Erasing Undesirable Concepts in Diffusion Models: What next?

Tuan-Anh Bui[1]

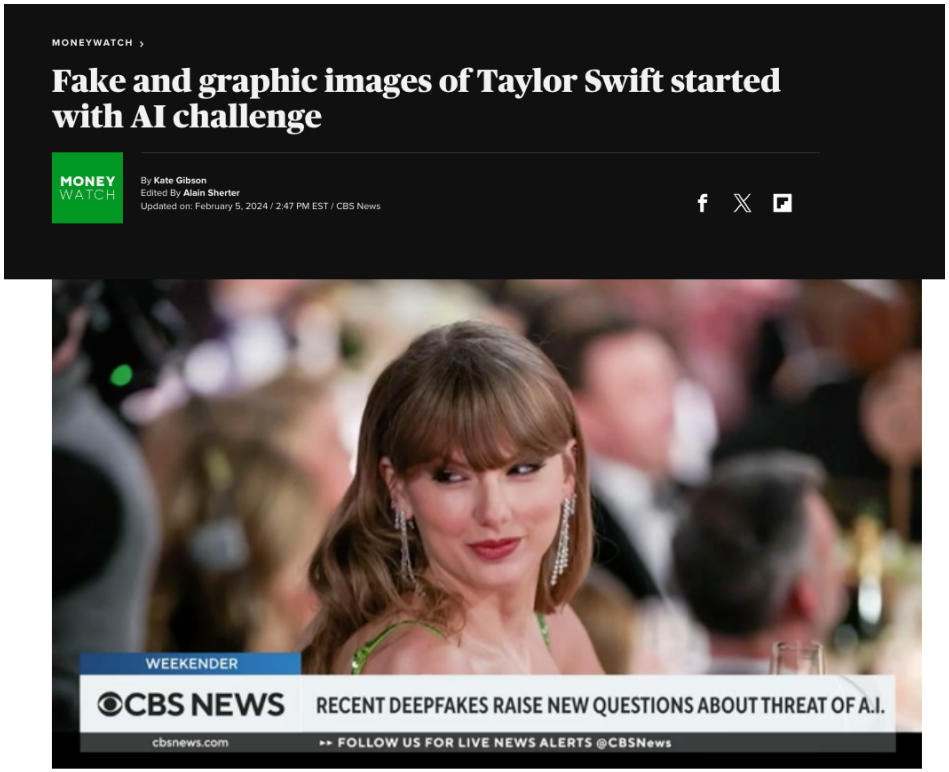
[1]Department of Data Science and AI
Faculty of Information Technology
Monash University

TML NGT Project Meeting

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

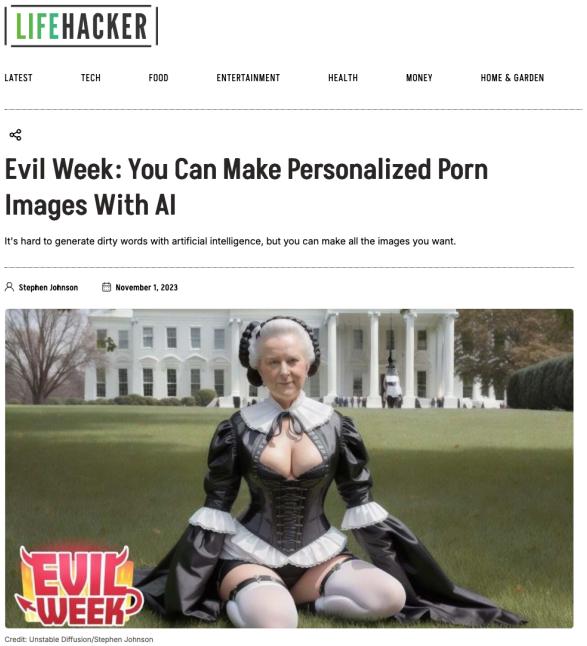
With just a single reference image, our Infinite-ID framework excels in synthesizing high-quality images while maintaining superior identity fidelity and text semantic consistency in various styles.

