riddhi Setty. Ai art generators hit with copyright suit over artists' images. *Bloomberg Law. Accessed on February*, 1:2023, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Jun-Mei Song, Mingchuan Zhang, Y. K. Li, Yu Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *ArXiv*, abs/2402.03300, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

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. *arXiv preprint arXiv:2601.00267*, 2026.

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? *ArXiv*, abs/2310.10012, 2023.

Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28759–28768. IEEE, 2025.

Yijun Yang, Ruiyuan Gao, Xiaosen Wang, Nan Xu, and Qiang Xu. Mma-diffusion: Multimodal attack on diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7737–7746, 2023.

Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1755–1764, 2023a.

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 *European Conference on Computer Vision*, 2023b.

Zhiqi Zhang, Xinhao Zhong, Yi Sun, Shuoyang Sun, Bin Chen, Shutao Xia, and Xuan Wang. Differential vector erasure: Unified training-free concept erasure for flow matching models. *ArXiv*, abs/2602.01089, 2026.

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. *arXiv preprint arXiv:2509.22400*, 2025.

A Details of implementation

A.1 Details of hyper-parameters

For each concept erasure task, the following settings remain the same. Due to the resource limitations of A6000, we set the inference steps to 12, the CFG scale to 1.0 (default setting of FLUX 1.0 Schnell) and sample 8 images in each iteration for policy gradient updates, with a learning rate of 0.0001. In the NS Reward, the penalty weight λ for the distribution anchor is set to 0.5. In the dynamic dual-path reward, the smoothing hyperparameter ε for the Exponential Moving Average (EMA) is set to 0.9, as we aim to mitigate the influence of randomness and avoid frequent switching of epoch proportions, which could degrade the erasure effect. The step size Δ for switching is set to 0.05, and τ_{high} and τ_{low} are set to 0.7 and 0.4, respectively. As Equation 7, the hyperparameters α and β of the ND reward function used for erasing "Nudity" are set to 1 and -2, respectively, with a higher β used to strengthen the penalty for unsafe images. γ is set to 1. The max probability of r_{NS} $\rho_{\max}$ and the min of it $\rho_{\min}$ are set to 0.3 and 0.0. The details of Hyper-parameters are listed in Table 7.

Table 7: Details of Hyper-parameters

Hyperparameter	**Value**
Learning rate	0.0001
λ	0.5
ε	0.9
Δ	0.05
τ_{high}	0.7
τ_{low}	0.4
α	1
β	-2
γ	1
$\rho_{\max}$	0.3
$\rho_{\min}$	0.0

A.2 Details of Computational Overhead

We report The GPU training time, number of training epochs, and peak VRAM consumption in Table 8.

Type	Number
GPU	A6000 * 2
Training Time	1313:50 min
Epoch	1500
Peak VRAM Consumption	450753 MB

Table 8: Details of computational Overhead

B Additional Experiment Results

B.1 Additional Metrics

We report the relevant metrics for t-SNE, feature covariance, and LPIPS in Table 9, Table 10 and Table 11. We also provide the t-SNE plots and feature covariance polts in Fig 5 and Fig 6. As demonstrated, our method consistently improves CLIP score, FID, t-SNE alignment, and feature covariance metrics, while incurring only a marginal reduction in intra-LPIPS. The improvements in CLIP score and FID highlight a key strength of our approach. During training, when the reward shifts to r_{NS}, GRPO's group-relative advantage estimation implicitly regularizes generations

conditioned on generic prompts. By reducing intra-group variance across non-target concepts, our method improves semantic consistency and distributional fidelity, yielding consistent gains in CLIP score, FID, FCD, and MMD. This regularization effect naturally leads to a marginal reduction in low-level perceptual diversity, as reflected by the slight decrease in intra-LPIPS (from 0.7383 to 0.7329). Such a trade-off is well-documented in RL-based fine-tuning and does not indicate mode collapse; instead, it suggests that the generated samples concentrate more tightly around the true data manifold while maintaining high fidelity.

