Erasing Undesirable Concepts in Diffusion Models with Adversarial Preservation

Anh Bui[1], Long Vuong[1], Khanh Doan[2], Trung Le[1], Paul Montague[3], Tamas Abraham[3], Dinh Phung[1]

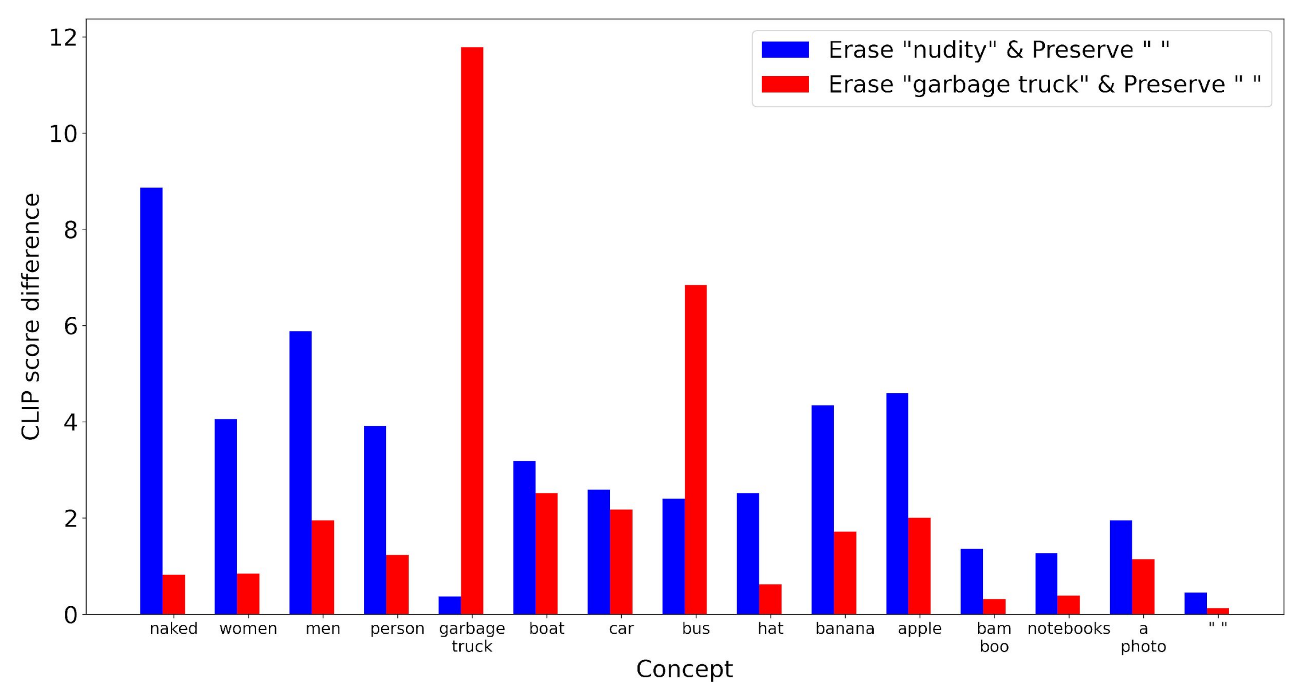
[1] Monash University, [2] VinAI Research, [3] DSTG