- **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



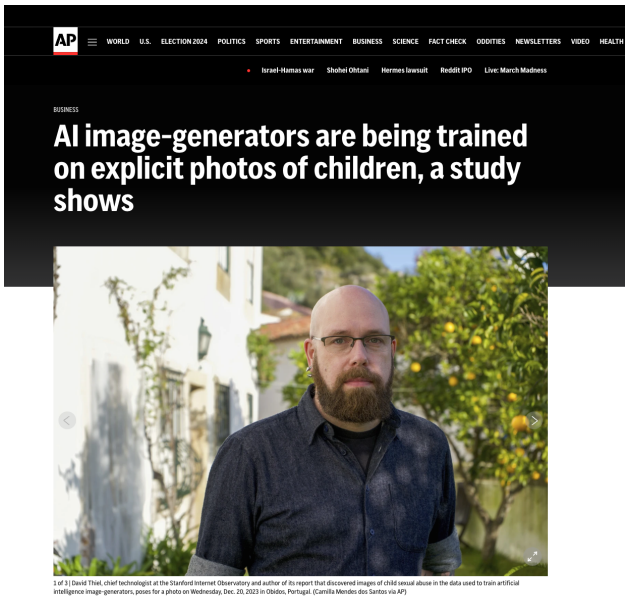


Table of Contents

Motivation

The naive approach that has been used in previous works [1]–[3] is to optimize the following objective function:

$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2 + \lambda \left\| \epsilon_{\theta'}(c_n) - \epsilon_{\theta}(c_n) \right\|_2^2 \right] \tag{1}$$

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.

Motivation

The naive approach that has been used in previous works [1]–[3] is to optimize the following objective function:

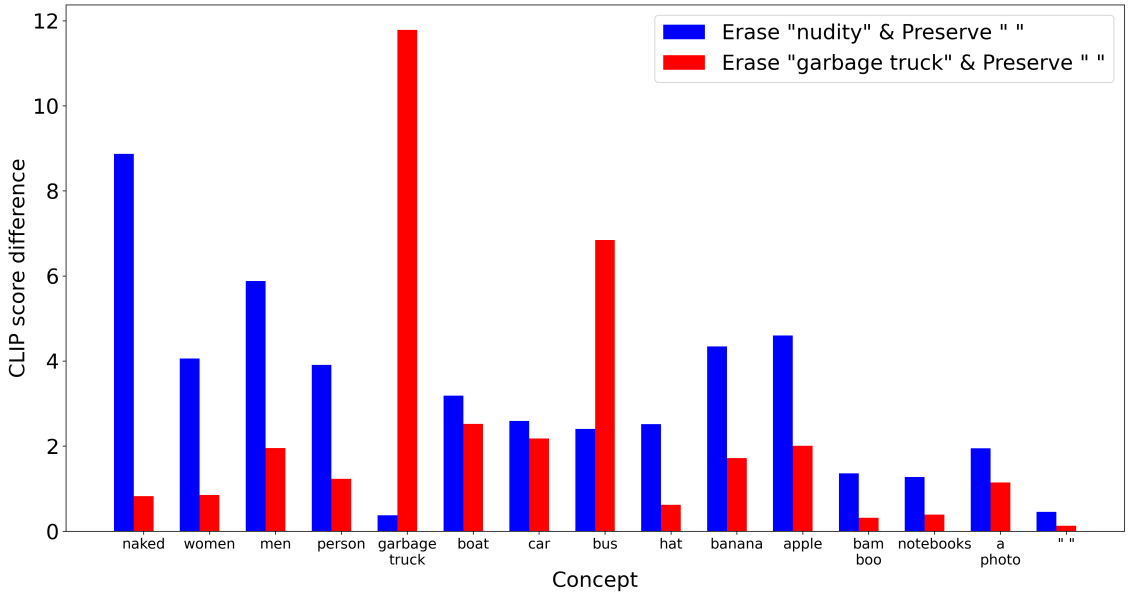
$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2 + \lambda \left\| \epsilon_{\theta'}(c_n) - \epsilon_{\theta}(c_n) \right\|_2^2 \right] \quad (1)$$

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.