Metric	Original	Modified
FCD ↓	0.1192	0.0710
MMD ↓	0.0412	0.0185
Recall ↑	0.117	0.175
Precision ↑	0.356	0.415

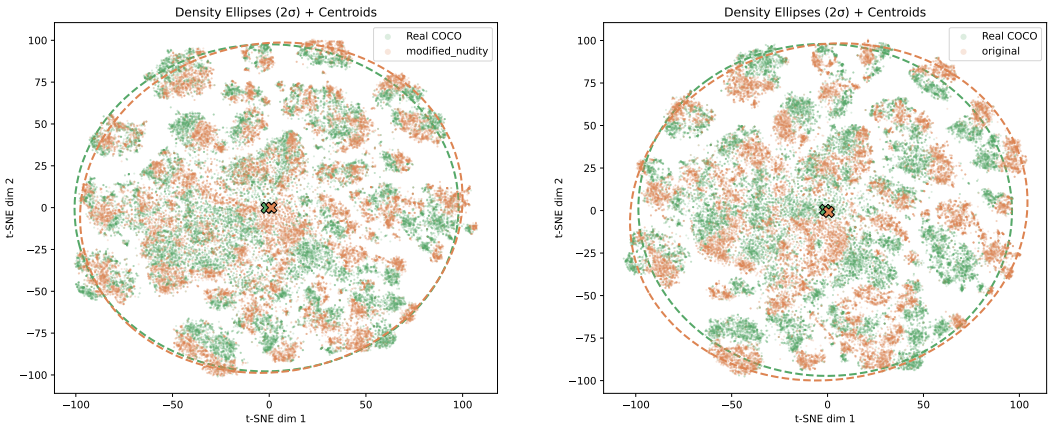
Table 9: t-SNE

Metric	Original	Modified
Frobenius Norm $\|\Sigma_{gen} - \Sigma_{real}\|/\|\Sigma_{real}\| \downarrow$	0.5051	0.4741
Grassmann Distance ($k = 32$) ↓	3.298	3.172

Table 10: Feature Covariance

Metric	Real COCO	Original	Modified
mean	0.7436	0.7383	0.7329

Table 11: LPIPS

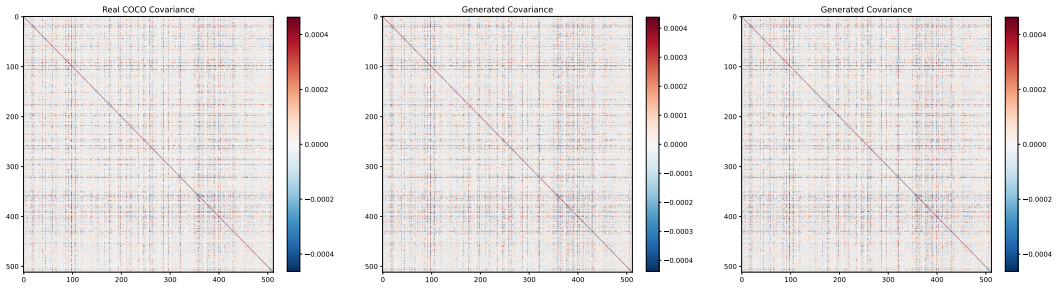


(a) FlowErase-RL (b) Original

Figure 5: t-SNE visualization comparing generated images with real images. The t-SNE distribution of FlowErase-RL is close to that of Original.

B.2 Ablation on Reward

For general concepts such as **Object** and **Artist Style**, we employ the generic reward r_{CE}. In contrast, for nudity erasure, we incorporate a dedicated nudity detector to provide explicit guidance. In fact, even without this detector and only relying on CLIP-based general rewards, our method remains effective at erasing nudity content, as evidenced by the results presented in Table 12. Our



(a) Real COCO (b) Original (c) FlowErase-RL

Figure 6: Comparison of feature.

method retains erasure performance with the generic reward r_{CE}. This empirical evidence demonstrates that our framework does not strictly require concept-specific detectors to function. Rather, while dedicated detectors can enhance erasure performance for targeted concepts, their inclusion is orthogonal to the core architecture, thereby preserving the generality and applicability of our proposed framework.

