Erasing Undesirable Concepts in Diffusion Models

Tuan-Anh (Tony) Bui

Department of Data Science and AI
Faculty of Information Technology
Monash University

Table of Contents

Introduction

Dr. Tuan-Anh (Tony) Bui, Research Fellow at Monash University.
Extensive experience in Generative AI and Trustworthy Machine Learning.

- Worked on GAN since 2017 at the Singapore University of Technology and Design (SUTD).
- Worked on Adversarial Machine Learning since 2019 at Monash University.
- Currently working at the intersection of Generative Models and Trustworthy ML (TML) since 2023.
  - Concept Unlearning/Editing in Foundation Models (NeurIPS 2024, ICLR 2025; ICLR 2025 Workshop).
  - Personalization and Anti-Personalization in Generative Models (Chief Investigator - Monash and Department of Defence collaboration recently funded with $800K).

Find me on:

- LinkedIn: https://www.linkedin.com/in/tuananhbui89/
- Website: https://tuananhbui89.github.io/
- Email: tuananh.bui@monash.edu

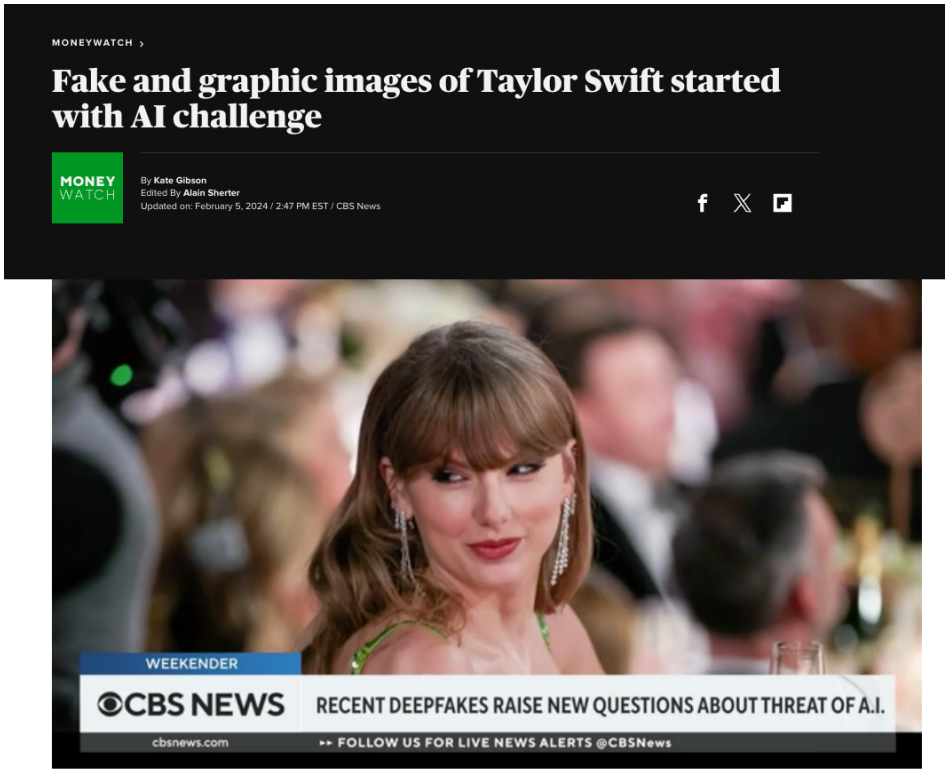
Why we need to erase concepts?

The current state of Generative AI

Everyone can generate high-quality content thanks to open-source Generative Models like Stable Diffusion.[1]

[1] Images from stability.ai

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

AFP

Home Services About us Crimes Jobs News Centre Police Checks Search this site

31 JULY 2024, 6:58PM

Media Release

Victorian man jailed for producing almost 800 AI-generated child abuse images

A Melbourne man has been sentenced to 13 months' imprisonment for online child abuse offences, including using an artificial intelligence program to produce child abuse images.

The man was sentenced by Melbourne County Court on 25 July, 2024, after pleading guilty to two offences.

The Victorian Joint Anti Child Exploitation Team (VIC JACET), which comprises members from the AFP and Victoria Police, linked the man to an online user engaging in sexualised conversations about children and transmitting child abuse material.

- A Melbourne man has been sentenced to 13 months' imprisonment for using AI to generate child abuse images.

Prevent misuse of AI-generated content

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

Naive Approaches

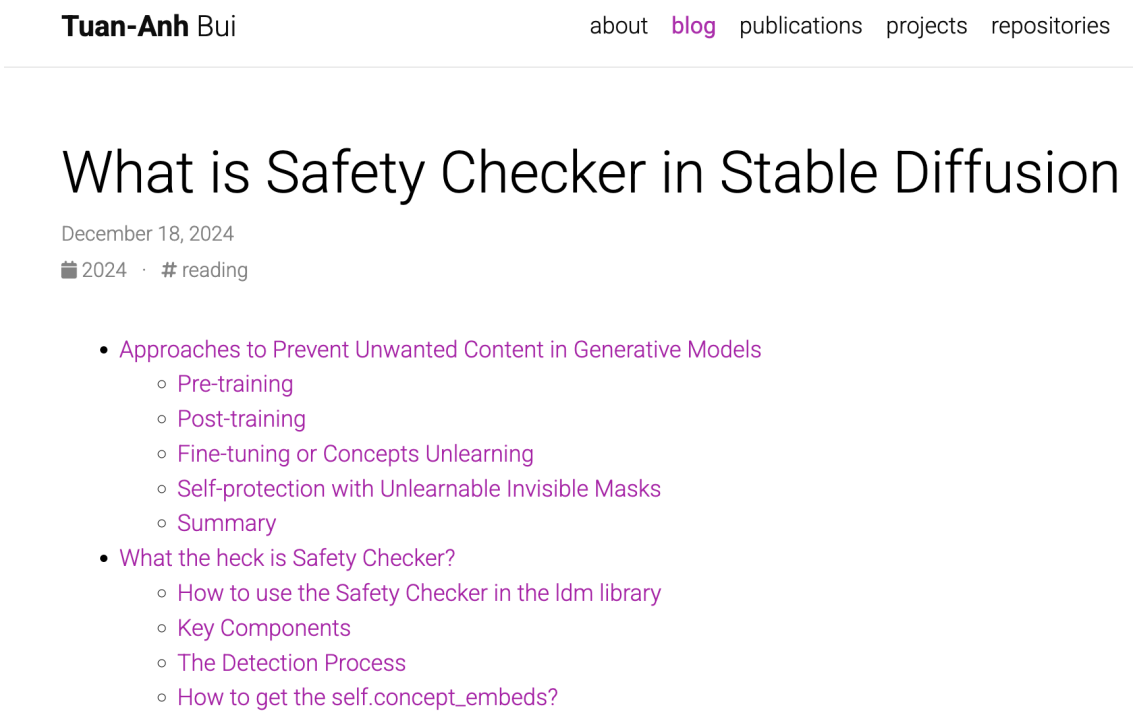
Pre-processing: Filtering all undesirable concepts from the training set and retraining the model from scratch.

But: Retraining is expensive! Queries come continuously!

Naive Approaches

✗Pre-processing: Filtering all undesirable concepts from the training set and retraining the model from scratch.

Post-processing: Detecting and censoring undesirable concepts in the generated images.



Tuan-Anh Bui about blog publications projects repositories

What is Safety Checker in Stable Diffusion

December 18, 2024

2024 · # reading

- Approaches to Prevent Unwanted Content in Generative Models
  - Pre-training
  - Post-training
  - Fine-tuning or Concepts Unlearning
  - Self-protection with Unlearnable Invisible Masks
  - Summary
- What the heck is Safety Checker?
  - How to use the Safety Checker in the ldm library
  - Key Components
  - The Detection Process
  - How to get the self.concept_embeds?

Naive Approaches

✗Pre-processing: Filtering all undesirable concepts from the training set and retraining the model from scratch.

