Erasing Undesirable Concepts in Diffusion Models with Adversarial Preservation

Tuan-Anh Bui[1]

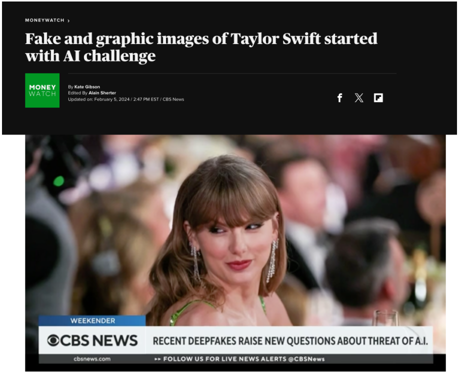
[1]Department of Data Science and AI
Faculty of Information Technology
Monash University

GenAI Reading, Mar 2024

Table of Contents

Table of Contents

Prevent misuse of AI-generated content



- **Sexually explicit AI-generated** images of Taylor Swift shared on X (Twitter). Attracted more than 45 million views, 24,000 reposts, remained live for about 17 hours before its removal. (The Verge)

Prevent misuse of AI-generated content

With just a single reference image, our Infinite-ID framework excels in synthesizing high-quality images while maintaining superior identity fidelity and text semantic consistency in various styles.

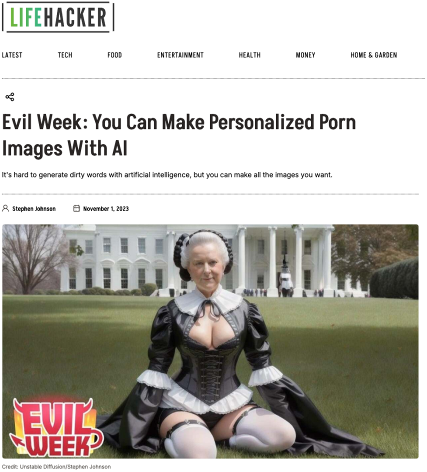
- **Personalization-GenAI** becomes extremely good[1]. The risk is now for everyone.

---

[1] Wu, et al. "Infinite-ID: Identity-preserved Personalization via ID-semantics Decoupling Paradigm." arxiv 2024

Prevent misuse of AI-generated content

- **Personalization-GenAI** becomes extremely good[1]. The risk is now for everyone. And it is already happening as reported here and here



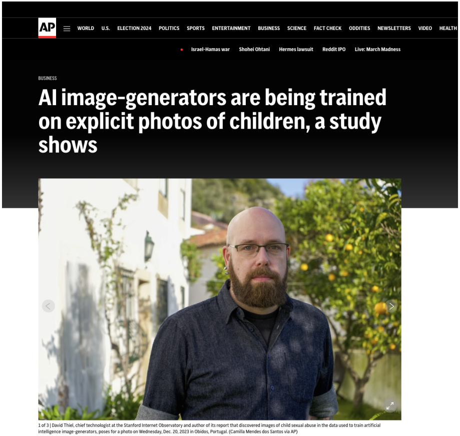


Table of Contents

Denoising Diffusion Models

In a nutshell, training a diffusion model involves two processes: a forward diffusion process where noise is gradually added to the input image, and a reverse denoising diffusion process where the model tries to predict a noise ϵ_t which is added in the forward process. More specifically, given a chain of T diffusion steps $x_0, x_1, ..., x_T$, the denoising process can be formulated as follows:

$$p_\theta(x_{T:0}) = p(x_T) \prod_{t=T}^{1} p_\theta(x_{t-1} \mid x_t) \tag{1}$$

The model is trained by minimizing the difference between the predicted noise ϵ_t and the true noise ϵ as follows:

$$\mathcal{L} = \mathbb{E}_{x_0 \sim p_{\text{data}}, t, \epsilon \sim \mathcal{N}(0, \mathbf{I})} \|\epsilon - \epsilon_\theta(x_t, t)\|_2^2 \tag{2}$$

where $\epsilon_\theta(x_t, t)$ is the predicted noise at step t by the denoising model θ.