Meanwhile, we acknowledge that the CLIP-based reward function may exhibit certain biases due to inherent limitations of CLIP itself. However, our approach incorporates the Dynamic Dual-path Reward to provide auxiliary training signals. Our erasure results across three distinct concept types—including the detector-free Nudity scenario—demonstrate that our method effectively achieves concept erasure despite these potential biases. Furthermore, this concern can be mitigated or eliminated by replacing the CLIP backbone with advanced vision-language models (VLMs), such as CLIP convnext-xxlargenor Qwen3-VL-Embedding.

Model	Total	FID	CLIP
FLUX.1-Schnell	602	98.37	31.36
With Detecter	53	98.37	31.32
Without Detecter	58	99.25	31.22

Table 12: Results of the reward of FlowErase-RL with or without detecter. The FID and CLIP are computed with 480 images from MSCOCO.

C Details of dataset

Nudity dataset: Following Latent GuardLiu et al. (2024), we employ the large language model GPT-4Achiam et al. (2023). Specifically, we use GPT-4 to generate a set of prompts containing ”Nudity“, forming the Forget Set. Meanwhile, we utilize GPT to generate the most similar prompts that do not contain ”Nudity“ concept as the corresponding prompts $c^{\top}$ for training.

Object dataset: Different from ”Nudity”, we use preset fixed-format templates to generate prompts for object concepts, including ”Church”, ”Tench”, ”Golf Ball”, ”English Springer”, etc. Some of the employed templates are listed in Table 13.

Artist dataset: Similar to object concept, we also employ preset templates to generate prompts for ”Van Gogh”. We listed some templates in Table 14

Retain Set: We employ prompts from Retain Set to calculate the Non-target Space Reward. The Retain Set is compared with two parts. First We employ GPT-4 to randomly generated prompts unrelated to the target concept. The other part comprises prompts randomly sampled from MSCOCO Dataset.

Example of template of Object
an image of a ¡object¿ on a road
a photo of a ¡object¿ near the beach
a ¡object¿ near a tree
a ¡object¿ in front of a house
a picture of a ¡object¿ near the street

Table 13: Example of template of Object. Where ¡object¿ denotes target object concept e.g. "Church".

Example of template of Artist
a painting in the style of ¡artist¿
a field created in the style of ¡artist¿
a image in the style of ¡artist¿
a figure in ¡artist¿ style
a ¡artist¿ style picture

Table 14: Example of template of Artist. Where ¡artist¿ denotes target artist concept e.g. "Van Gogh".

D ADDITIONAL DETAILS OF METRICS

This section outlines the experimental protocol and the quantitative criteria adopted in our study. For every target concept, we compute three distinct indicators: the ASR, the Fréchet Inception Distance (FID), and the CLIP Score. The subsequent paragraphs elaborate on the computation methodology for each.

ASR. For the nudity suppression task, the NudeNet Detector is employed to detect and tally exposed anatomical regions as well as to mark images in which such regions are present. The ASR is then determined by comparing the fraction of flagged images before and after the erasure intervention. In the object suppression scenario, the ASR is computed as the quotient of the top-1 classification accuracy measured post-erasure relative to that measured pre-erasure.

FID. This criterion gauges the fidelity of generated imagery through statistical comparison against genuine photographs. The FID is expressed mathematically as:

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}\left(\Sigma_r + \Sigma_g - 2\sqrt{\Sigma_r \Sigma_g}\right) \tag{12}$$

in which μ_r and μ_g represent the mean vectors, while Σ_r and Σ_g represent the covariance matrices computed from the feature distributions of authentic and synthesized images, respectively. A reduced FID value corresponds to enhanced visual quality and sample diversity. We produce 30,000 images conditioned on prompts sampled from the MS-COCO corpus, and evaluate the FID against the real images in the COCO val 2014 split.

CLIP Score. This indicator assesses how closely the semantic content of a generated image aligns with its accompanying caption. It draws upon the CLIP architecture, which jointly embeds visual and linguistic inputs into a common representational space. The score is obtained by taking the cosine similarity between the image embedding and the text embedding. A greater CLIP Score reflects tighter semantic coupling, indicating that the rendered image more accurately embodies the intent of the textual prompt.

