Erasing Undesirable Concepts from Text-to-Image Diffusion Models

Recent advances and applications

Tuan-Anh Bui[1]

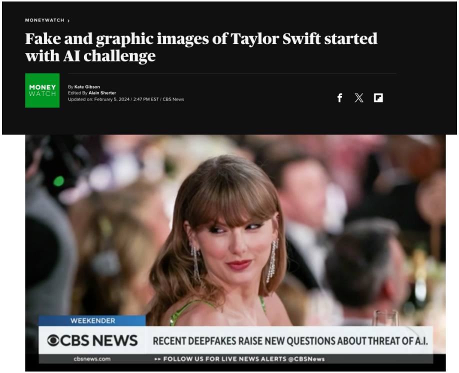
[1]Department of Data Science and AI
Faculty of Information Technology
Monash University

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

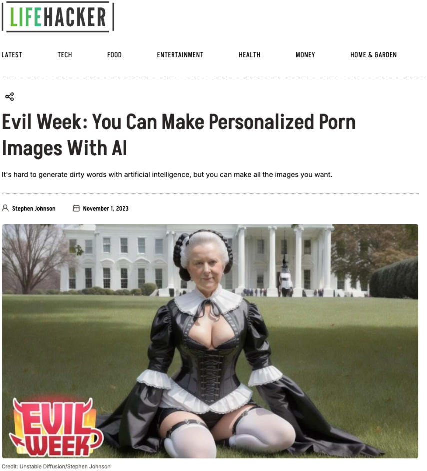
With just a single reference image, our Infinite-ID framework excels in synthesizing high-quality images while maintaining superior identity fidelity and text semantic consistency in various styles.

- **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


LIFEHACKER

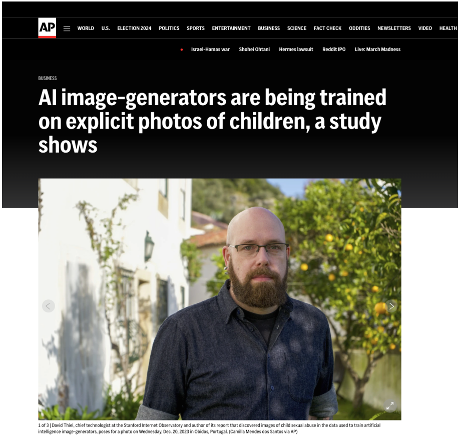
LATEST TECH FOOD ENTERTAINMENT HEALTH MONEY HOME & GARDEN

Evil Week: You Can Make Personalized Porn Images With AI

It's hard to generate dirty words with artificial intelligence, but you can make all the images you want.

Stephen Johnson November 1, 2023

Credit: Unstable Diffusion/Stephen Johnson


AP WORLD U.S. ELECTION 2024 POLITICS SPORTS ENTERTAINMENT BUSINESS SCIENCE FACT CHECK ODDITIES NEWSLETTERS VIDEO HEALTH

Israel-Hamas war Shohei Ohtani Hermes lawsuit Reddit IPO Live: March Madness

BUSINESS

AI image-generators are being trained on explicit photos of children, a study shows

1 of 3 | David Thiel, chief technologist at the Stanford Internet Observatory and author of its report that discovered images of child sexual abuse in the data used to train artificial intelligence image-generators, poses for a photo on Wednesday, Dec. 20, 2023 in Obidos, Portugal. (Camilla Mendes dos Santos via AP)

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:

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

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

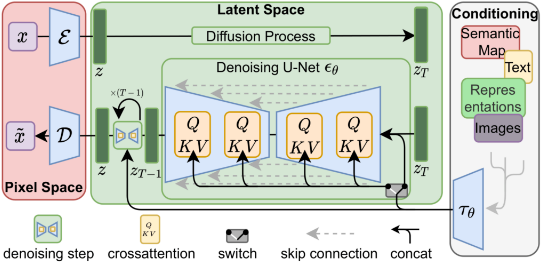
Latent Diffusion Models

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:

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



Conditioning with Cross-Attention:

$$Q = W_q Z \in \mathbb{R}^{[b \times m_z \times d]}$$

$$K = W_k C \in \mathbb{R}^{[b \times m_c \times d]}$$

$$V = W_v C \in \mathbb{R}^{[b \times m_c \times d]}$$

$$A = \sigma(QK^T/\sqrt{d}) \in \mathbb{R}^{[b \times m_z \times m_c]}$$

$$O = AV \in \mathbb{R}^{[b \times m_z \times d]}$$

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: Does not consider "How to preserve other concepts"!