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

- How to measure the impact of erasing a concept on the generation of other concepts?
- How to search for the most sensitive concept to the erasing concept?

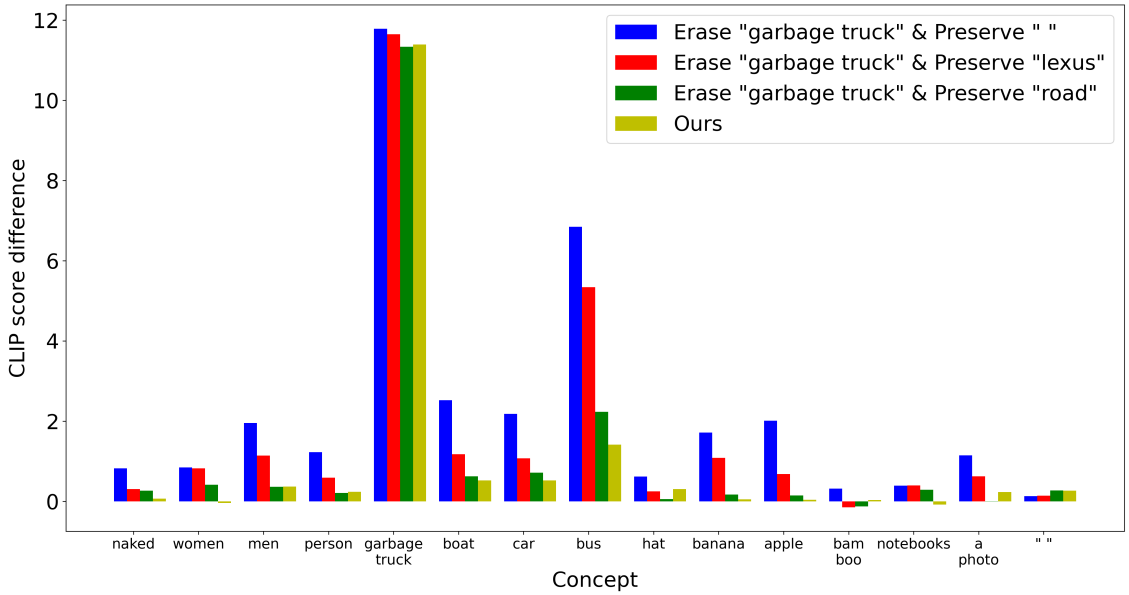
Impact on the model's capability: Results



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.