Latent Diffusion Models

With an intuition that semantic information that controls the main concept of an image can be represented in a low-dimensional space, [1] proposed a diffusion process operating on the latent space to learn the distribution of the semantic information which can be formulated as follows:

$$p_\theta(z_{T:0}) = p(z_T) \prod_{t=T}^{1} p_\theta(z_{t-1} \mid z_t) \tag{3}$$

where $z_0 \sim \varepsilon(x_0)$ is the latent vector obtained by a pre-trained encoder ε. The objective function of the latent diffusion model as follows:

$$\mathcal{L} = \mathbb{E}_{z_0 \sim \varepsilon(x), x \sim p_{\text{data}}, t, \epsilon \sim \mathcal{N}(0,\mathbf{I})} \left\| \epsilon - \epsilon_\theta(z_t, t) \right\|_2^2 \tag{4}$$

Table of Contents

Naive Approaches

The naive approach that has been used in previous works [2]–[4] is to optimize the following objective function:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2 \right] \tag{5}$$

Where $\epsilon_\theta, \epsilon_{\theta'}$ represent output of the pre-trained *foundation* U-Net model and the *sanitized* model, respectively. c_e, c_n represent to-be-erased concept and a neutral/null input (e.g., "A photo" or " "), respectively.
Advantage: Simple yet effective in erasing concepts.
Drawback: Degradation in the quality of other concepts.

Naive Approaches

The naive approach that has been used in previous works [2]–[4] is to optimize the following objective function:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2 \right] \tag{5}$$

Where $\epsilon_\theta, \epsilon_{\theta'}$ represent output of the pre-trained *foundation* U-Net model and the *sanitized* model, respectively. c_e, c_n represent to-be-erased concept and a neutral/null input (e.g., "A photo" or " "), respectively.

Advantage: Simple yet effective in erasing concepts.

Drawback: Degradation in the quality of other concepts.

Idea: Instead of preserving *neutral concepts*, can we preserve the *most sensitive concepts* to the erasing concept?

Naive Approaches

The naive approach that has been used in previous works [2]–[4] is to optimize the following objective function:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2 \right] \tag{5}$$

Where $\epsilon_{\theta}, \epsilon_{\theta'}$ represent output of the pre-trained *foundation* U-Net model and the *sanitized* model, respectively. c_e, c_n represent to-be-erased concept and a neutral/null input (e.g., "A photo" or " "), respectively.
Advantage: Simple yet effective in erasing concepts.
Drawback: Degradation in the quality of other concepts.
Idea: Instead of preserving *neutral concepts*, can we preserve the *most sensitive concepts* to the erasing concept?
Question: *But how to measure the impact of erasing a concept on the generation of other concepts?*

Impact on the model's capability: How to measure?

Settings:

- $\epsilon_\theta(z_t, c, t)$ is the output of the model at step t with the input z_t and the concept c.
- $\mathcal{C}, \mathbf{E} \subset \mathcal{C}, \mathcal{R} = \mathcal{C} \setminus \mathbf{E}$: the entire concept space, the set of to be erased and remaining concepts, respectively.
- $\epsilon_{\theta'}(z_t, c, t)$ is the output of the *sanitized* model by removing the set of concepts $\mathbf{E}$ from the model $\epsilon_\theta(z_t, c, t)$.
- $c_e \in \mathbf{E}$, $c_n \in \mathcal{R}$: the to-be-erased and neutral concepts ("a photo" or " "), respectively.

Impact on the model's capability: How to measure?

Measuring Generation Capability with CLIP Alignment Score:

- For each concept $c \in \mathcal{C}$, generate k samples $\{G(\theta, c, z_T^i)\}_{i=1}^k$.
- Compute the CLIP alignment score $S_{\theta,i,c} = S(G(\theta, c, z_T^i), c)$
- Intepretation: The higher the score, the better the model can generate the concept c.

Impact on the model's capability: How to measure?

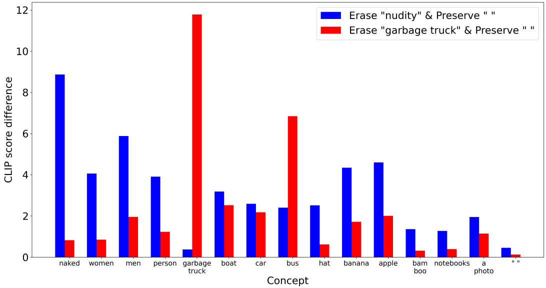
Measuring Generation Capability with CLIP Alignment Score:

- For each concept $c \in \mathcal{C}$, generate k samples $\{G(\theta, c, z_T^i)\}_{i=1}^k$.
- Compute the CLIP alignment score $S_{\theta,i,c} = S(G(\theta, c, z_T^i), c)$
- Intepretation: The higher the score, the better the model can generate the concept c.