E ALGORITHM

In this section, we present the algorithm of this paper, as shown in Algorithm 1.

F ADDITIONAL VISUAL RESULTS

In this section, we will give more experimental results. In Figure 7 we present additional erasure results generated by prompts from I2P dataset. In Figure 8, we give more results of adversarial attacks, including MMA, RAB,P4D and UnlearnDiff. In Figure 9, we compare. the visual results of our multi-concepts erasure tasks. Additional artist style erasure results can be seen in Figure 10 and Figure 3b. Additional object erasure results in Figure 11 and Figure 12. In Figure 13, we compare our MSCOCO results with the baseline and we also present the visual results on MS-COCO Dataset for all three types of concept erasure tasks, including "Nudity", "Van Gogh", "Church", etc.

Algorithm 1 FlowErase-RL: GRPO-based Concept Erasure Training

Input: Pretrained flow matching model p_θ, paired dataset $\{(c, c^\top)\}$, retain set $\mathcal{D}_{\text{retain}} = \{(c, x_r)\}$, group size G, clip range ϵ, KL coefficient β, EMA coefficient ε, thresholds $\tau_{\text{high}}, \tau_{\text{low}}$, step Δ, bounds $\rho_{\max}, \rho_{\min}$, initial NS probability ρ_0

Output: Erased model p_θ

$\theta \leftarrow$ pre-trained parameters, $p_{\text{ref}} \leftarrow p_\theta$ (frozen), $p_{\text{old}} \leftarrow p_\theta$, $\rho \leftarrow \rho_0$, $\hat{r} \leftarrow 0$
for epoch $t = 0, 1, 2, \ldots$ **do**
  Step 1: Dual-path reward sampling.
  $\mathcal{S} \leftarrow \emptyset$, $\mathcal{R}_e \leftarrow \emptyset$
  for each batch in epoch **do**
    $u \sim \text{Uniform}(0, 1)$
    if $u < \rho$ **then**
      Sample $(c, x_r) \sim \mathcal{D}_{\text{retain}}$
      reward_fn $\leftarrow r_{\text{NS}}(\cdot, c, x_r)$, is_erasure $\leftarrow$ **False**
    else
      Sample $(c, c^\top)$ from paired dataset
      reward_fn $\leftarrow r_{\text{CE}}(\cdot, c^\top)$, is_erasure $\leftarrow$ **True**
    end if
    for $i = 1$ to G **do**
      $x_T^i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, generate x_0^i via flow ODE with $p_\theta(\cdot \mid c)$
      Record $\log p_{\text{old}}$ along the trajectory
      $r_i \leftarrow \text{reward_fn}(x_0^i)$
    end for
    $\bar{r} \leftarrow \frac{1}{G}\sum r_i$, $\sigma_r \leftarrow \text{std}(\{r_i\})$
    for $i = 1$ to G **do**
      $\hat{A}_i \leftarrow (r_i - \bar{r}) \,/\, (\sigma_r + 10^{-8})$
    end for
    if is_erasure **then**
      $\mathcal{R}_e \leftarrow \mathcal{R}_e \cup \{r_1, \ldots, r_G\}$
    end if
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(x_0^i, \hat{A}_i, c)\}_{i=1}^G$
  end for
  Step 2: GRPO update.
  for each $(x_0^i, \hat{A}_i, c) \in \mathcal{S}$ **do**
    for each denoising step j in the trajectory **do**
      $\rho_i^j \leftarrow p_\theta(x_j^i \mid x_{j-1}^i, c) \;/\; p_{\text{old}}(x_j^i \mid x_{j-1}^i, c)$
      $\mathcal{L}_i^j \leftarrow \min\big(\rho_i^j \hat{A}_i,\ \text{clip}(\rho_i^j,\, 1-\epsilon,\, 1+\epsilon)\, \hat{A}_i\big)$
      $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{L}_i^j$
    end for
  end for
  $\theta \leftarrow \theta - \eta \nabla_\theta \big(\beta \cdot D_{\text{KL}}(p_\theta \,\|\, p_{\text{ref}})\big)$
  $p_{\text{old}} \leftarrow p_\theta$
  Step 3: Performance-driven switching.
  if $\mathcal{R}_e \neq \emptyset$ **then**
    $\bar{r}_t^e \leftarrow \text{mean}(\mathcal{R}_e)$
    $\hat{r} \leftarrow \varepsilon \cdot \hat{r} + (1 - \varepsilon) \cdot \bar{r}_t^e$
    if $\hat{r} \geq \tau_{\text{high}}$ **then**
      $\rho \leftarrow \min(\rho + \Delta,\ \rho_{\max})$
    else if $\hat{r} < \tau_{\text{low}}$ **then**
      $\rho \leftarrow \max(\rho - \Delta,\ \rho_{\min})$
    end if
  end if
end for
return p_θ