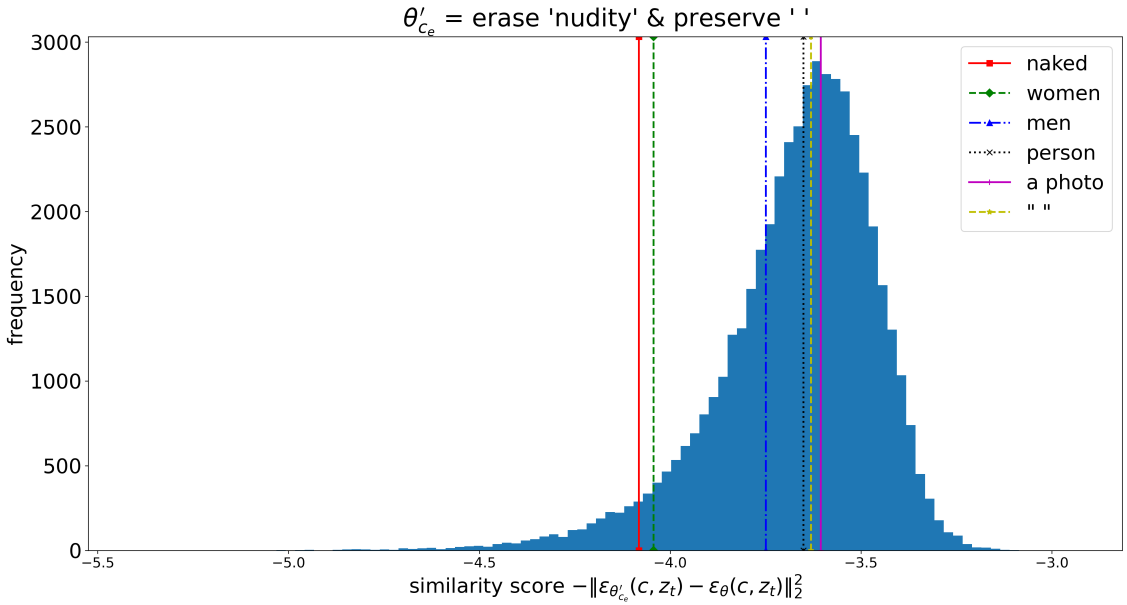
Impact on the model's capability: Results



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.

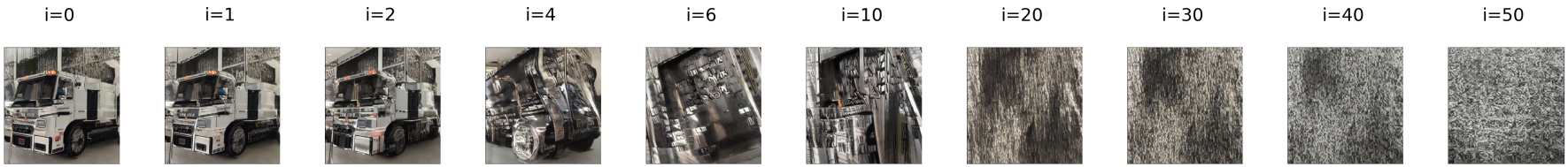
Neutral concepts lie in the middle of the spectrum. Again, not a good choice to preserve!

Objective Function: First Attempt

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

- Inner Max L_2 w.r.t. c_a: Searching for the most sensitive concept to the erasing concept c_e.
- Outer Min $L_1 + L_2$: Erasing the concept c_e and preserving the adversarial concept c_a, simultaneously.

Objective Function: Solving with PGD



$$\min_{\theta'} \max_{c_a \in \mathcal{R}} \mathbb{E}_{c_e \in \mathbf{E}} \Big[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_a) - \epsilon_{\theta}(c_a) \right\|_2^2}_{L_2} \Big] \tag{3}$$

Solving the optimization problem with PGD:

- Init $c_{a,t=0} = c_e = \tau(\text{"garbage truck"})$.
- Iteratively update $c_{a,t}$ with the gradient of L_2 w.r.t. c_a.

The adversarial concept c_a quickly converges to background noise type of concept. *Continuos concept space is not suitable for adversarial preservation.*