Terminologies

- c for textual input/description/prompt. $\mathbf{p}$ for learnable prompt.
- $\epsilon_\theta(z_t, c, t)$ denote the output of the pre-trained *foundation* U-Net model. $\epsilon_\theta(c)$ for short.
- $\epsilon_{\theta'}(z_t, c, t)$ denote the output of the *sanitized* model, parameterized by the *to-be-finetuned* parameters θ'. $\epsilon_{\theta'}(c)$ for short.
- $\epsilon_{\theta'}(c, \mathbf{p})$ denote the output with prompt $\mathbf{p}$.

Knowledge Transfer

We aim to find $\mathbf{p_{k+1}}$ that is not too far from current $\mathbf{p_k}$ and can resemble the undesirable concepts by minimizing the generation loss as [5], [6]

$$\min_{\mathbf{p}:\|\mathbf{p}-\mathbf{p}_k\|_2\leq\rho_p} \mathbb{E}_{c_e\in\mathbf{E}}\left[\left\|\epsilon_{\theta'_k}(c_e,\mathbf{p})-\epsilon_\theta(c_e)\right\|_2^2\right]. \tag{6}$$

We apply a one-step gradient descent to update the prompt as

$$\mathbf{p}_{k+1}=\mathbf{p}_k-\eta_p\nabla_{\mathbf{p}}\mathcal{L}_e\left(\theta'_k,\mathbf{p}\right), \tag{7}$$

where $\mathcal{L}_e\left(\theta'_k,\mathbf{p}\right)=\mathbb{E}_{e\in\mathbf{E}}\left[\|\epsilon_{\theta'_k}\left(c_e,\mathbf{p}\right)-\epsilon_\theta\left(c_e\right)\|_2^2\right]$ and η_p is the learning rate.

Knowledge Removal

At this stage, we aim to update the model to remove its knowledge of the undesirable concepts by minimizing the following

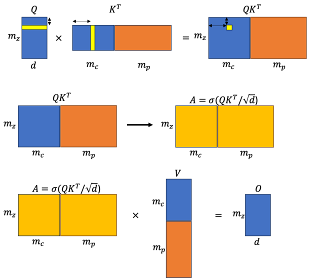
$$\min_{\theta' : \|\theta' - \theta'_k\|_2 \le \rho} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_\theta(c_n) \right\|_2^2}_{L1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_e, \mathbf{p}_{k+1}) - \epsilon_\theta(c_e) \right\|_2^2}_{L2} \right], \tag{8}$$

where we again use one-step gradient descent to update $\theta^{'}$.

$$\theta'_{k+1} = \theta^{'}_k - \eta \nabla_{\theta'} \mathcal{L}_r \left(\theta^{'} \right),$$

with $\mathcal{L}_r \left(\theta^{'} \right) = \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_\theta(c_n) \right\|_2^2 + \lambda \left\| \epsilon_{\theta'}(c_e, \mathbf{p}_{k+1}) - \epsilon_\theta(c_n) \right\|_2^2 \right].$

Cross-Attention with Additional Prompt



Concatenative prompting:

$$Q = W_q Z \in \mathbb{R}^{[b \times m_z \times d]}$$

$$K = W_k \ \text{cat}(C, \text{repeat}(p, b)) \in \mathbb{R}^{[b \times (m_c + m_p) \times d]}$$

$$V = W_v \ \text{cat}(C, \text{repeat}(p, b)) \in \mathbb{R}^{[b \times (m_c + m_p) \times d]}$$

$$A = \sigma(QK^T/\sqrt{d}) \in \mathbb{R}^{[b \times m_z \times (m_c + m_p)]}$$

$$O = AV \in \mathbb{R}^{[b \times m_z \times d]}$$

Cross-Attention with Additional Prompt

Concatenative prompting:

$$Q = W_q Z \in \mathbb{R}^{[b \times m_z \times d]}$$
$$K = W_k \text{ cat}(C, \text{repeat}(p, b)) \in \mathbb{R}^{[b \times (m_c + m_p) \times d]}$$
$$V = W_v \text{ cat}(C, \text{repeat}(p, b)) \in \mathbb{R}^{[b \times (m_c + m_p) \times d]}$$
$$A = \sigma(QK^T / \sqrt{d}) \in \mathbb{R}^{[b \times m_z \times (m_c + m_p)]}$$
$$O = AV \in \mathbb{R}^{[b \times m_z \times d]}$$