KEY OBSERVATIONS

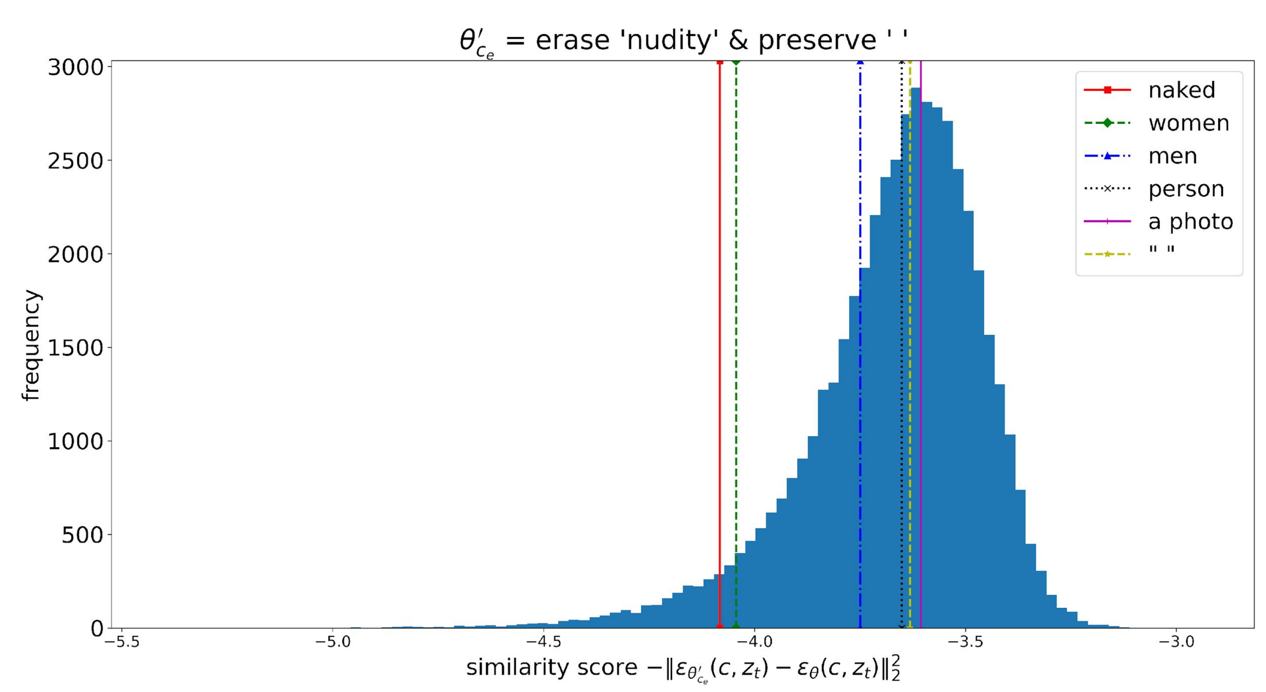
How to measure the Side-Effect of Concept Erasure

- CLIP alignment score $S_{\theta,i,c} = S(G(\theta, c, z_T^i), c)$ → the higher score, the better model can generate concept c
- $\delta_{c_e}(c) = \frac{1}{k}\sum_{i=1}^{k}(S_{\theta,i,c} - S_{\theta'_{c_e},i,c})$ → the larger different, the higher side-effect (negatively) to model's capability

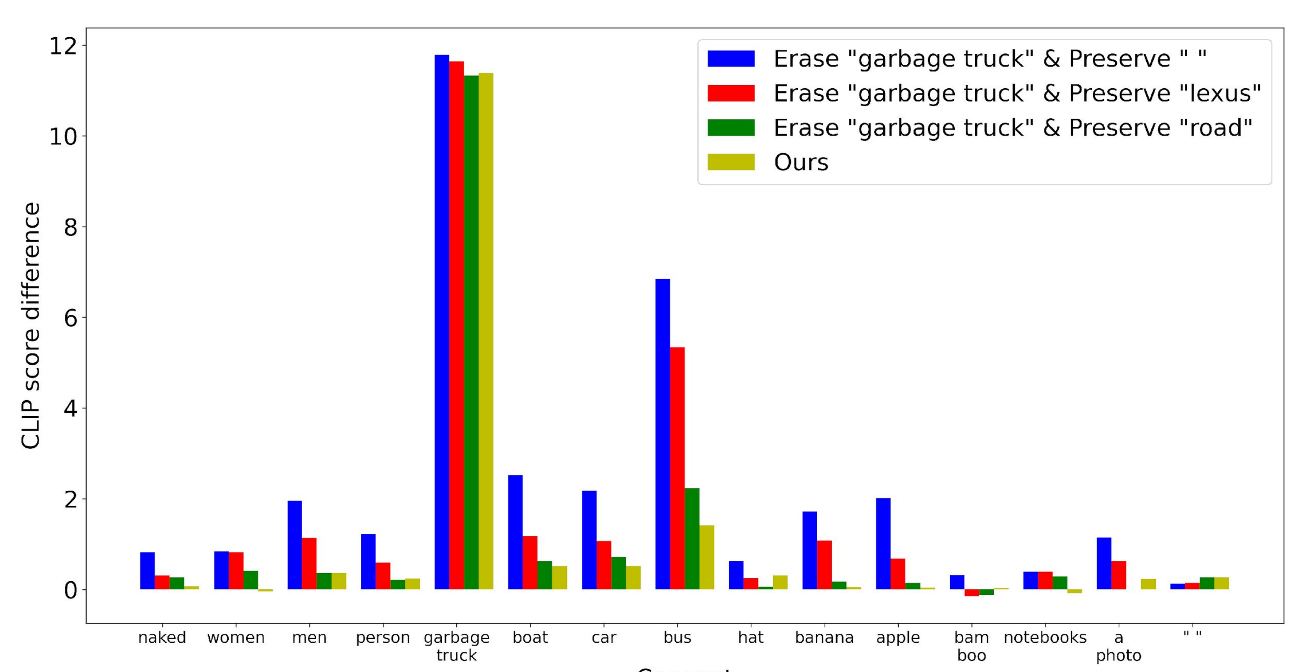
1 - Erasing Different Concepts Leads to Different Side-Effects