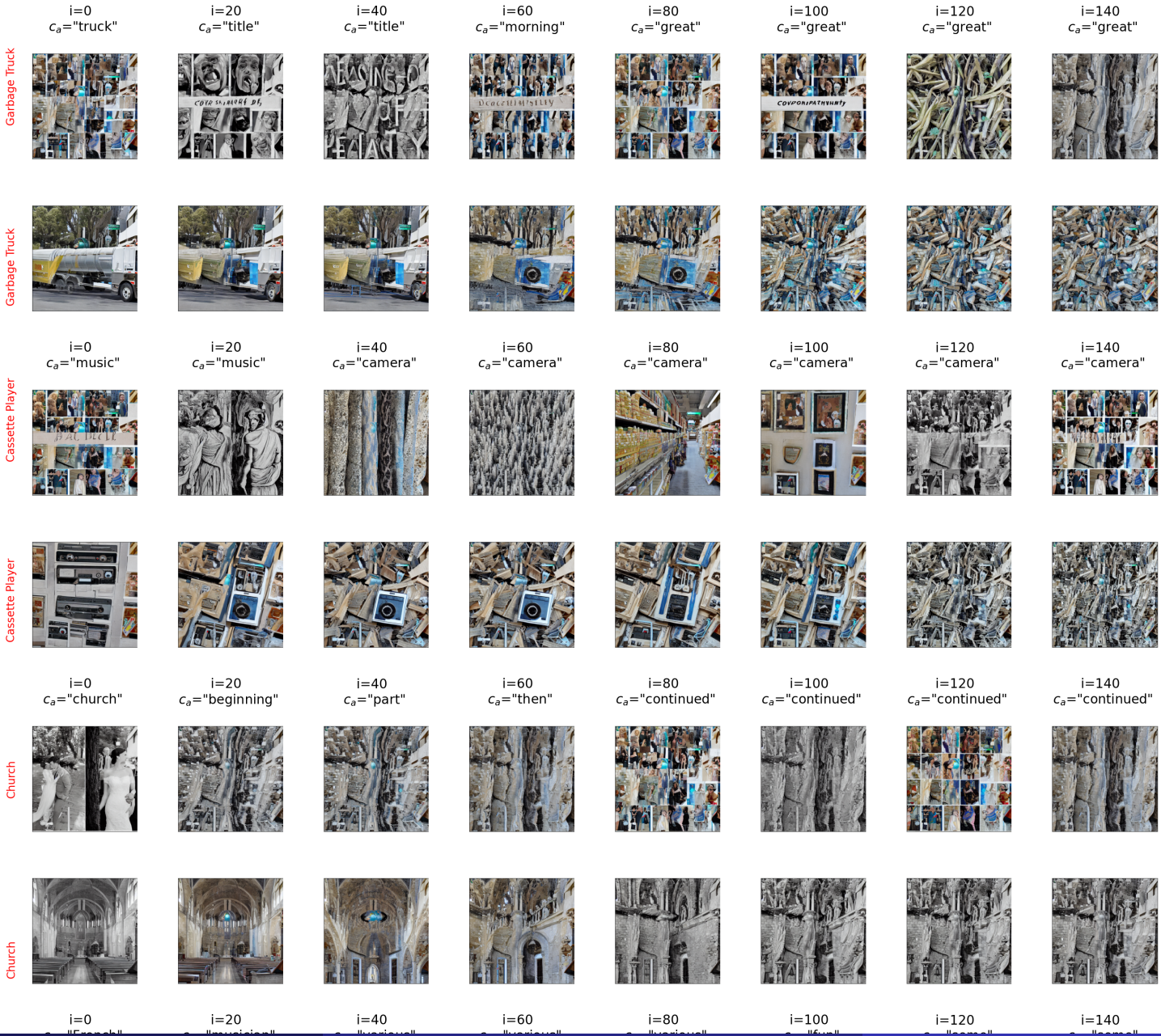
Objective Function: Relaxation with Gumbel-Softmax

$$\min_{\theta'} \max_{\pi \in \Delta_{\mathcal{R}}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\|\epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n)\|_2^2}_{L_1} + \lambda \underbrace{\|\epsilon_{\theta'}(\mathbf{G}(\pi) \odot \mathcal{R}) - \epsilon_{\theta}(\mathbf{G}(\pi) \odot \mathcal{R})\|_2^2}_{L_2} \right] \tag{4}$$

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 $\Delta_{\mathcal{R}}$.

Table of Contents

Idea 1 - Collapse of Concepts



Idea 1 - Collapse of Concepts

$$\min_{\theta'} \max_{c_a \in \mathcal{R}} \mathbb{E}_{c_e \in \mathbf{E}} \left[\underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_a) - \epsilon_{\theta}(c_a) \right\|_2^2}_{L_2} \right] \tag{5}$$

Intuition: Using the same concept c_n for all erasing tasks in $\mathbf{E}$ may lead to the collapse of concepts.