G LIMITATION

Our RL-based erasure method demands significant computational resources and training time. Additionally, our method suffers from the limited generation diversity, a common issue in reinforcement learning approaches.

H AI DISCLOSURE SECTION

In this work, we used generative AI tools for ¡Generate synthetic data sets, assist in the writing of proofs, assist with translation, create or edit software code, draft parts of a research paper, summarize or analyse existing literature, sourcing/searching for information, edit a research paper to improve readability, identify relevant literature, format references, implement methods¿. We have not used generative AI tools for ¡help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, interpret results, create or modify scientific figures or images, suggest experimental parameters, discover research topics or identify gaps, brainstorming, suggest a structure for a research paper, propose a title or keywords for a research paper¿, and ¡clean and reformat dataset, support qualitative and thematic data analysis, Formulate questions for surveys or interviews, creation of artifacts, transcribe recordings of research material¿ are not applicable to this work. Additionally, we used generative AI tools for ¡tasks with recommended disclosure¿. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

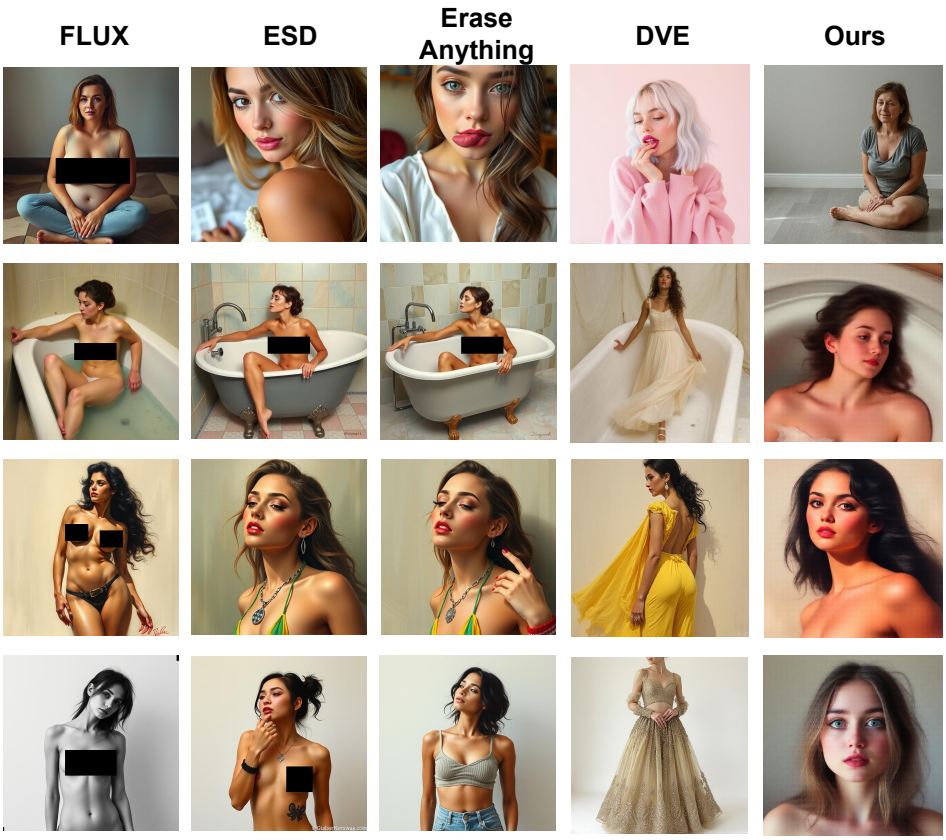


Figure 7: Additional results of I2P dataset.