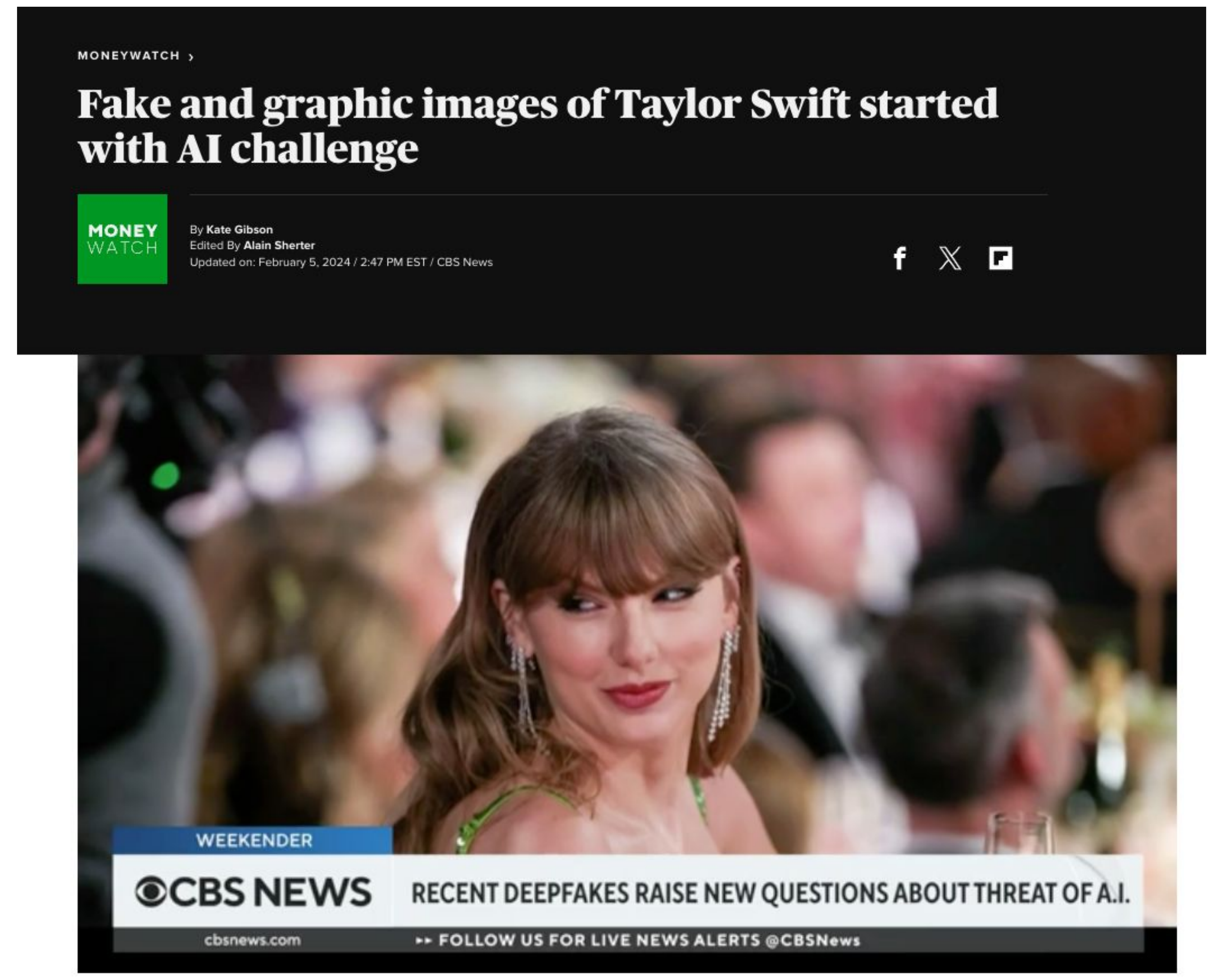


2 - Neutral Concepts lie in the Middle of the Sensitivity Spectrum



3 – What Concept Should be Kept to Minimize the Side-Effect





How to prevent AI-generated "po*n" content?

CONCEPT ERASURE

Naïve Approach

$$\min_{\theta'} \mathbb{E}_{c_e\in E}\left[\|\epsilon_{\theta'}(c_e) - \epsilon_\theta(c_n)\|_2^2\right] + L_2$$

Where:

- θ, θ': original and sanitized models
- $c_e \in E$: concept to-be-erased (e.g., `nudity')
- c_n: neutral/generic concept (e.g., `a photo')
- $\epsilon_\theta(c)$: noise-prediction function
- L_2: preservation loss

$$L_2 = \|\epsilon_{\theta'}(c_n) - \epsilon_\theta(c_n)\|_2^2 \text{ or } \|\theta' - \theta\|_2^2$$

From Observations to Motivation:

- Observation 2 → Preserving a neutral/generic concept c_n is sub-optimal.
- Observation 1 → to-be-preserved concept should be adaptive.
- Observation 3 → to-be-preserved concept should be related to the to-be-erased concepts.