How to measure the impact of erasing a concept c_e on the generation of other concepts $c \in \mathcal{R}$?

- Obtained the sanitized model $\theta^{'}$ by erasing the concept c_e.
- Compute the CLIP alignment score $S_{\theta^{'},i,c} = S(G(\theta^{'}, c, z_T^i), c)$.
- Compute the difference $\delta_{c_e}(c) = \frac{1}{k} \sum_{i=1}^k \left(S_{\theta,i,c} - S_{\theta'_{c_e},i,c} \right)$.
- Intepretation: The higher the score, the more the model's capability is affected by erasing the concept c_e (negatively).

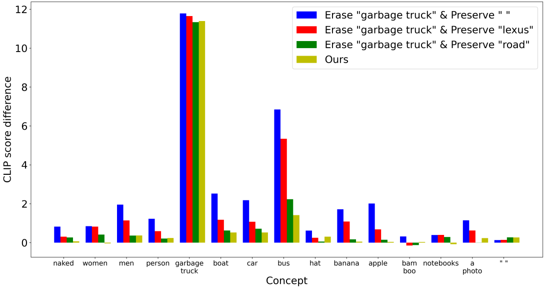
Impact on the model's capability: Results



Impact of erasing "nudity" or "garbage truck" to other concepts:

- The impact varies across different concepts.
- Affecting more related concepts than unrelated ones, i.e., erasing "nudity" affects "women", "men" than "bamboo", "notebooks", while erasing "garbage truck" affects "bus".
- Neutral concepts are very resistant to changes.

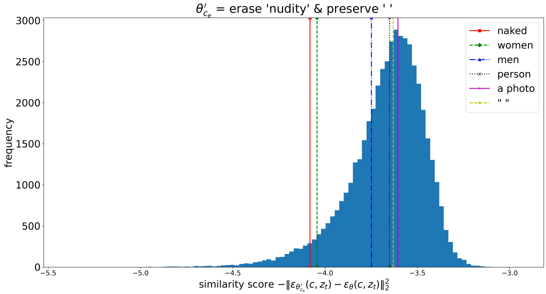
Impact on the model's capability: Results



Impact of choosing different concepts to preserve:

- Choosing the right concept to preserve is crucial.
- Preserving "road" > "lexus" > " " in maintaining the quality of other concepts.
- Early advertisment: our adaptive preservation is the best :D

Sensitivity Spectrum



Sensitivity spectrum of concepts to the target concept "nudity":

- Scanning through entire 50k concepts.
- Similarity score $-\|\epsilon_{\theta'_{c_e}}(c, z_t) - \epsilon_\theta(c, z_t)\|_2^2$. Intepretation: the higher the score, the more similar the output of two models, i.e., the less the impact of erasing the concept c_e on the concept c.

Neutral concepts lie in the middle of the spectrum. Again, not a good choice to preserve!

Table of Contents

Objective Function: First Attempt

$$\min_{\theta'} \max_{c_a \in \mathcal{R}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_\theta(c_n) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_a) - \epsilon_\theta(c_a) \right\|_2^2}_{L_2} \right] \tag{6}$$

- Minimizing L_1: Erasing the concept c_e.
- Minimizing L_2: Preserving the adversarial concept c_a.
- Maximizing L_2 w.r.t. c_a: Searching for the most sensitive concept to the erasing concept c_e.

Objective Function: Solving with PGD

$$\min_{\theta'} \max_{c_a \in \mathcal{R}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_a) - \epsilon_{\theta}(c_a) \right\|_2^2}_{L_2} \right] \tag{7}$$

Solving the optimization problem with PGD:

- Init $c_{a,t=0} = c_e = \tau($"garbage truck"$)$.
- The adversarial concept c_a quickly converges to background noise type of concept.

Continuos concept space is not suitable for adversarial preservation.