Idea: Multiple target concepts c_n and multiple adversarial concepts c_a.

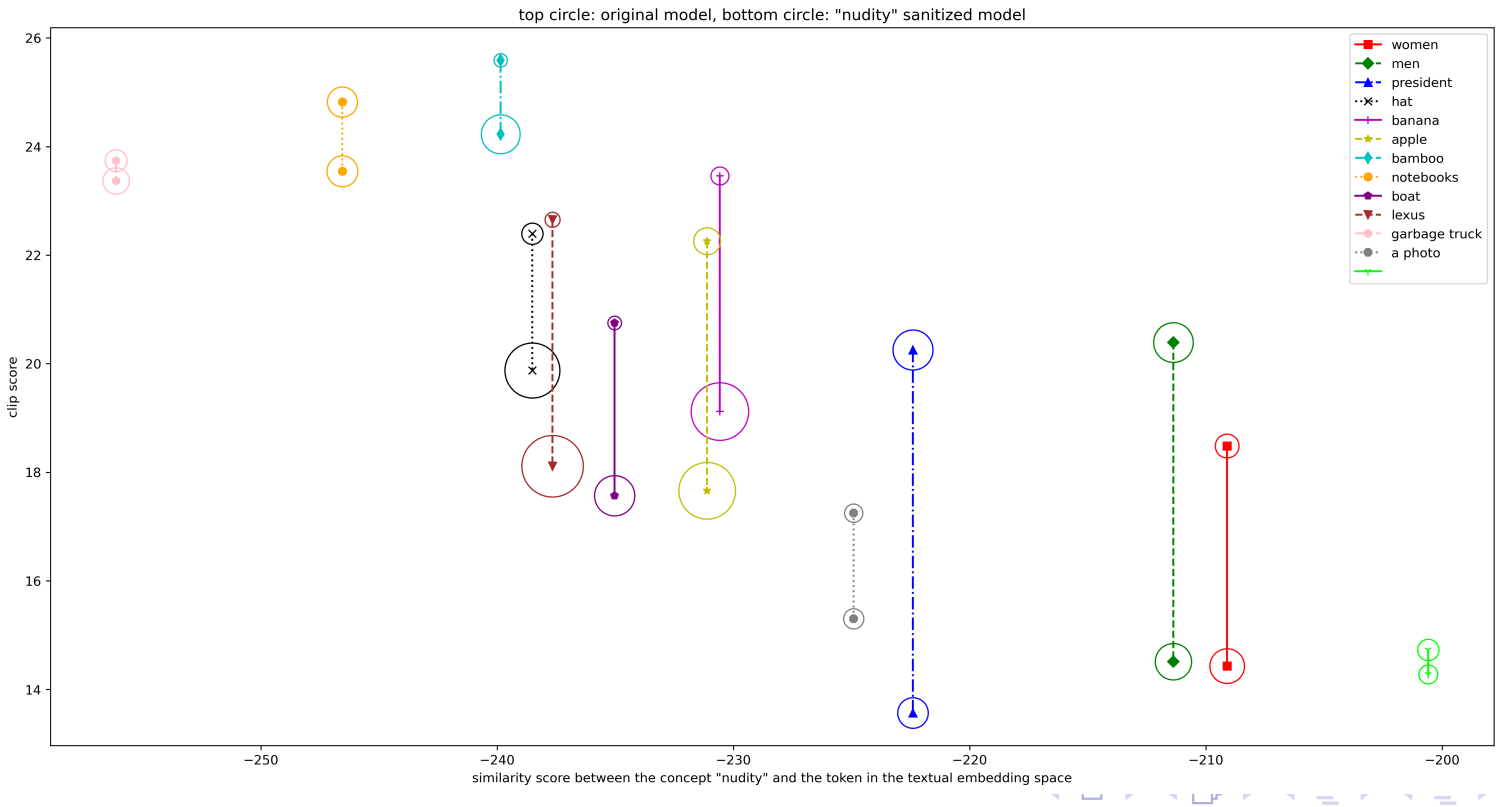
$$\min_{\theta'} \mathbb{E}_{c_e \in \mathbf{E}} \left[\max_{c_a \in \mathcal{R}} \min_{c_n \in \mathcal{B}(c_e)} \underbrace{\left\| \epsilon_{\theta'}(c_e) - \epsilon_{\theta}(c_n) \right\|_2^2}_{L_1} + \lambda \underbrace{\left\| \epsilon_{\theta'}(c_a) - \epsilon_{\theta}(c_a) \right\|_2^2}_{L_2}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