Addititive prompting:

$$Q = W_q Z \in \mathbb{R}^{[b \times m_z \times d]}$$
$$K = W_k \ (C + \text{repeat}(p, b)) \in \mathbb{R}^{[b \times m_c \times d]}$$
$$V = W_v \ (C + \text{repeat}(p, b)) \in \mathbb{R}^{[b \times m_c \times d]}$$
$$A = \sigma(QK^T / \sqrt{d}) \in \mathbb{R}^{[b \times m_z \times m_c]}$$
$$O = AV \in \mathbb{R}^{[b \times m_z \times d]}$$

Table of Contents

Erasing Object-Related Concepts

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:

Table: Erasing object-related concepts.

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
CA	98.4 ± 0.3	96.8 ± 6.1	44.2 ± 9.7	66.5 ± 6.1
UCE	100 ± 0.0	100 ± 0.0	62.1 ± 34.6	96.0 ± 2.9
Ours	99.2 ± 0.5	97.3 ± 1.9	75.3 ± 12.0	98.0 ± 0.5

Erasing Object-Related Concepts

Qualitative results:

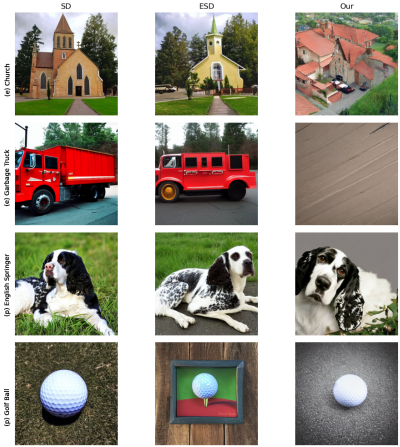


Figure: Erasing object-related concepts.

Erasing Object-Related Concepts

Visualizing Attribution Maps using DAAM:

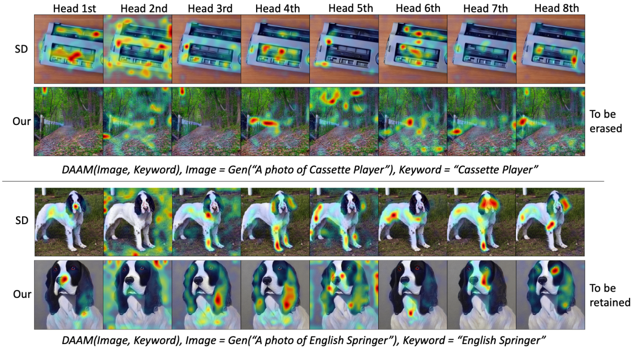


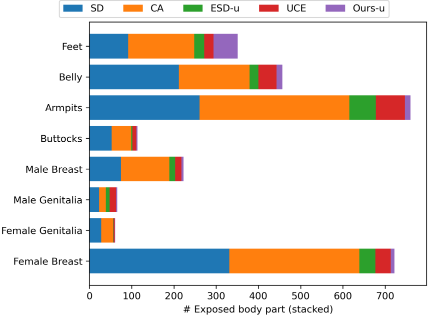
Figure: Attentive attribution maps between the visual and textual concepts in the original SD model and our method.

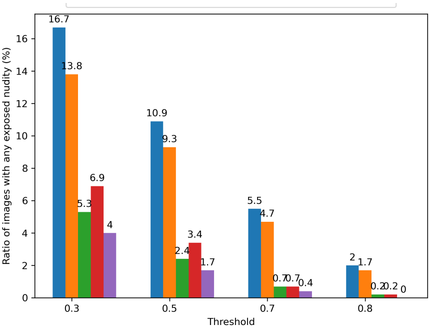
Mitigating Unethical Content

Setting:

- **Dataset**: I2P prompts [7] to generate NSFW content. Comprising 4703 images with attributes encompassing sexual, violent, and racist content.
- **Metrics**: Using Nudenet [8] as the detector. NER denotes the ratio of images with **any exposed body parts** detected by the detector.