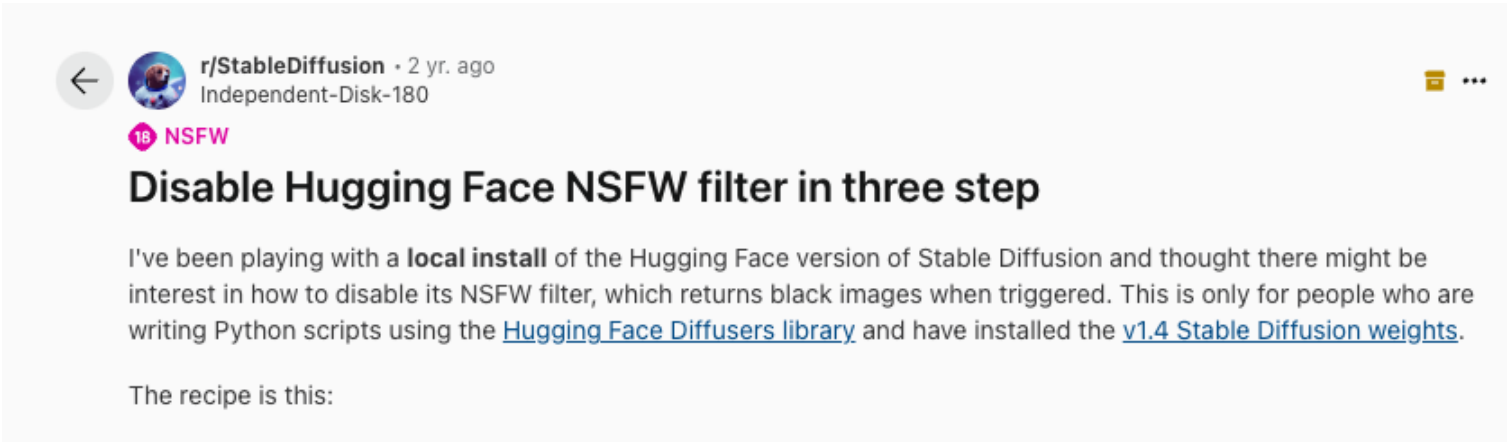
Post-processing: Detecting and censoring undesirable concepts in the generated images.

Key Components

- CLIP Vision Model: Uses a pretrained CLIP vision encoder to extract features from images
- Visual Projection Layer: Projects the CLIP features into a specific embedding space
- Two sets of learned concept embeddings:
  - `concept_embeds`: 17 different concepts (presumably NSFW concepts)
  - `special_care_embeds`: 3 special concepts that require extra attention
  - These embeddings are preloaded and marked as non-trainable (`requires_grad=False`).
- Corresponding weights for both types of embeddings:
  - `concept_embeds_weights`: 17 weights for the 17 concepts
  - `special_care_embeds_weights`: 3 weights for the 3 special concepts
  - These weights determine how strictly the model filters specific concepts.

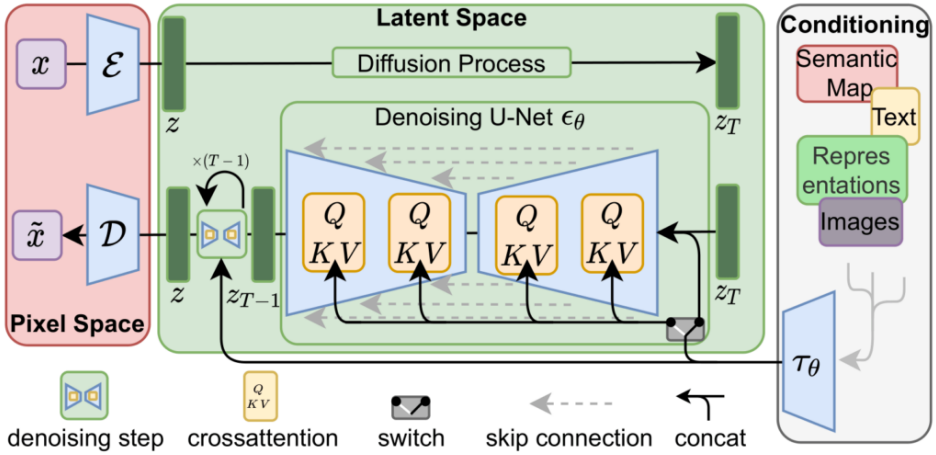
Naive Approaches

✗Pre-processing: Filtering all undesirable concepts from the training set and retraining the model from scratch.
Post-processing: Detecting and censoring undesirable concepts in the generated images.
But: Detectors can be bypassed easily in open-source models!



Naive Approaches

Erasing by fine-tuning the text encoder?



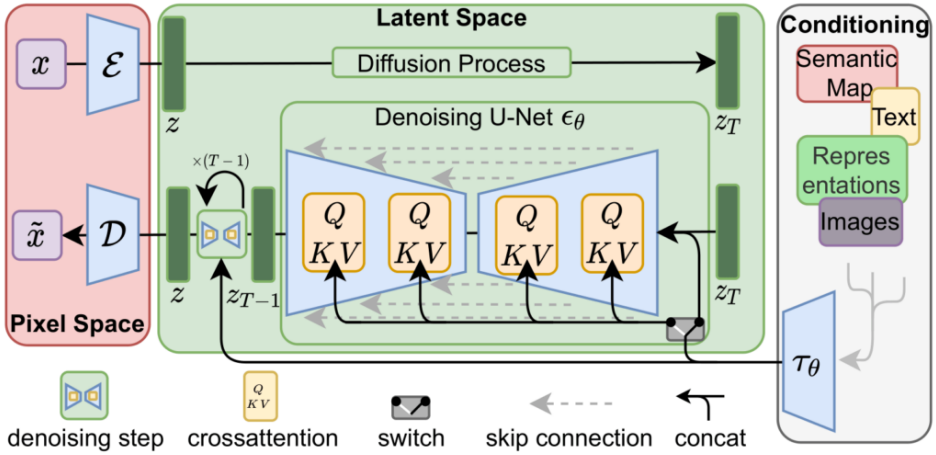
Cross-attention mechanism in Stable Diffusion:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

where $Q = z_t W_Q$, $K = \tau_\phi(c) W_K$, $V = \tau_\phi(c) W_V$, W_Q, W_K, W_V are the weights of the linear layers, and $\tau_\phi(c)$ is the pre-trained text encoder.

Naive Approaches

Erasing by fine-tuning the text encoder?



Cross-attention mechanism in Stable Diffusion:

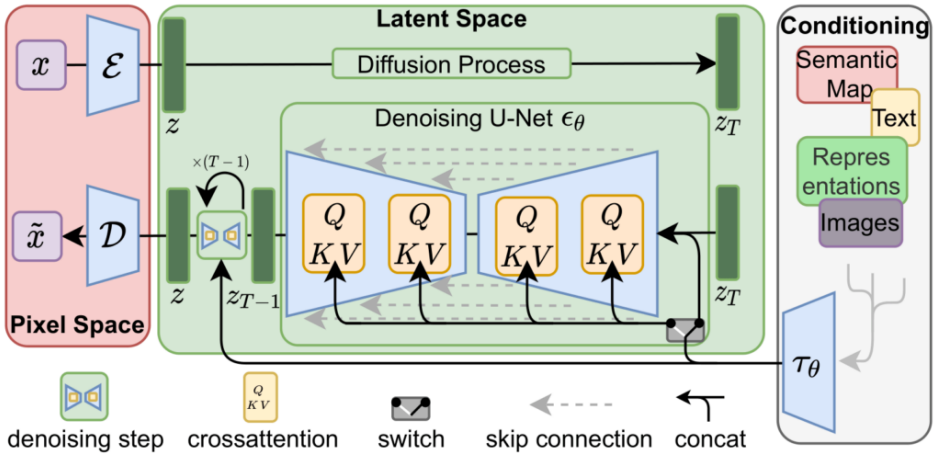
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{1}$$

where $Q = z_t W_Q$, $K = \tau_\phi(c) W_K$, $V = \tau_\phi(c) W_V$, W_Q, W_K, W_V are the weights of the linear layers, and $\tau_\phi(c)$ is the pre-trained text encoder.
Unlearning by mapping embedding of $c_e =$ "nudity" to the embedding of $c_t =$ "a person".

$$\phi^* = \arg\min_{\phi'} \left\| \tau_{\phi'}(c_e) - \tau_{\phi'}(c_t) \right\|_2^2 + \lambda \left\| \phi' - \phi \right\|_2^2 \tag{2}$$

Naive Approaches

✗Pre-processing: Filtering all undesirable concepts from the training set and retraining the model from scratch.
✗Post-processing: Detecting and censoring undesirable concepts in the generated images.
Erasing by fine-tuning the text encoder?
But: The text encoder τ_{ϕ^*} can be replaced by the original one τ_ϕ!



Naive Approaches

✗Pre-processing: Filtering all undesirable concepts from the training set and retraining the model from scratch.
✗Post-processing: Detecting and censoring undesirable concepts in the generated images.
✗Erasing by fine-tuning the text encoder?
✓Erasing by fine-tuning the U-Net model!