Idea 2 - Modeling the Impact Function

Question: Can we use the similarity in the textual embedding space to find the most sensitive concept? i.e., $c_a \approx \arg\max_{c \in \mathcal{R}} \text{sim}(c, c_e)$.
Why? Because multimodal models like CLIP which were trained on large-scale datasets paired with text and images, have shown that the textual embedding space is semantically meaningful and can be used to search for similar concepts.

Idea 2 - Modeling the Impact Function

Question: Can we use the similarity in the textual embedding space to find the most sensitive concept? i.e., $c_a \approx \arg\max_{c \in \mathcal{R}} \text{sim}(c, c_e)$.
Why? Because multimodal models like CLIP which were trained on large-scale datasets paired with text and images, have shown that the textual embedding space is semantically meaningful and can be used to search for similar concepts.
Short answer: No. The similarity in the textual embedding space does not reflect the similarity in the visual space, i.e., $\text{sim}(c_i, c_j) \neq \text{sim}(G(c_i), G(c_j))$.



Idea 2 - Modeling the Impact Function

Question: Can we use the similarity in the textual embedding space to find the most sensitive concept? i.e., $c_a \approx \arg\max_{c \in \mathcal{R}} \text{sim}(c, c_e)$.
Short answer: No. The similarity in the textual embedding space does not reflect the similarity in the visual space, i.e., $\text{sim}(c_i, c_j) \neq \text{sim}(G(c_i), G(c_j))$.
Intuition: Because of the **cross-attention mechanism**

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

Such that $\text{sim}(c_i, c_j) \neq \text{sim}(Wc_i, Wc_j) \neq \text{sim}(O(c_i), O(c_j))$

Idea 2 - Modeling the Impact Function

Idea: Can we learn a model $f_\phi()$ to predict the impact on concept c_j when erasing c_i for any arbitrary pair of concepts c_i, c_j? I.e., $f_\phi(c_i, c_j) \approx \delta_{c_i}(c_j)$, where $\delta_{c_i}(c_j)$ is the groundtruth impact function.
Implications: We can use $f_\phi()$ to search for the most sensitive concept to the erasing concept c_i and many other implications, i.e., long-tail distribution, continual learning.

Idea 2 - Modeling the Impact Function

Idea: Can we learn a model $f_\phi()$ to predict the impact on concept c_j when erasing c_i for any arbitrary pair of concepts c_i, c_j? I.e., $f_\phi(c_i, c_j) \approx \delta_{c_i}(c_j)$, where $\delta_{c_i}(c_j)$ is the groundtruth impact function.
Implications: We can use $f_\phi()$ to search for the most sensitive concept to the erasing concept c_i and many other implications, i.e., long-tail distribution, continual learning.

Difficulties:
(1) What is the impact function $\delta_{c_i}(c_j)$? "The impact of erasing the concept c_i on the generation of the concept c_j", is an ambiguous definition? For example, how to measure the generation capability of the concept c_j? What are desired properties of the impact function $\delta_{c_i}(c_j)$?
(2) The **empirical** impact function $\delta_{c_i}(c_j)$ is very expensive to obtain. It is because: (i) requiring a sanitized model $\theta'_{c_i} = \mathcal{A}(\theta, c_i)$ and (ii) $G(\theta'_{c_i}, c_j, z_T)$ depends on the initial z_T which is stochastic, current solution is to generate K samples for each concept c_j for each pair of concepts c_i, c_j.