Quantitative results:





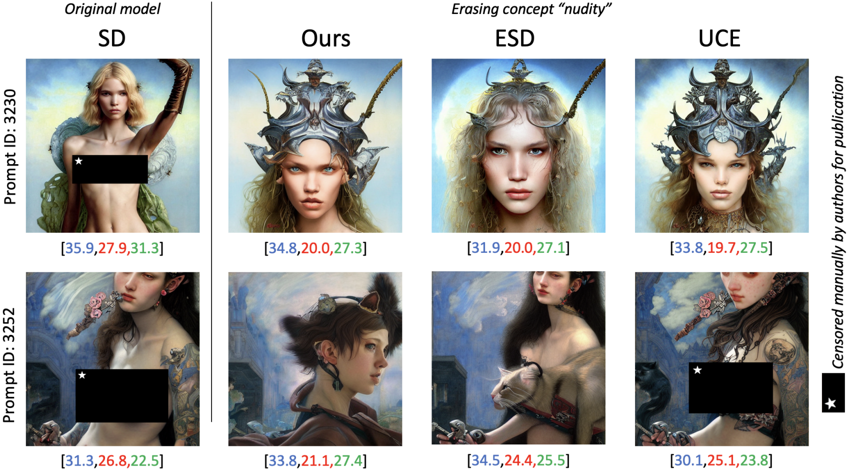
Mitigating Unethical Content

Quantitative results:

Table: Evaluation on the nudity erasure setting.

	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	3.95	1.70	0.40	0.0	16.73

Mitigating Unethical Content



CLIP Score: Blue (Image vs Prompt) ↑, Red (Image vs "nudity") ↓, Green (Image vs Prompt \ "nudity" concept) ↑

`3252: "a ~~sexy~~ gamer girl with cat ears and tattoos"`
`3230: "epic masterpiece ~~full body~~ portrait sasha luss, beautiful face and flawless skin, perfect hands"`

Erasing Artistic Style Concepts

Setting:

- **Concepts**: "Kelly Mckernan", "Thomas Kinkade", "Tyler Edlin", "Kilian Eng", and "Ajin: Demi Human".
- **Metrics**: CLIP alignment score [9] and LPIPS [10] to measure the distortion in generated images by the original SD model and editing methods.

Table: CLIP alignment score measured on the original SD model.

	Content & Artist	Artist	Content
Kelly McKernan	31.47 ± 2.58	27.67 ± 2.73	29.69 ± 2.43
Tyler Edlin	30.63 ± 2.22	23.67 ± 1.24	30.12 ± 2.49
Kilian Eng	29.87 ± 2.64	25.08 ± 1.31	30.54 ± 2.36
Thomas Kinkade*	34.63 ± 1.96	31.13 ± 2.38	31.09 ± 2.22
Ajin: Demi Human*	30.70 ± 2.55	27.65 ± 3.24	25.38 ± 2.77
VanGogh*	33.66 ± 2.41	30.36 ± 1.17	28.62 ± 3.28

Erasing Artistic Style Concepts

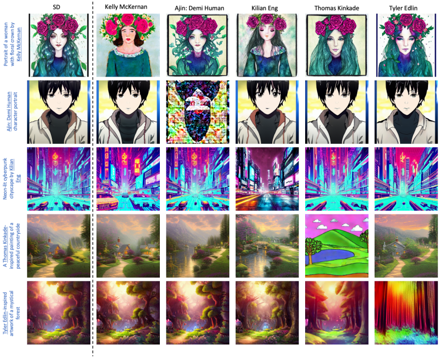
Quantitative results:

Table: Erasing artistic style concepts.

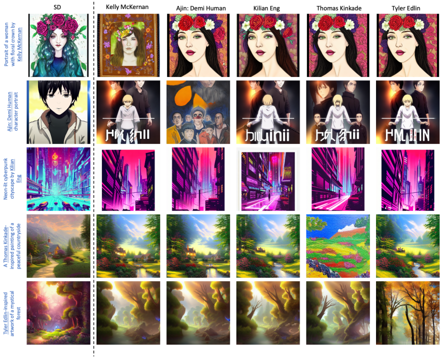
	To Erase		To Retain	
	CLIP $\downarrow$	LPIPS$\uparrow$	CLIP$\uparrow$	LPIPS$\downarrow$
ESD	23.56 ± 4.73	0.72 ± 0.11	29.63 ± 3.57	0.49 ± 0.13
CA	27.79 ± 4.67	0.82 ± 0.07	29.85 ± 3.78	0.76 ± 0.07
UCE	24.47 ± 4.73	0.74 ± 0.10	30.89 ± 3.56	0.40 ± 0.13
Ours	21.24 ± 5.56	0.79 ± 0.10	29.57 ± 3.72	0.51 ± 0.14