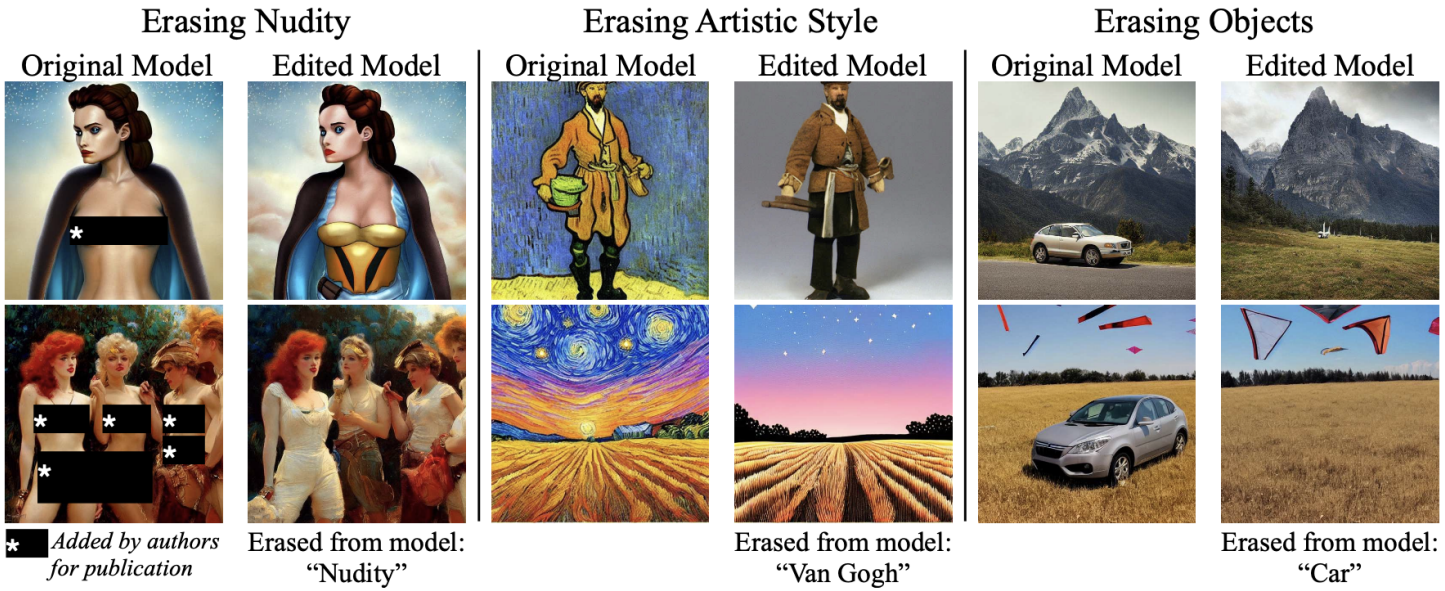


Figure 1: Given only a short text description of an undesired visual concept and no additional data, our method fine-tunes model weights to erase the targeted concept. Our method can avoid NSFW content, stop imitation of a specific artist's style, or even erase a whole object class from model output, while preserving the model's behavior and capabilities on other topics.

Two Approaches to Erasing

Two main approaches of erasing: Attention-based and Output-based.
Attention-based erasing [1], [2]: Change the projection of the text embedding to the latent space. The objective function can be Eq. 5 (TIME) or Eq. 6 (UCE):

$$\min_{W} \sum_{c_i \in E} \|Wc_i - v_i^*\|_2^2 + \lambda \|W - W_0\|_2^2 \tag{3}$$

$$\min_{W} \sum_{c_i \in E} \|Wc_i - v_i^*\|_2^2 + \sum_{c_j \in P} \|Wc_j - W_0 c_j\|_2^2 \tag{4}$$

Where E is the set of to-be-edited concepts, P is the set of preserved concepts, W_0 is the original projection matrix (W_0^K and W_0^V). $v_i^* = Wc_t$ is the target embedding of the concept c_t, which depends on the application.
Note of terminology: c is the text embedding of a concept, e.g., $c = \tau_\phi(\text{"nudity"})$, but we omit τ_ϕ for simplicity.

Two Approaches to Erasing

Attention-based erasing [1], [2]: Change the projection of the text embedding to the latent space. The objective function can be Eq. 5 (TIME) or Eq. 6 (UCE):

$$\min_{W} \sum_{c_i \in E} \| W c_i - v_i^* \|_2^2 + \lambda \| W - W_0 \|_2^2 \tag{5}$$

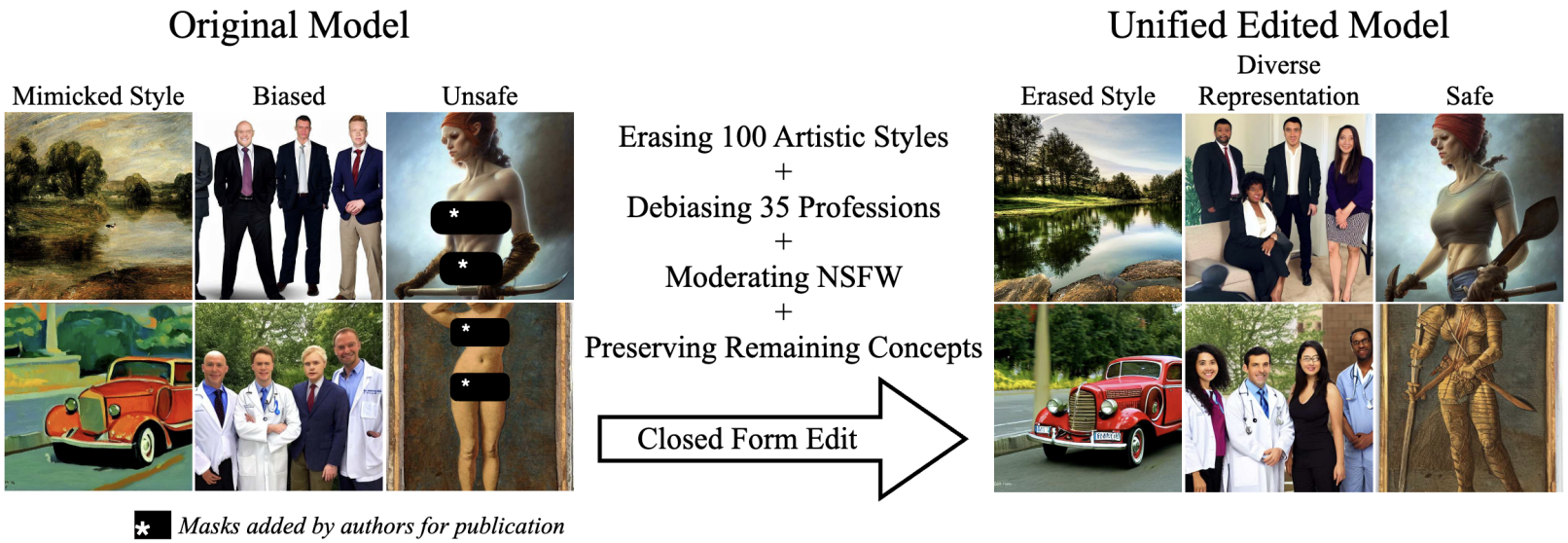
$$\min_{W} \sum_{c_i \in E} \| W c_i - v_i^* \|_2^2 + \sum_{c_j \in P} \| W c_j - W_0 c_j \|_2^2 \tag{6}$$

Where E is the set of to-be-edited concepts, P is the set of preserved concepts, W_0 is the original projection matrix (W_0^K and W_0^V). $v_i^* = W c_t$ is the target embedding of the concept c_t, which depends on the application.

Note of terminology: c is the text embedding of a concept, e.g., $c = \tau_\phi(\text{"nudity"})$, but we omit τ_ϕ for simplicity.

- **Erasing/Moderation**: $c_e = \text{"nudity"} \rightarrow c_t = \text{"a person"}$ or $c_e = \text{"Van Gogh"} \rightarrow c_t = \text{"art"}$.
- **Editing**: $c_e = \text{"Mercedes"} \rightarrow c_t = \text{"BMW"}$ or $c_e = \text{"car"} \rightarrow c_t = \text{"bicycle"}$.
- **Debiasing**: Choose $v^* = W(c_i + \sum_{t=1}^{p} \alpha_t a_t)$, where c_i is "doctor" and a_t is attributes that we want to distribute across such as "white", "asian", "black". By this way, the original concept "doctor" no longer only associated with "white" but also with "asian" and "black".

Two Approaches to Erasing



Pros and Cons:

- Pros: Fast and effective to erase/edit (multiple) concepts. Have a close-form solution.
- Cons: Strongly depends on the preservation set P. Requires some tricks to avoid matrix non-invertible.

Two Approaches to Erasing

Output-based erasing: Change the output of the diffusion model to remove/edit the concept. The most naive approach [3] is to optimize the following objective function:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2 \right] + \lambda \left\| \epsilon_{\theta'}(c_n) - \epsilon_{\theta}(c_n) \right\|_2^2 \tag{7}$$

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

Pros and Cons:

- Pros: Simple yet effective in erasing concepts. Working on the output space directly.
- Cons: Degradation in the quality of other concepts. Strongly depends on the choice of the preservation set P.

Two Approaches to Erasing

Output-based erasing: Change the output of the diffusion model to remove/edit the concept. The most naive approach [3] is to optimize the following objective function:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2 \right] + \lambda \left\| \epsilon_{\theta'}(c_n) - \epsilon_{\theta}(c_n) \right\|_2^2 \tag{7}$$

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

Pros and Cons:

- Pros: Simple yet effective in erasing concepts. Working on the output space directly.
- Cons: Degradation in the quality of other concepts. Strongly depends on the choice of the preservation set P.

Our NeurIPS 2024: Instead of preserving *neutral concepts*, can we preserve the *most sensitive concepts* to the erasing concept?