Objective Function: Relaxation with Gumbel-Softmax

$$\min_{\theta'} \max_{\pi \in \Delta_{\mathcal{R}}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\|\epsilon_{\theta'}(c_e) - \epsilon_\theta(c_n)\|_2^2}_{L_1} + \lambda \underbrace{\|\epsilon_{\theta'}(\mathbf{G}(\pi) \odot \mathcal{R}) - \epsilon_\theta(\mathbf{G}(\pi) \odot \mathcal{R})\|_2^2}_{L_2} \right] \tag{8}$$

Where $\mathbb{P}_{\mathcal{R},\pi} = \sum_{i=1}^{|\mathcal{R}|} \pi_i \delta_{e_i}$ is the distribution over the concept space $\mathcal{R}$, $\mathbf{G}(\pi)$ is the Gumbel-Softmax distribution over the concept space $\mathcal{R}$.
Instead of directly searching c_a in the continuous concept space, we switch to searching for the embedding distribution π on the simplex $\Delta_{\mathcal{R}}$.

Adversarial Concept Preservation Algorithm

Algorithm 1 Find Adversarial Concept

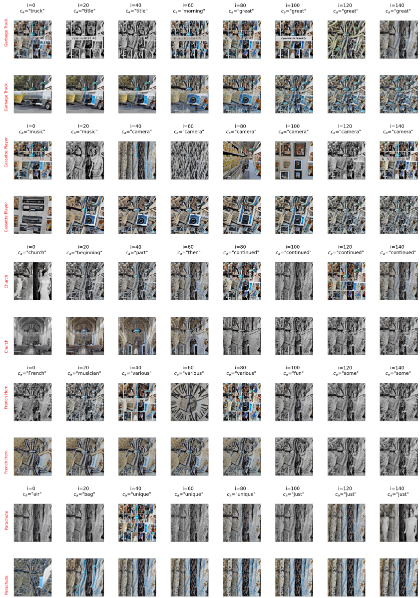
Input: $\theta, \mathcal{R}$. Searching hyperparameters: η, N_{iter}. Current state θ'_k
Output: Adversarial concept c_a
for $i = 1$ to N_{iter} **do**
  $\pi \leftarrow \pi + \eta \nabla_{\pi} \left[\|\epsilon_{\theta'}(\mathbf{G}(\pi) \odot \mathcal{R}) - \epsilon_{\theta}(\mathbf{G}(\pi) \odot \mathcal{R})\|_2^2 \right]$ ▷ Maximize L_2
end for
$c_a = \mathbf{G}(\pi^*) \odot \mathcal{R}$

Algorithm 2 Adversarial Erasure Training

Input: $\theta, \mathcal{R}, \mathbf{E}, \lambda$. Searching hyperparameters: η, N_{iter}.
Output: θ'
$k \leftarrow 0, \theta'_k \leftarrow \theta$
while Not Converged **do**
  $c_e \sim \mathbf{E}$
  $c_a \leftarrow$ FindAdversarialConcept$(\theta'_k, \theta, \mathcal{R}, \eta, N_{\text{iter}})$
  $\theta'_{k+1} \leftarrow \theta'_k - \alpha \nabla_{\theta'} [\|\epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n)\|_2^2 + \lambda \|\epsilon_{\theta'}(c_a) - \epsilon_{\theta}(c_a)\|_2^2]$ ▷ Outer min
end while

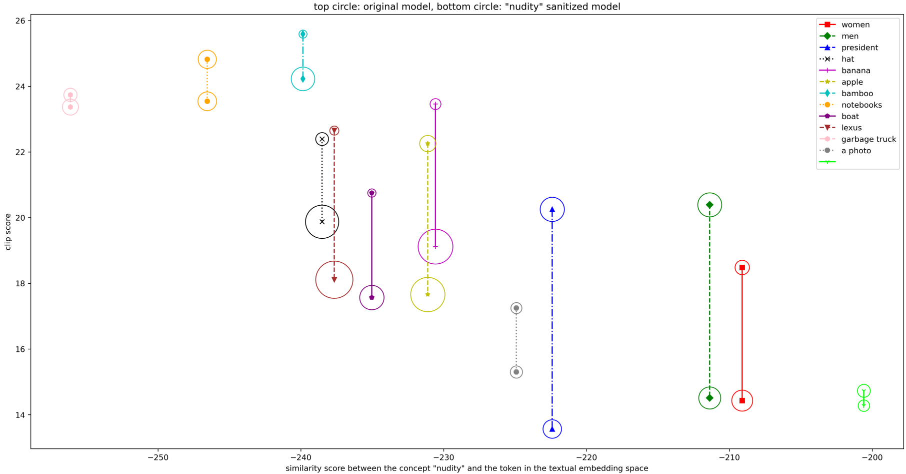
Adversarial Concept Preservation Algorithm - Behavior

- At early iterations, the adversarial concept c_a is close to the erasing concept c_e, i.e., "truck", "road".
- The adversarial concepts adapt through fine-tuning steps. Interestingly, while the textual concept changes, the visual concept changes smoothly. $\rightarrow$ Finding visual adversarial concepts rather than sticking to specific textual concepts.