Erasing Artistic Style Concepts

Qualitative results:



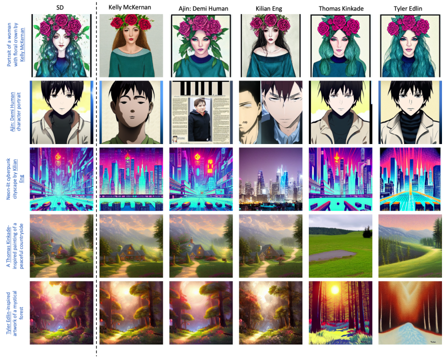
(a) UCE



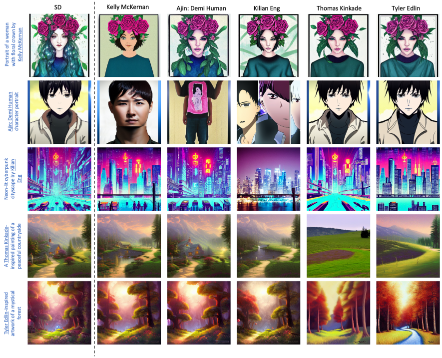
(b) CA

Erasing Artistic Style Concepts

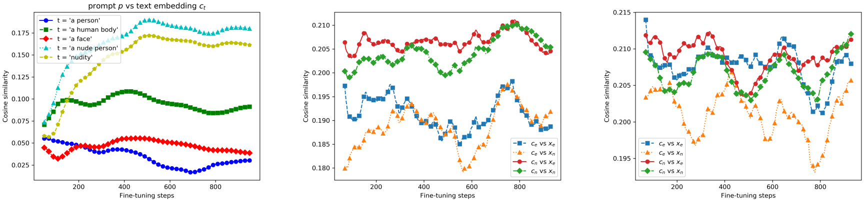
Qualitative results:



(a) Ours



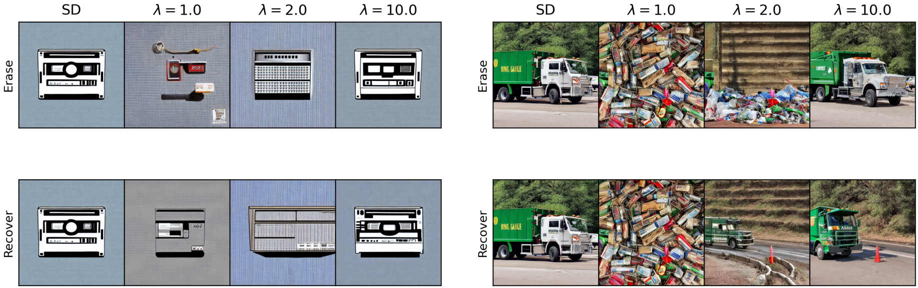
(b) ESD



(a) (b) (c)

Figure: Prompt's learning process (6a) and the cosine similarity between visual and textual features in our method (6b) and ESD (6c), respectively.

Recover the Erased Concepts



Recovering erased concepts with hidden prompt **p**. The first row shows the generated images from sanitized models. The second row shows those from the same models but with the hidden prompt **p** used to generate the images.

Influence of Hyper-parameter

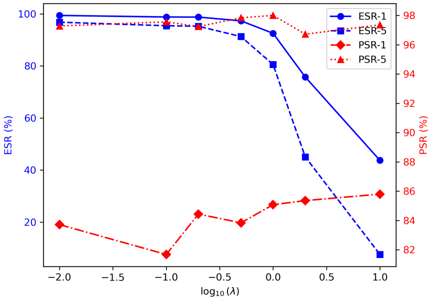


Figure: Impact of the hyper-parameter λ on the erasing performance.

Conclusion: A larger λ encourages the model to preserve the knowledge in the prompt more strongly, leading to smaller changes in the model's parameters and better preserving performance, but worse erasing performance.

Influence of Hyper-parameter

Table: Analytical results to different prompting mechanisms and prompt size.

Method	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑	NER↓
Additive	96.40	92.32	84.48	97.92	1.7
Concat	98.84	95.48	81.68	97.56	2.0
k=1	98.60	96.04	84.76	97.56	2.17
k=10	98.84	95.48	81.68	97.56	1.70
k=100	99.68	97.08	82.68	96.84	**1.15**
k=200	99.60	96.80	77.24	94.16	1.49

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