ADVERSARIAL PRESERVATION

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

Where:

- θ, θ': original and sanitized models
- c_a: `Adversarial` concept, i.e., the concept will be affected most by the erasure
- $\mathcal{R}$: Concept space to search c_a

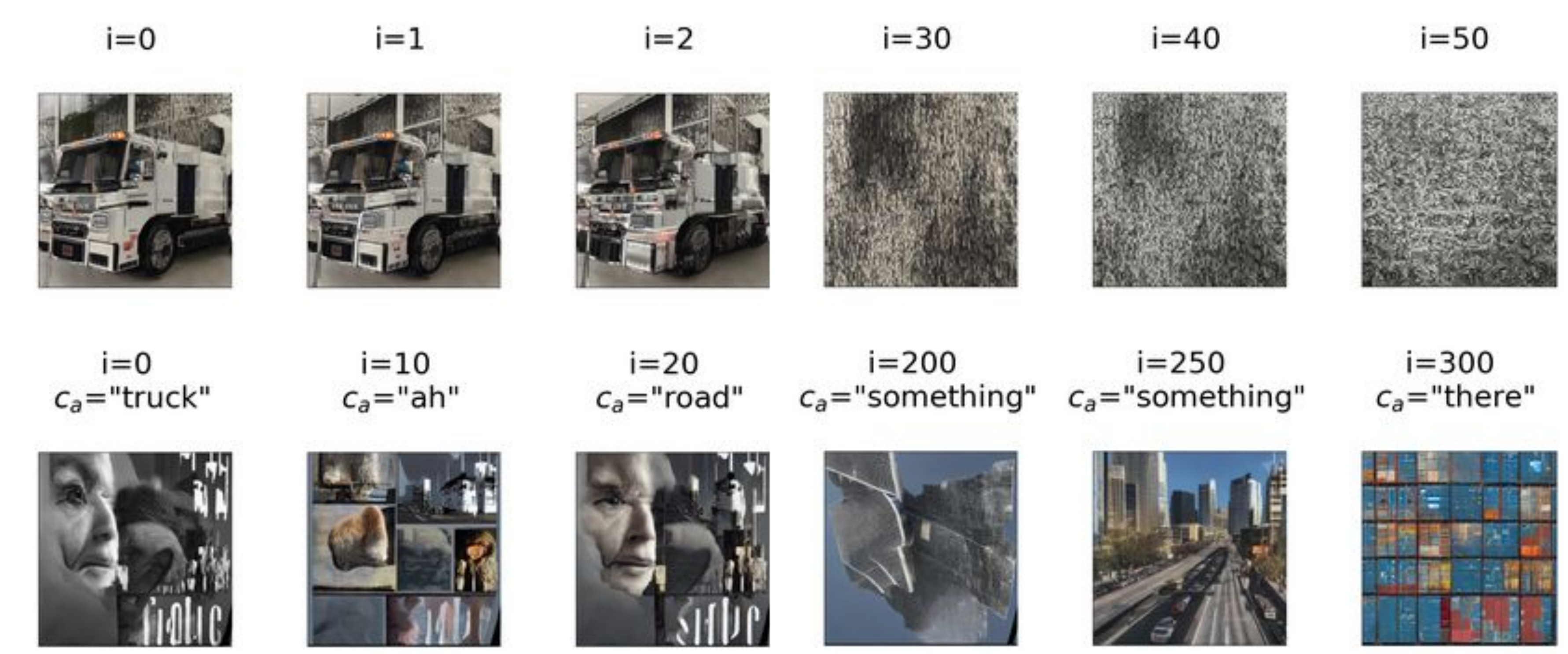
Interpretation:

- Inner-Max: Find adversarial concept that is affected most by the erasure
- Outer-Min: Update model to erasure E and preserve c_a, simultaneously.

Finding Adversarial Concept with PGD

- Init $c_{a,t=0} = c_e$, e.g., $\triangleq \tau(\text{``garbarge truck''})$
- Iteratively update $c_{a,t+1} = c_a + \eta \nabla_{c_a} L_2$

However, c_a quickly converges to background noise/non-sense type of concept



Relaxation with Gumbel-Softmax

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

- Modelling c_a as a distribution over the concept space $\mathcal{R}$
- Searching π on the simplex $\Delta_{\mathcal{R}}$

EXPERIMENTAL RESULTS AND MORE