Adversarial Preservation

(NeurIPS 2024)

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?

Challenge: How to detect (robustly/reliably) whether or not a concept is in the generated image?

Solution: Measuring the CLIP alignment score between the generated image and the concept.

- 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?

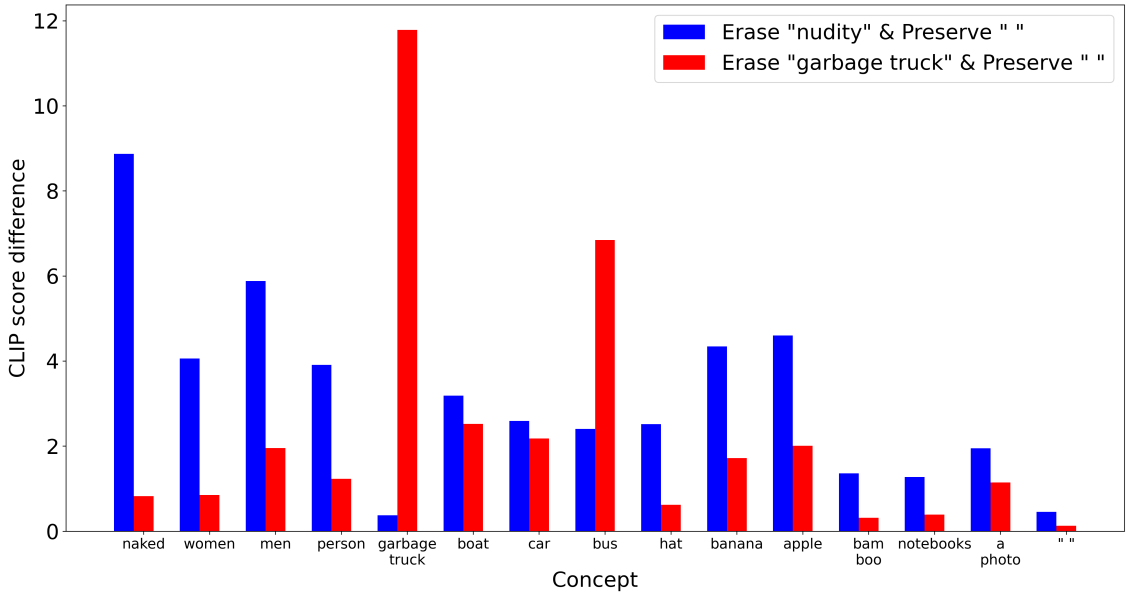
Measuring Generation Capability with CLIP Alignment Score:

- 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 θ' 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).

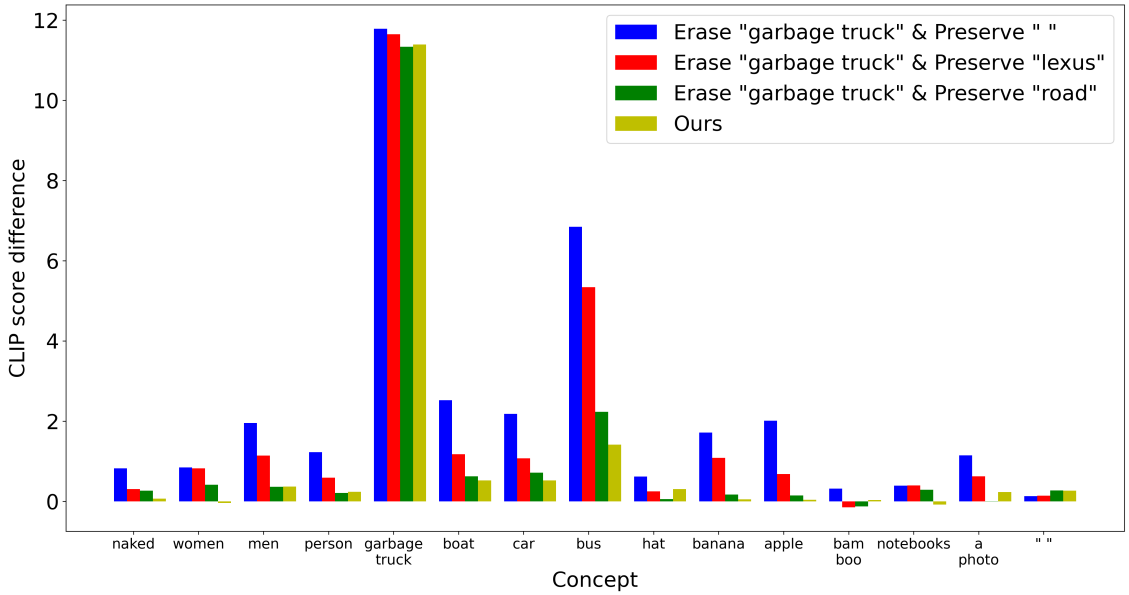
Impact on the model's capability: Results



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:

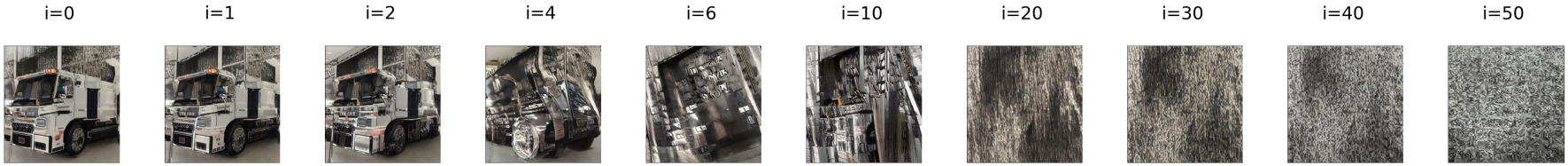
- Choosing the right concept to preserve is crucial.
- Preserving "road" $>$ "lexus" $>$ " " in maintaining the quality of other concepts.
- Early advertisment: our adaptive preservation is the best :D

Adversarial Preservation

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

- 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.

Adversarial Preservation: Solving with PGD



$$\min_{\theta'} \max_{c_a \in \mathcal{R}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_a) - \epsilon_{\theta}(c_a) \right\|_2^2}_{L_2} \right] \tag{9}$$

Solving the optimization problem with PGD:

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

Continuos concept space is not suitable for adversarial preservation.

Adversarial Preservation: Relaxation with Gumbel-Softmax

$$\min_{\theta'} \max_{\pi \in \Delta_{\mathcal{R}}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\|\epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t)\|_2^2}_{L_1} + \lambda \underbrace{\|\epsilon_{\theta'}(\mathbf{G}(\pi) \odot \mathcal{R}) - \epsilon_{\theta}(\mathbf{G}(\pi) \odot \mathcal{R})\|_2^2}_{L_2} \right] \tag{10}$$

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 Preservation: Relaxation with Gumbel-Softmax

Implementation of Gumbel-Softmax operator:

$$y_i = \exp(\frac{l_i + g_i}{\tau}) / \sum_{j=1}^{|\mathcal{R}|} \exp(\frac{l_j + g_j}{\tau}) \tag{11}$$

where g_i is the Gumbel noise and τ is the temperature. $l \in \mathbb{R}^{|\mathcal{R}|}$ is the logits of the Gumbel-Softmax distribution, and is a learnable parameter. If $\tau \to 0$, y_i becomes one-hot.
Hard version: represent the adversarial concept as the most likely concept, i.e.,

$$\hat{y} = \text{sg}((\text{one-hot}(\text{argmax}(y)) - y)) + y \tag{12}$$

where sg denotes the stop-gradient (detach) operation.
Soft version: represent the adversarial concept as a combination of multiple concepts, i.e., $\hat{y} = y$

Adversarial Concept Preservation Algorithm

Algorithm 1 Find Adversarial Concept

Input: $\theta, \mathcal{R}$. Searching hyperparameters: η, N_{iter}. Current state $\theta_k^{'}$
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