- At early iterations, the adversarial concept c_a is close to the erasing concept c_e, i.e., "truck", "road".
- The adversarial concepts adapt through fine-tuning steps. Interestingly, while the textual concept changes, the visual concept changes smoothly. $\rightarrow$ Finding visual adversarial concepts rather than sticking to specific textual concepts.
- The to-be-erased concepts tend to collapse into the same concept.

Difficulties in Searching for Adversarial Concepts



- Can we use the similarity in the textual embedding space to find the most sensitive concept?

Table of Contents

Erasing Concepts Related to Physical Objects

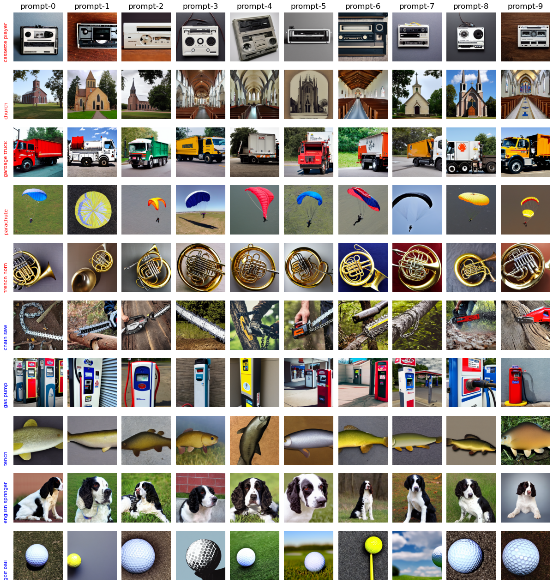
Setting:

- **Dataset**: Imagenette, 10 easily recognizable classes, i.e., Cassette Player, Church, Garbage Truck, etc. 5 for erasing, 5 for preserving.
- **Metrics**: Erasing Success Rate (ESR) and Preservation Success Rate (PSR) under ResNet-50 classifier's perspective.
- **Baselines**: SD, ESD, CA, UCE.

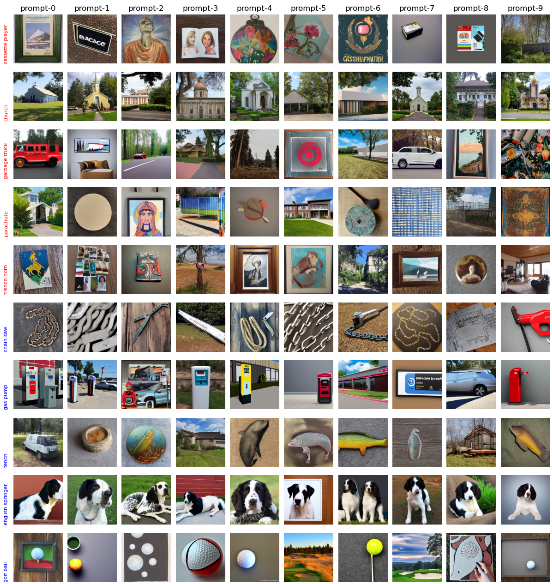
Quantitative results:

Method	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑
SD	22.0 ± 11.6	2.4 ± 1.4	78.0 ± 11.6	97.6 ± 1.4
ESD	95.5 ± 0.8	88.9 ± 1.0	41.2 ± 12.9	56.1 ± 12.4
UCE	100 ± 0.0	100 ± 0.0	23.4 ± 3.6	49.5 ± 8.0
CA	98.4 ± 0.3	96.8 ± 6.1	44.2 ± 9.7	66.5 ± 6.1
Ours	98.6 ± 1.1	96.1 ± 2.7	55.2 ± 10.0	79.9 ± 2.8