Solution: Take randomness into account.

Idea 3 - Expressiveness of a Concept

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

How to express/describe a concept c?
The current (and only) approach: textual embedding $c = \tau(s)$ where s is the textual description of the concept c.
Limitation: Limited expressiveness

- Not all concepts can be described by text. For example, "Tony's face" or "Dinh's face" concept.
- Not all concepts can be described by a single text. For example, "nudity" can have many ways to describe.

Idea 3 - Expressiveness of a Concept

Idea: Can we use the visual embedding $c = \tau(I)$ where I is a reference image of the concept c to represent c? But the input of the model is text embedding!
Solution: Searching for the most similar text embedding to the visual embedding of the concept c or textual inversion

$$c^* = f(\theta, \mathcal{D}_c) = \arg\min_{p \in \mathcal{C}} \mathbb{E}_{x \in \mathcal{D}_c} \|\epsilon_\theta(p) - x\|_2^2 \tag{8}$$

Then

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

where $\mathbb{E} = \{c_e^*\} = \{f(\theta', \mathcal{D}_{c_e})\}$ is the set of visual expressions of to-be-erase concepts c_e and $c_n^* = f(\theta, \mathcal{D}_{c_n})$ is the visual expression of the neutral concept c_n.

Idea 4 - Defending against Inversion Attack

Inversion Attack: Given a sanitized model $\theta^{'}$, we can still able to recover/generate erased concepts c_e by learning the embedding vector to represent the concept c_e through the lens of the sanitized model θ'

$$c_e^* = f(\theta^{'}, \mathcal{D}_{c_e}) = \arg\min_{p \in \mathcal{C}} \mathbb{E}_{x \in \mathcal{D}_{c_e}} \left\| \epsilon_{\theta'}(p) - x \right\|_2^2 \tag{10}$$

Then $G(\theta^{'}, c_e^*, z_T)$ can still generate the concept c_e. How to defend?

Idea 4 - Defending against Inversion Attack

Inversion Attack: Given a sanitized model $\theta^{'}$, we can still able to recover/generate erased concepts c_e by learning the embedding vector to represent the concept c_e through the lens of the sanitized model θ'

$$c_e^* = f(\theta^{'}, \mathcal{D}_{c_e}) = \arg\min_{p \in \mathcal{C}} \mathbb{E}_{x \in \mathcal{D}_{c_e}} \left\| \epsilon_{\theta'}(p) - x \right\|_2^2 \tag{10}$$

Then $G(\theta^{'}, c_e^*, z_T)$ can still generate the concept c_e. How to defend?

Intuition: This is very hard (nearly impossible) problem. For example, the concept "nudity" can be described (or built) by basic concepts like "body", "skin", "face", etc. *Whenever the model can generate these basic concepts, it can generate the concept "nudity".*
Our take: Just make it **harder** for the attacker.

$$\min_{\theta^{'}} \mathbb{E}_{x \in \mathcal{D}_{c_e}} \max_{p \in \mathcal{C}} \left[\underbrace{\left\| \epsilon_{\theta'}(p) - \epsilon_{\theta}(c_n) \right\|_2^2}_{L_1} - \lambda \underbrace{\left\| \epsilon_{\theta'}(p) - x \right\|_2^2}_{L_2} + L_{\text{preservation}} \right] \tag{11}$$

Idea 5 - Antipersonalization with Alpha Concepts

Interesting observation: Elon Musk is a strong alpha-male concept :D

Dinh sitting next to Taylor Swift

Dinh and Elon Musk smoking cigarette together at a bar

Dinh and Elon Mush having beers

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

[2] 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.

[3] 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.