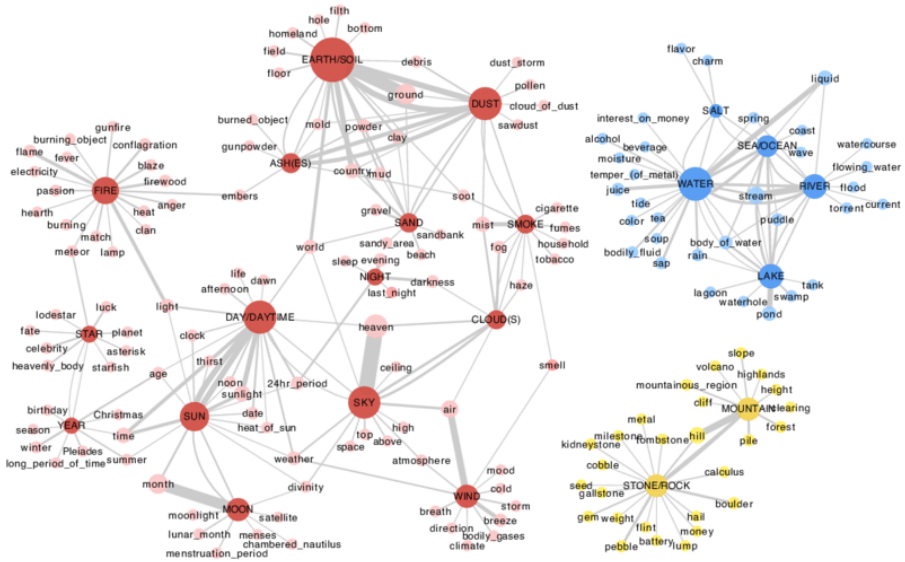
Input: $\theta, \mathcal{R}, \mathbf{E}, \lambda$. Searching hyperparameters: η, N_{iter}.
Output: $\theta^{'}$
$k \leftarrow 0, \theta_k^{'} \leftarrow \theta$
while Not Converged **do**
$c_e \sim \mathbf{E}$
$c_a \leftarrow \text{FindAdversarialConcept}(\theta_k^{'}, \theta, \mathcal{R}, \eta, N_{\text{iter}})$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

Adversarial Guide Erasure

(ICLR 2025)

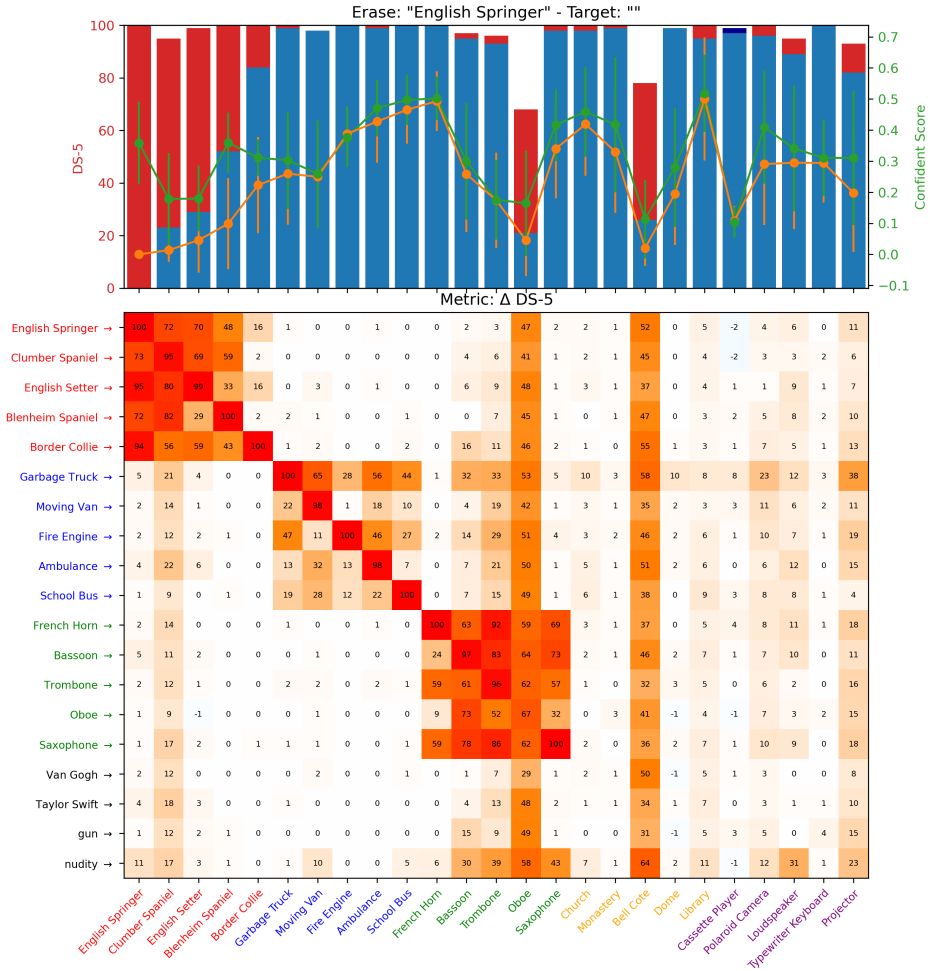
Motivation

Observation from our NeurIPS 2024 work: Erasing a concept c_e will affect the related concepts more than the unrelated ones.



Modeling the concept space as a graph where each node is a concept and each edge is the generation capability of one concept when erasing another concept.
It might have a geometric property, but it is not well understood.

The Concept Graph



- $G_{c_e}(c_i)$ is the generation capability of concept c_i when erasing concept c_e. $G_0(c_i)$ is that of the original model (red-bar + blue-bar).
- $\Delta_{c_e}(c_i) = G_{c_e}(c_i) - G_0(c_i)$ measures the impact of c_e on c_i.
- **Locality:** The concept graph is sparse and localized, i.e., $\Delta_{c_e}(c_i) \geq \Delta_{c_e}(c_j)$ if $d(c_i, c_e) \leq d(c_j, c_e)$.
- **Asymmetry:** The concept graph is asymmetric, i.e., $\Delta_{c_e}(c_i) \neq \Delta_{c_i}(c_e)$.
- **Abnormal:** If $G_0(c_i)$ low, then $\Delta_{c_e}(c_i)$ high for any c_e, e.g., "Bell Cote" and "Oboe"

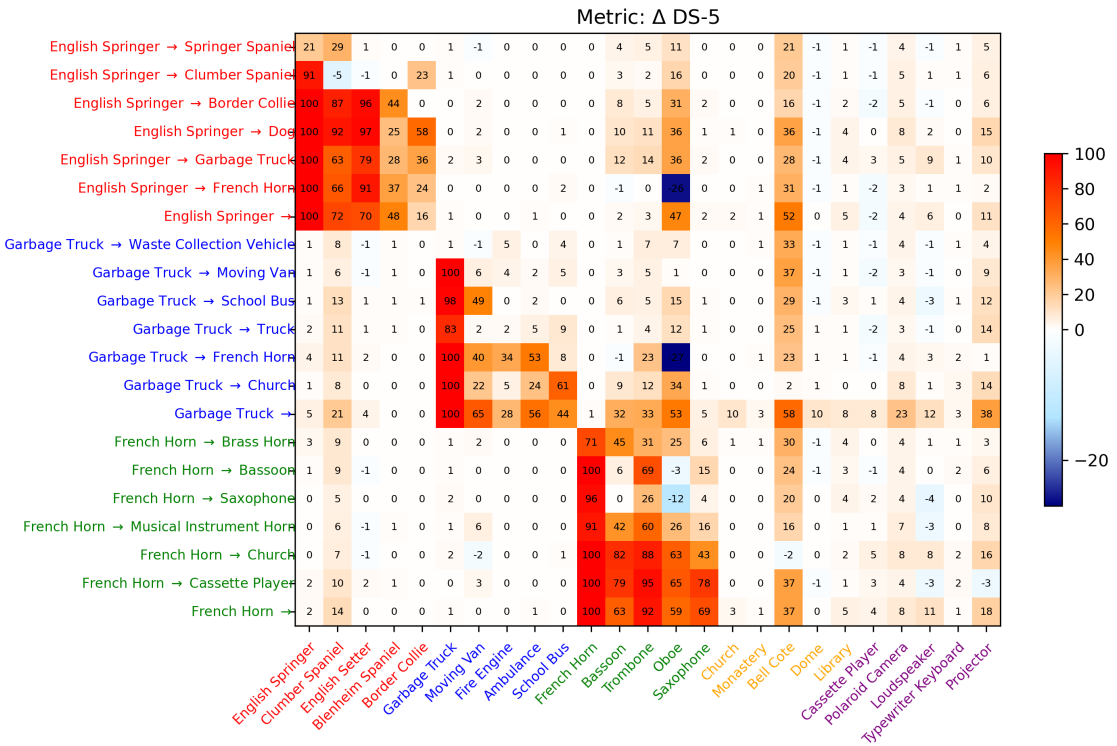
The Concept Graph

The concept graph has sparse and localized structure.
$\rightarrow$ Suggesting that we should map the erasing concept c_e to some specific target concepts c_t that are close to c_e rather than unrelated or neutral concepts.

Recall the objective function of the naive method:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2 \right] + \lambda \left\| \epsilon_{\theta'}(c_n) - \epsilon_{\theta}(c_n) \right\|_2^2 \tag{13}$$

The Fantastic Targets



- **Locality:** $\Delta_{c_e}(c_i) \geq \Delta_{c_e}(c_j)$ if $d(c_i, c_e) \leq d(c_j, c_e)$ regardless of target concept c_t.
- **Abnormal:** If $G_0(c_i)$ low, then $\Delta_{c_e}(c_i)$ high for any c_e, regardless of target concept c_t.
- **Synonym ✗✗✗:** Best at preserving, worst at erasing, i.e., $G_{c_e}(c_i) \approx G_0(c_i) \forall c_i$ including c_e.
- **Unrelated ✗✗:** Bad at preserving, i.e., $G_{c_e}(c_e) \approx 0$ but $G_{c_e}(c_i) < G_0(c_i) \forall c_i \neq c_e$.
- **General ✗:** moderate at erasing and preserving.
- **In-class ✓:** Good at both erasing and preserving, i.e., $G_{c_e}(c_e) \approx 0$ and $G_{c_e}(c_i) \approx G_0(c_i) \forall c_i \neq c_e$.