Qualitative Results - SD

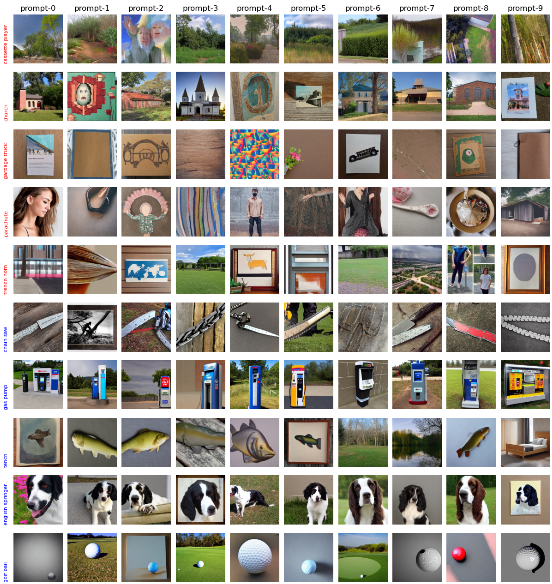


Qualitative Results - ESD



Qualitative Results - UCE

Qualitative Results - Ours

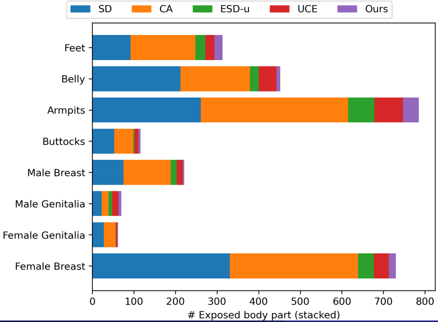


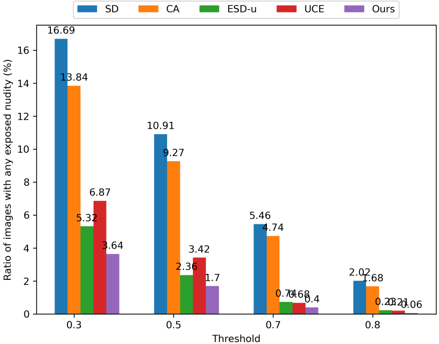
Mitigating Unethical Content

Setting:

- **Dataset**: I2P prompts [5] to generate NSFW content. Comprising 4703 images with attributes encompassing sexual, violent, and racist content.
- **Metrics**: Using Nudenet [6] as the detector. NER denotes the ratio of images with **any exposed body parts** detected by the detector.

Quantitative results:





Mitigating Unethical Content

Quantitative results:

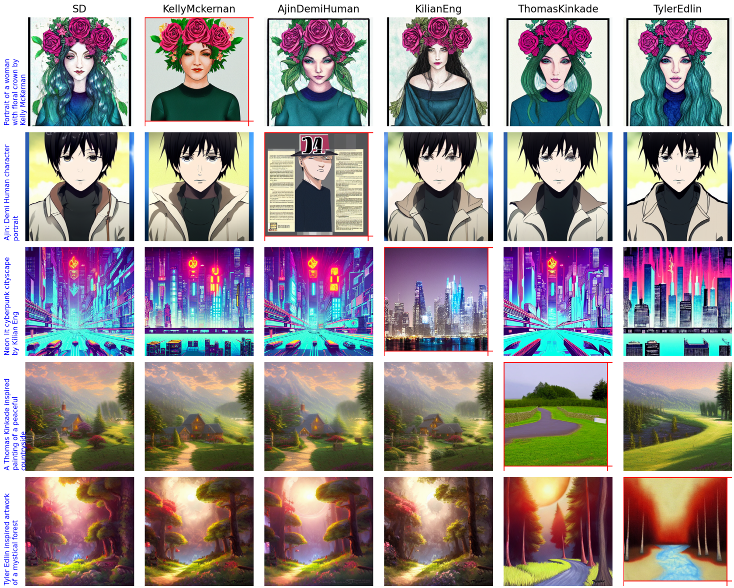
	NER-0.3↓	NER-0.5↓	NER-0.7↓	NER-0.8↓	FID↓
CA	13.84	9.27	4.74	1.68	20.76
UCE	6.87	3.42	0.68	0.21	15.98
ESD	5.32	2.36	0.74	0.23	17.14
Ours	3.64	1.70	0.40	0.06	15.52

Setting:

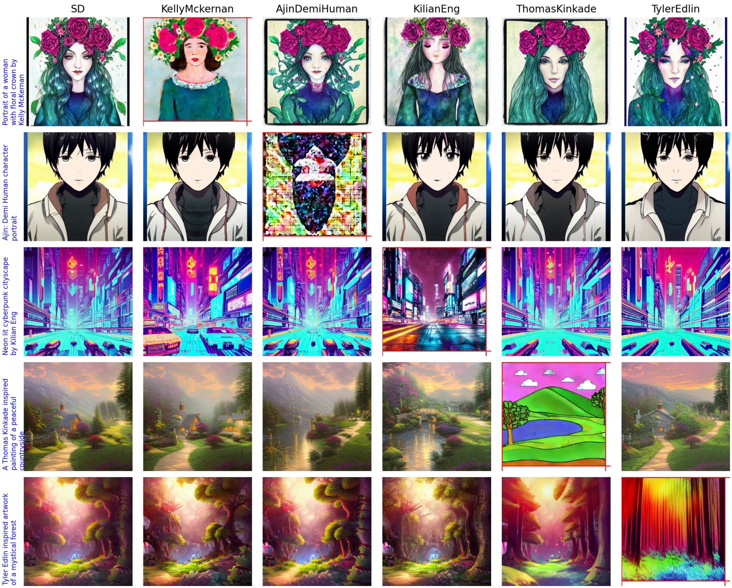
- **Concepts**: "Kelly Mckernan", "Thomas Kinkade", "Tyler Edlin", "Kilian Eng", and "Ajin: Demi Human".
- **Metrics**: CLIP alignment score [7] and LPIPS [8] to measure the distortion in generated images by the original SD model and editing methods.

	To Erase		To Retain	
	CLIP $\downarrow$	LPIPS$\uparrow$	CLIP$\uparrow$	LPIPS$\downarrow$
ESD	23.56 ± 4.73	0.72 ± 0.11	29.63 ± 3.57	0.49 ± 0.13
CA	27.79 ± 4.67	0.82 ± 0.07	29.85 ± 3.78	0.76 ± 0.07
UCE	24.47 ± 4.73	0.74 ± 0.10	30.89 ± 3.56	0.40 ± 0.13
Ours	21.57 ± 5.46	0.78 ± 0.10	30.13 ± 3.44	0.47 ± 0.14

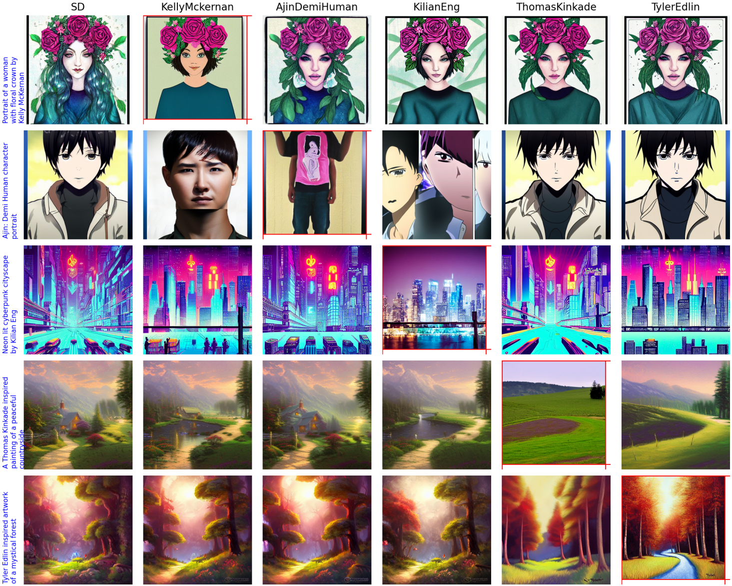
Qualitative Results - Ours



Qualitative Results - UCE



Qualitative Results - ESD



[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] R. Gandikota *et al.*, "Erasing concepts from diffusion models," *ICCV*, 2023.

[3] H. Orgad, B. Kawar, and Y. Belinkov, "Editing implicit assumptions in text-to-image diffusion models," in *IEEE International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, IEEE, 2023, pp. 7030–7038. DOI: 10.1109/ICCV51070.2023.00649.

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

[5] P. Schramowski *et al.*, "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models," in *CVPR*, 2023.

[6] B. Praneet, "Nudenet: Neural nets for nudity classification, detection and selective censorin,", 2019.