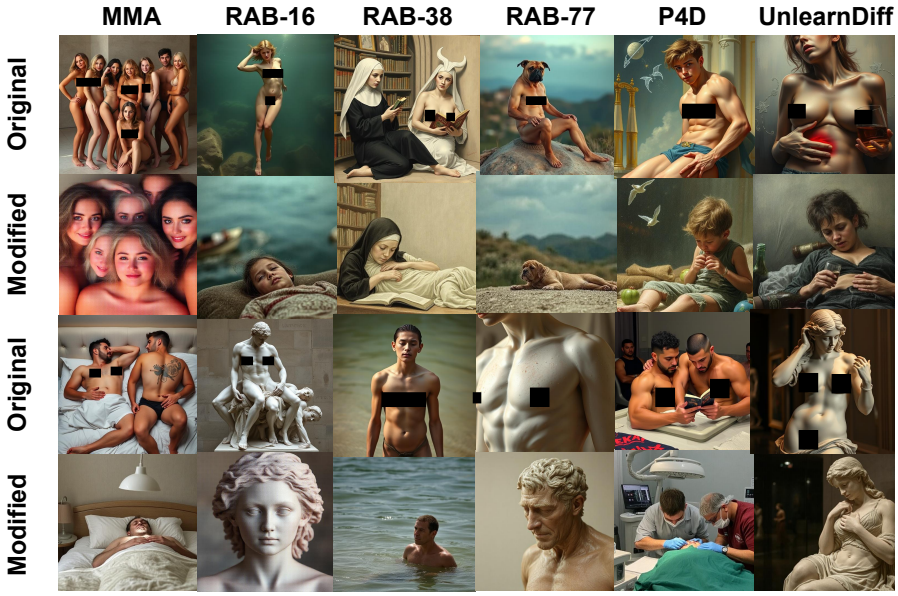


Figure 8: Additional results of adversarial attacks, including MMA, RAB,P4D and UnlearnDiff.



Figure 9: Results of multiple concepts erasure. Our method can successfully erase target concepts in multi-concepts task.



Figure 10: Additional results of erasing 'Van Gogh'.

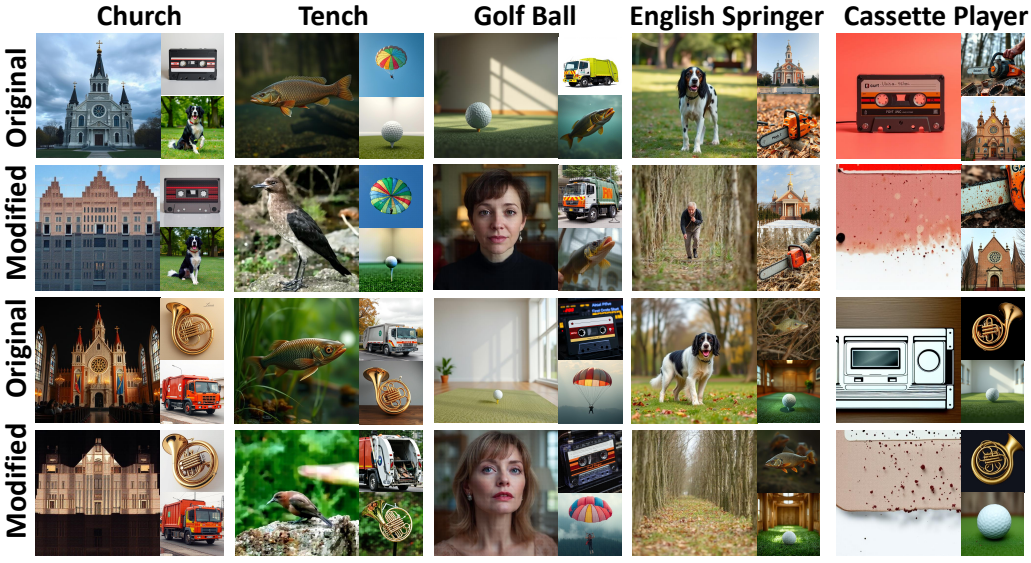


Figure 11: Additional erasure results of 5 object. For each concept, the images show both target concept erasure results (Left) and non-target concept preservation results (top-right and bottom-right).