Adversarial Guide Erasure

The 'Fantastic Targets': Closely related to the erasing concept c_e and but not its synonyms.

Adversarial Guide Erasure

The 'Fantastic Targets': Closely related to the erasing concept c_e and but not its synonyms.

$$\min_{\theta'} \mathop{\mathbb{E}}_{c_e \in \mathbf{E}} \max_{c_t \in \mathcal{C}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_t) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_2} \right] \tag{14}$$

- Minimizing L_1 w.r.t. θ': Erasing the concept c_e.
- Minimizing L_2 w.r.t. θ': Preserving the target concept c_t.
- Maximizing L_2 w.r.t. c_t: Searching for the most sensitive concept to the erasing concept c_e.
- Maximizing L_1 w.r.t. c_t: Ensuring that $c_t \neq c_e$

Adversarial Guide Erasure - Behavior

Figure: Intermediate results of the search process, with images generated from the most sensitive concepts c_t found by our method and c_e at the same optimization step.

- At early iterations, the adversarial concept c_t is close to the erasing concept c_e, i.e., "model", "icon".
- The adversarial concepts adapt through fine-tuning steps.

Erasing Concepts Related to Physical Objects

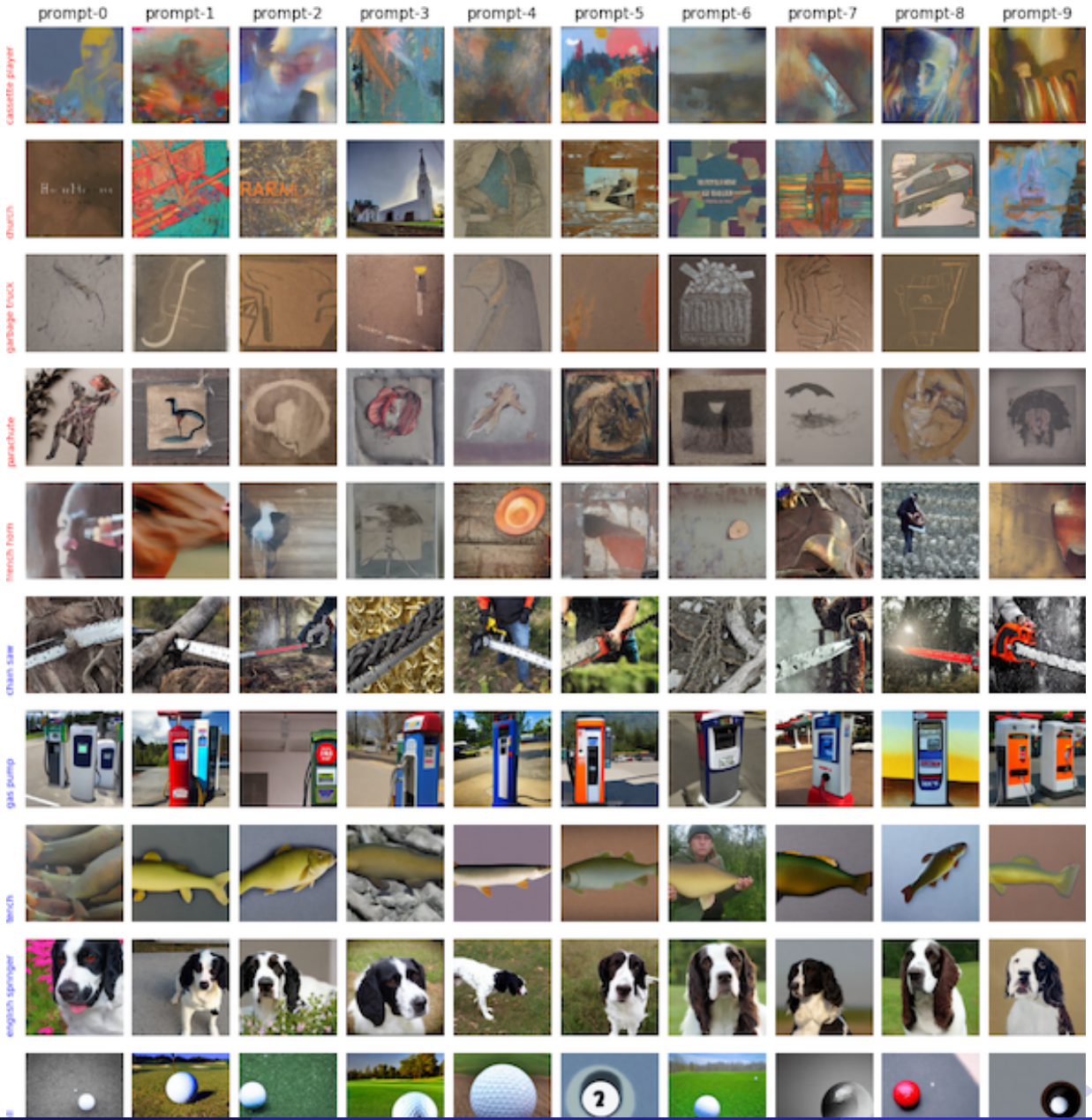
Setting:

- **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↑	FID↓	CLIP↑
SD	22.0 ± 11.6	2.4 ± 1.4	78.0 ± 11.6	97.6 ± 1.4	16.1	26.4
ESD	95.5 ± 0.8	88.9 ± 1.0	41.2 ± 12.9	56.1 ± 12.4	17.9	24.5
UCE	$\mathbf{100 \pm 0.0}$	$\mathbf{100 \pm 0.0}$	23.4 ± 3.6	49.5 ± 8.0	19.1	21.4
CA	98.4 ± 0.3	96.8 ± 6.1	44.2 ± 9.7	66.5 ± 6.1	16.6	25.8
MACE	$\underline{99.3 \pm 0.3}$	$\underline{97.6 \pm 1.2}$	47.4 ± 12.0	72.8 ± 10.5	16.9	24.9
Ours-AP	98.6 ± 1.1	96.1 ± 2.7	$\underline{55.2 \pm 10.0}$	$\underline{79.9 \pm 2.8}$	$\underline{16.3}$	$\underline{26.1}$
Ours-AGE	98.1 ± 1.1	95.7 ± 2.5	$\mathbf{73.6 \pm 9.8}$	$\mathbf{95.6 \pm 1.1}$	**16.1**	**26.0**

Erasing Concepts Related to Physical Objects

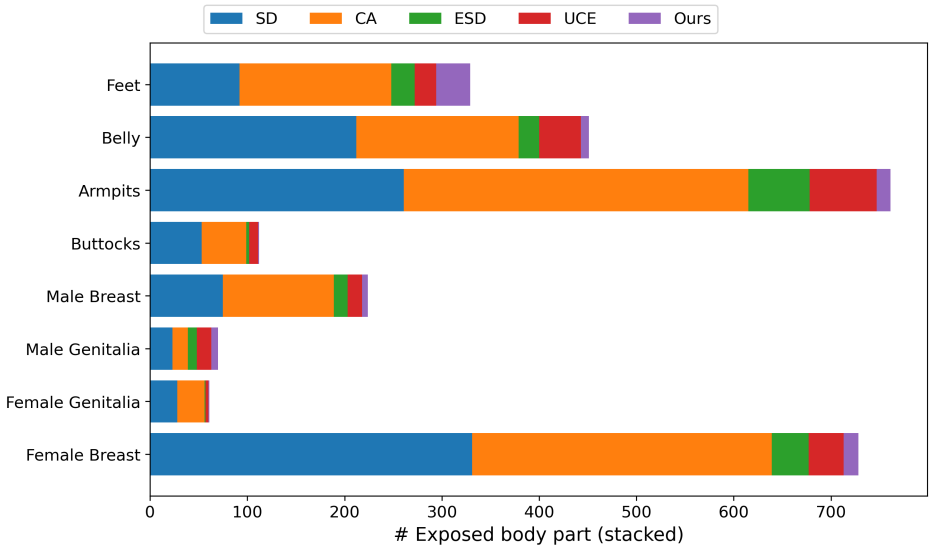


Mitigating Unethical Content

Setting:

- **Dataset**: I2P prompts [4] to generate NSFW content. Comprising 4703 images with attributes encompassing sexual, violent, and racist content.
- **Metrics**: Using Nudenet [5] as the detector. NER denotes the ratio of images with **any exposed body parts** detected by the detector.

Quantitative results:



Mitigating Unethical Content

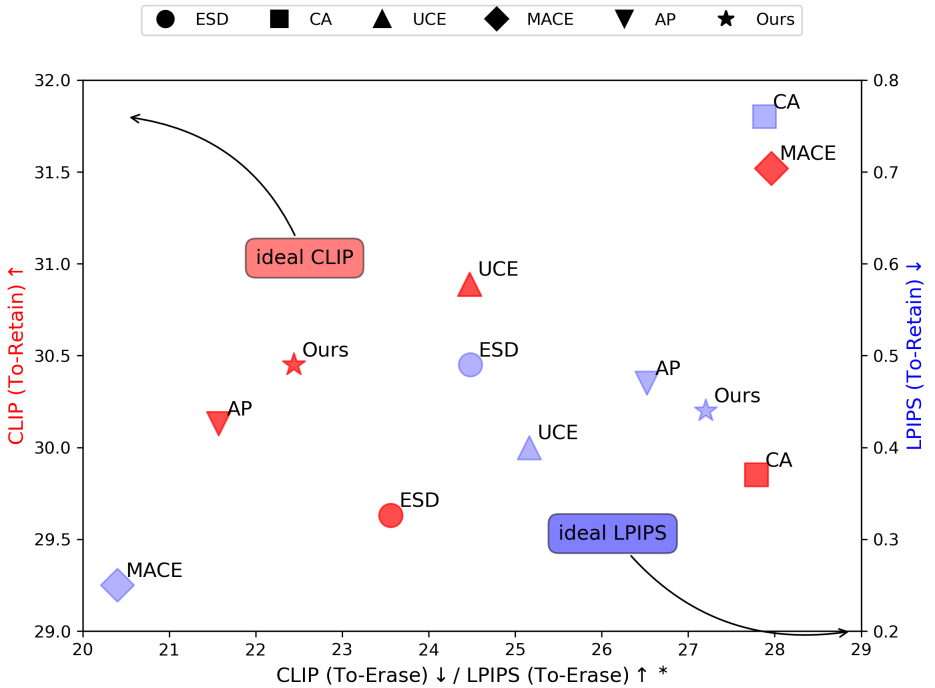
Quantitative results:

	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-AP	3.64	1.70	0.40	0.06	15.52
Ours-AGE	5.06	1.53	0.32	0.04	14.20

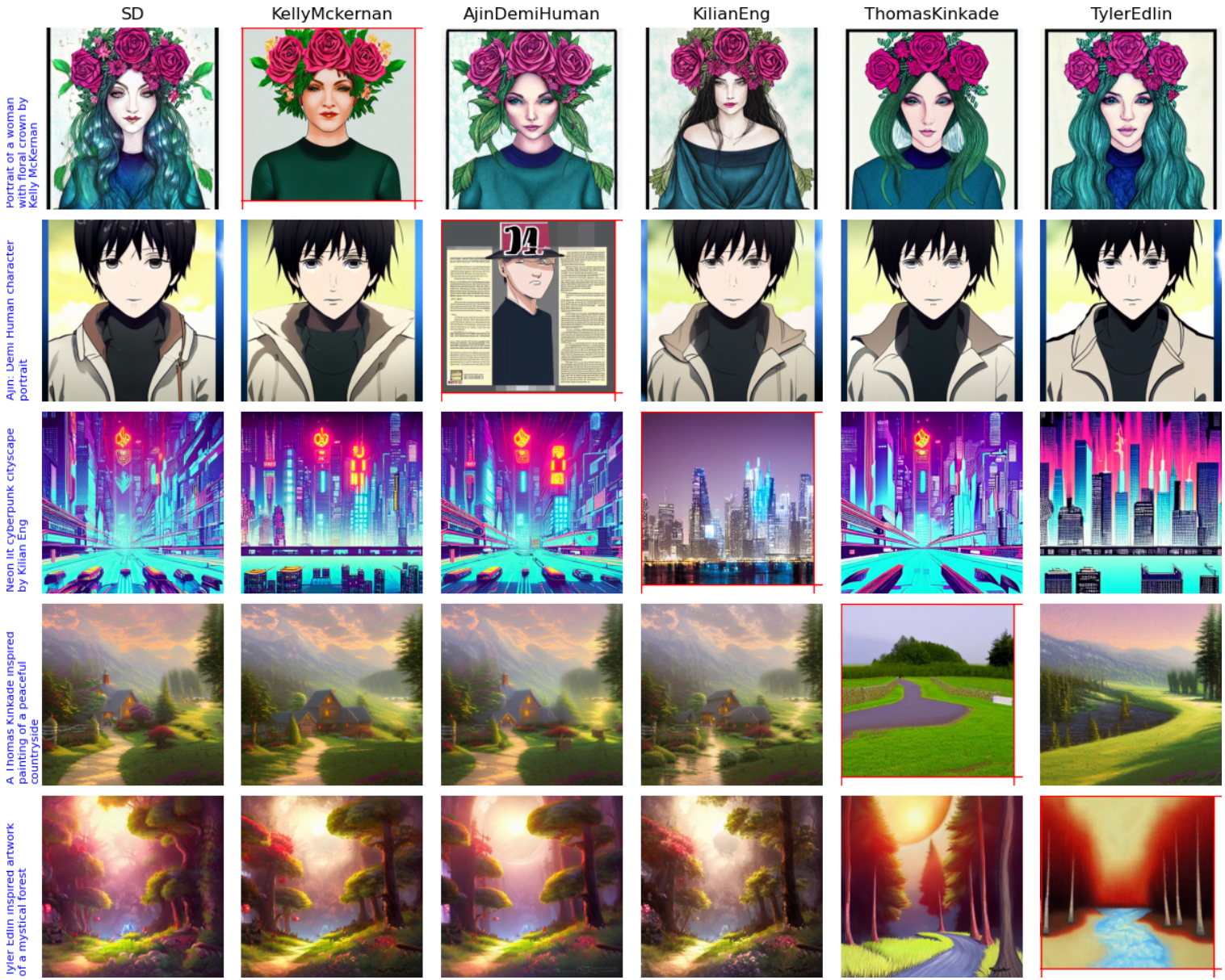
Erasing Artistic Concepts

Setting:

- **Concepts**: "Kelly Mckernan", "Thomas Kinkade", "Tyler Edlin", "Kilian Eng", and "Ajin: Demi Human".
- **Metrics**: CLIP alignment score [6] and LPIPS [7] to measure the distortion in generated images by the original SD model and editing methods.



Erasing Artistic Concepts



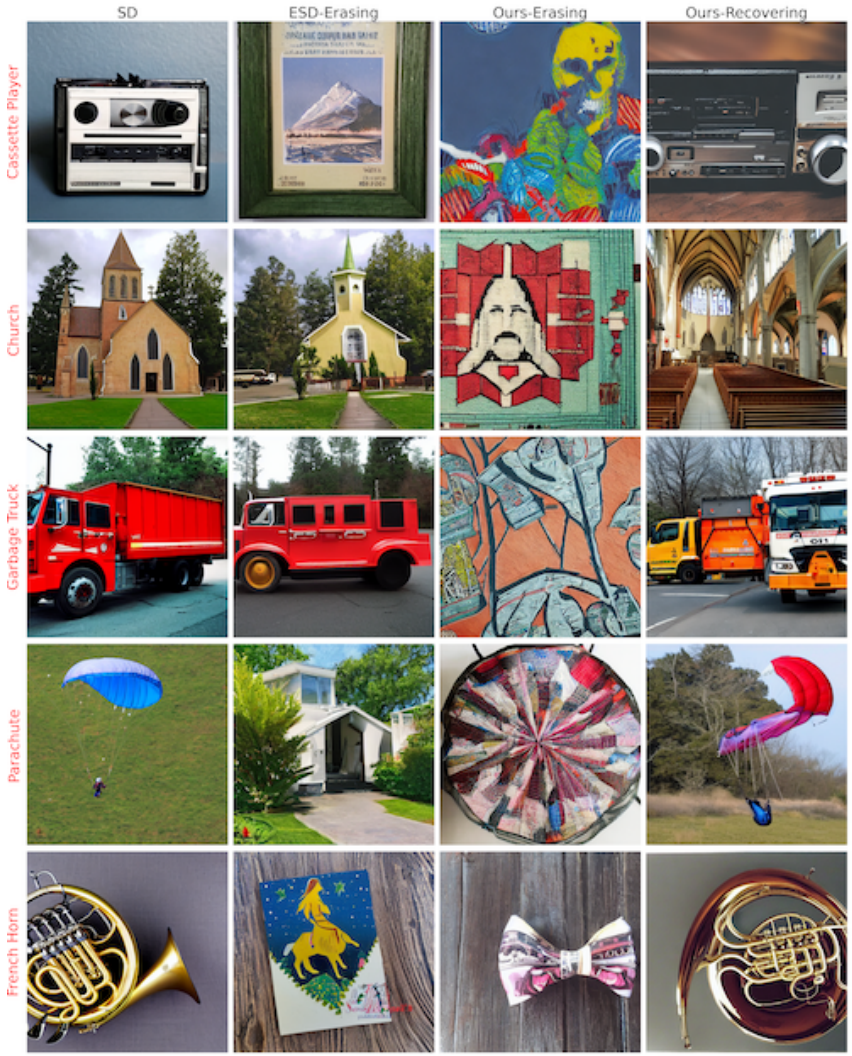
Hiding and Recovering Knowledge in Text-to-Image Diffusion Models via Learnable Prompts

(ICLR 2025 - Delta Workshop)

Motivation

Hiding concepts instead of permanently removing them:

- **Flexibility & Controlled Access**: Provide more flexibility, especially useful when regulations evolve, controlled access is required or different levels of access can be granted.
- **Enhancing Security Against Backdoor Attacks**: This method can be considered as an internal backdoor attack.