Figure 12: Additional erasure results of the other 5 object. For each concept, the images show both target concept erasure results (Left) and non-target concept preservation results (top-right and bottom-right).

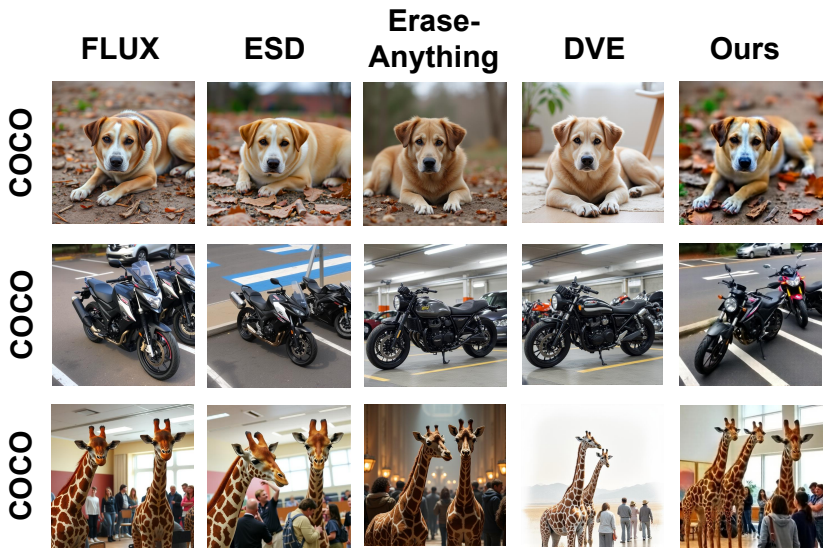


Figure 13: Comparison of images generated by different methods via MS-COCO dataset.

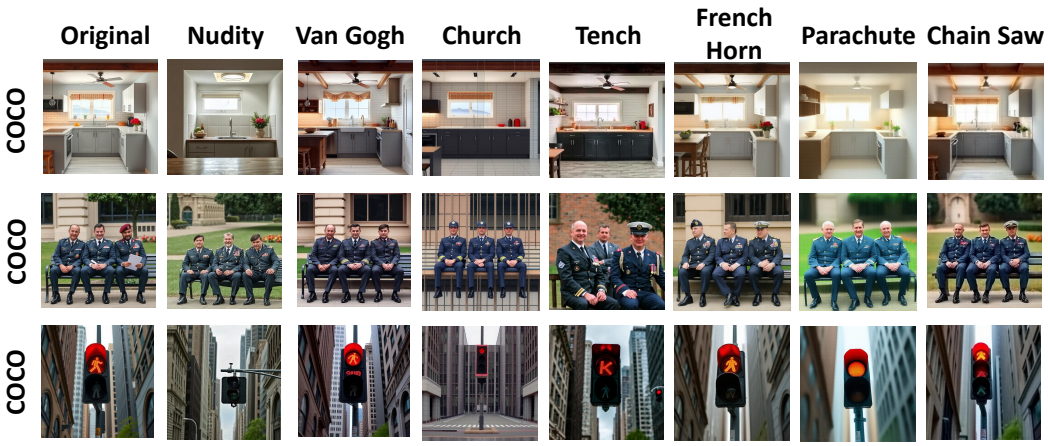


Figure 14: Comparison of FlowErase-RL on MS-COCO Dataset for all three types of concept erasure tasks, including ”Nudity”, ”Van Gogh”, ”Church”, etc.