Our Approach - Knowledge Hiding and Recovery

We introduce an additional prompt $\mathbf{p}_{c_e}$ acts as an external memory to store the knowledge of undesirable concepts $c_e \in \mathbf{E}$.

Stage 1 - Knowledge Hiding: Hide the undesirable concepts $c_e \in \mathbf{E}$ from the model θ'.

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_e, \mathbf{p}_{c_e}) - \epsilon_{\theta}(c_e) \right\|_2^2}_{L2} \right] \tag{15}$$

Stage 2 - Knowledge Recovery: Recover the undesirable concepts $c_e \in \mathbf{E}$ so that the model θ' can generate images of c_e when prompted with $\mathbf{p}_{c_e}$.

$$\mathbf{p}_{c_e} = \underset{\mathbf{p}: \|\mathbf{p} - c_e\|_2 \leq \rho}{\operatorname{argmin}} \left\| \epsilon_{\theta'}(c_e, \mathbf{p}) - \epsilon_{\theta}(c_e) \right\|_2^2 \tag{16}$$

Results

Table: Erasing object-related concepts. Ours* denote the results with the setting with the knowledge of to-be-preserved concepts.

Method	ESR-1↑	ESR-5↑	PSR-1↑	PSR-5↑	RSR-1↑	RSR-5↑
SD	22.0 ± 11.6	2.4 ± 1.4	78.0 ± 11.6	97.6 ± 1.4	N/A	N/A
ESD	95.5 ± 0.8	88.9 ± 1.0	41.2 ± 12.9	56.1 ± 12.4	N/A	N/A
CA	98.4 ± 0.3	96.8 ± 6.1	44.2 ± 9.7	66.5 ± 6.1	N/A	N/A
UCE	100 ± 0.0	100 ± 0.0	23.4 ± 3.6	49.5 ± 8.0	N/A	N/A
UCE*	100 ± 0.0	100 ± 0.0	62.1 ± 34.6	96.0 ± 2.9	N/A	N/A
Ours	99.5 ± 0.3	98.0 ± 1.9	26.6 ± 5.7	47.8 ± 5.0	72.0 ± 11.2	97.2 ± 2.4
Ours*	99.2 ± 0.5	97.3 ± 1.9	75.3 ± 12.0	98.0 ± 0.5	71.5 ± 9.7	95.3 ± 3.6

What's Next?

The implications of our works

$$\min_{\theta'} \mathop{\mathbb{E}}_{c_e \in \mathbf{E}} \max_{c_t \in \mathcal{C}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_t) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_2} \right] \tag{17}$$

Our framework can be apply directly to the model editing/debiasing problems.

- Debiasing: $c_e =$ 'doctor' then choose $\mathcal{C} = \{$'female doctor', 'male doctor', 'black doctor', 'asian doctor', 'old doctor', 'young doctor', etc$\}$

The implications of our works

$$\min_{\theta'} \mathop{\mathbb{E}}_{c_e \in \mathbf{E}} \max_{c_t \in \mathcal{C}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_t) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_2} \right] \tag{17}$$

Our framework can be apply directly to the model editing/debiasing problems.

- Debiasing: $c_e =$ 'doctor' then choose $\mathcal{C} = \{$'female doctor', 'male doctor', 'black doctor', 'asian doctor', 'old doctor', 'young doctor', etc$\}$
- Editing: $c_e =$ 'doctor' to $c_t =$ 'nurse' then choose $\mathcal{C} = \{$'nurse', 'white-nurse', 'asian-nurse', etc$\}$

The implications of our works

$$\min_{\theta'} \mathop{\mathbb{E}}_{c_e \in \mathbf{E}} \max_{c_t \in \mathcal{C}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_t) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_2} \right] \tag{17}$$

Our framework can be apply directly to the model editing/debiasing problems.

- Debiasing: $c_e =$ 'doctor' then choose $\mathcal{C} = \{$'female doctor', 'male doctor', 'black doctor', 'asian doctor', 'old doctor', 'young doctor', etc$\}$
- Editing: $c_e =$ 'doctor' to $c_t =$ 'nurse' then choose $\mathcal{C} = \{$'nurse', 'white-nurse', 'asian-nurse', etc$\}$
- Visual Editing: $c_e =$ TI(images of 'dog') maps to $c_t =$ 'cat' where TI is a Textual Inversion method.

The implications of our works

$$\min_{\theta'} \mathop{\mathbb{E}}_{c_e \in \mathbf{E}} \max_{c_t \in \mathcal{C}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_t) - \epsilon_{\theta}(c_t) \right\|_2^2}_{L_2} \right] \tag{17}$$

Our framework can be apply directly to the model editing/debiasing problems.

- Debiasing: $c_e =$ 'doctor' then choose $\mathcal{C} = \{$'female doctor', 'male doctor', 'black doctor', 'asian doctor', 'old doctor', 'young doctor', etc$\}$
- Editing: $c_e =$ 'doctor' to $c_t =$ 'nurse' then choose $\mathcal{C} = \{$'nurse', 'white-nurse', 'asian-nurse', etc$\}$
- Visual Editing: $c_e = \text{TI}($images of 'dog'$)$ maps to $c_t =$ 'cat' where TI is a Textual Inversion method.
- Mask Editing: $c_e = \text{TI}($images of 'dog and cat'$)$ maps to $c_t = \text{TI}(\mathcal{M}_{\text{dog}}$ (images of 'dog and cat')) where $\mathcal{M}$ is a mask.

Textual Inversion

Images from https://textual-inversion.github.io/

Textual Inversion

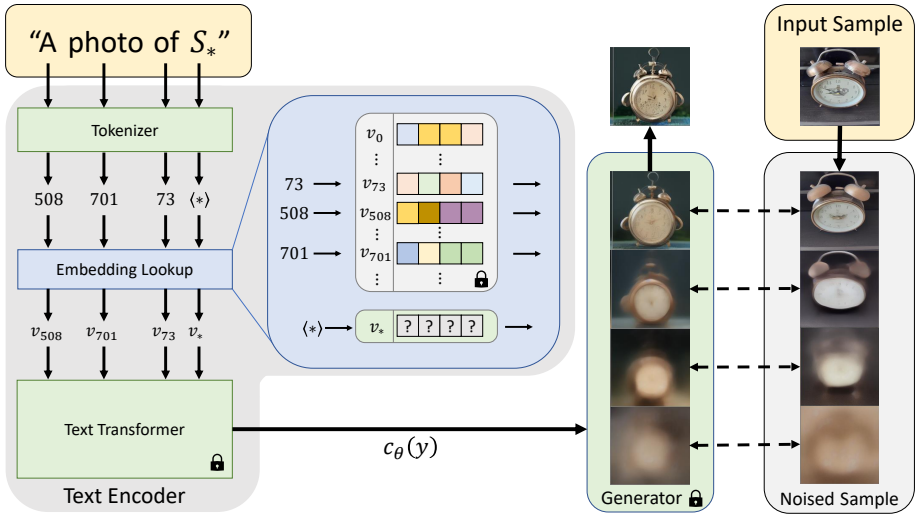


Figure: **Textual Inversion [8].** A string S_* containing our placeholder word is first converted into tokens. These tokens are converted to continous vector representations (the "embeddings", v).

$$v_* = \text{argmin}_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right], \tag{18}$$

where v_* is the embedding vector associated with the placeholder word S_*.

The implications of our works

Model analysis: Given a model θ and new fine-tuned model $\theta^{'}$, how to identify the difference between them?
Approach: Design an inspecting concept space $\mathcal{C}$ and solve the following optimization problem:

$$\max_{c_t \in \mathcal{C}} \left[\underbrace{\|\epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_t)\|_2^2}_{L_1} + \lambda \underbrace{\|\epsilon_{\theta'}(c_t) - \epsilon_{\theta}(c_t)\|_2^2}_{L_2} \right] \tag{19}$$

This helps us identify the most sensitive concepts (but not the ones that are changed c_e) after fine-tuning $\theta \rightarrow \theta^{'}$.
By changing λ, we can control the similarity of the inspecting concepts to the changed concepts.
This approach can be applied to many tasks:

- Machine Unlearning: Identifying which concepts are critical and should be preserved.
- Personalization: Identifying which concepts are being changed most after personalization.

[1] 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.

[2] 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.

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

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

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

[6] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[7] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.

[8] R. Gal, Y. Alaluf, Y. Atzmon, *et al.*, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.