

Best Paper
Honorable
Mention

Concept Arithmetics for Circumventing Concept Inhibition in Diffusion Models



Vitali Petsiuk, Kate Saenko
Boston University



BOSTON
UNIVERSITY



Diffusion models for image generation

the good



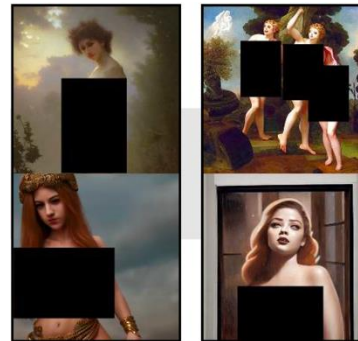
Diffusion
model

the ugly

copyrighted

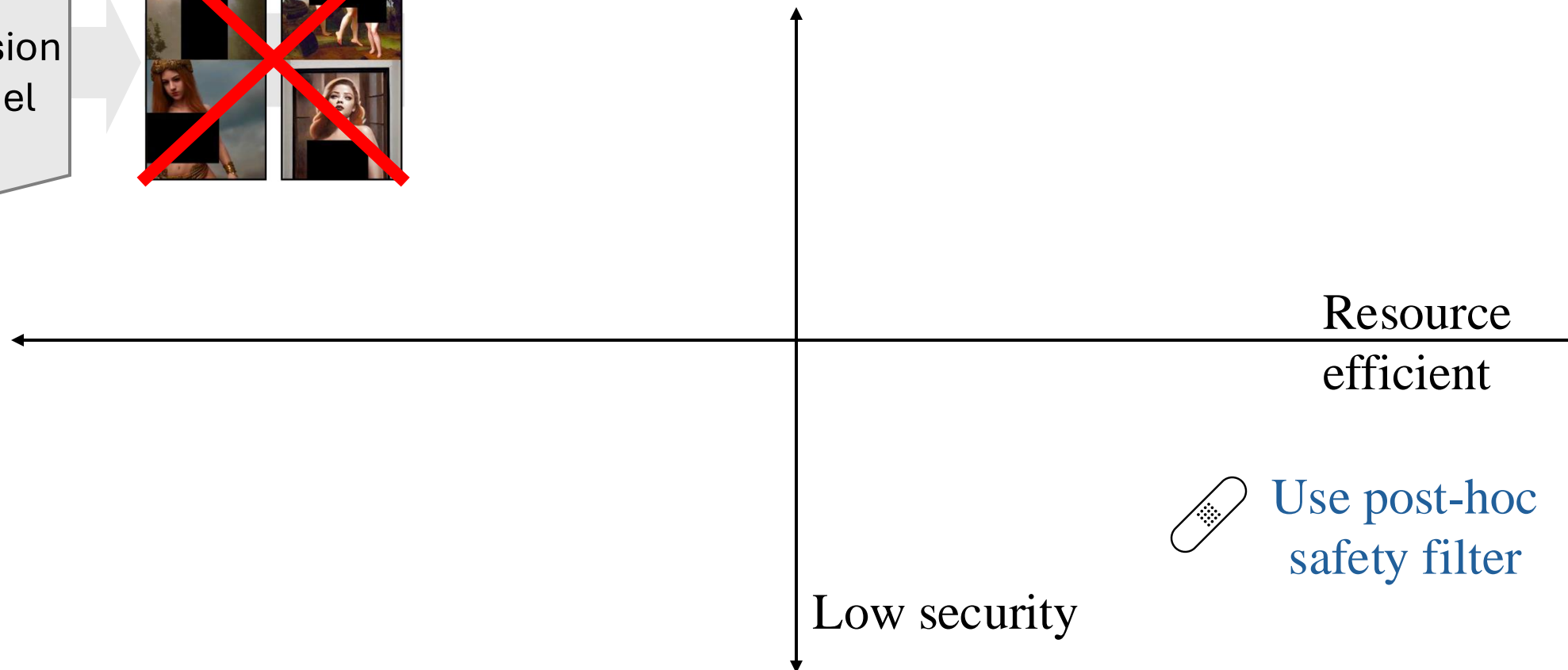
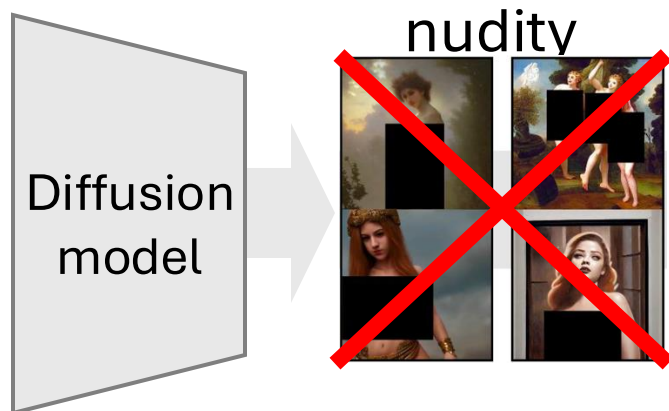


nudity

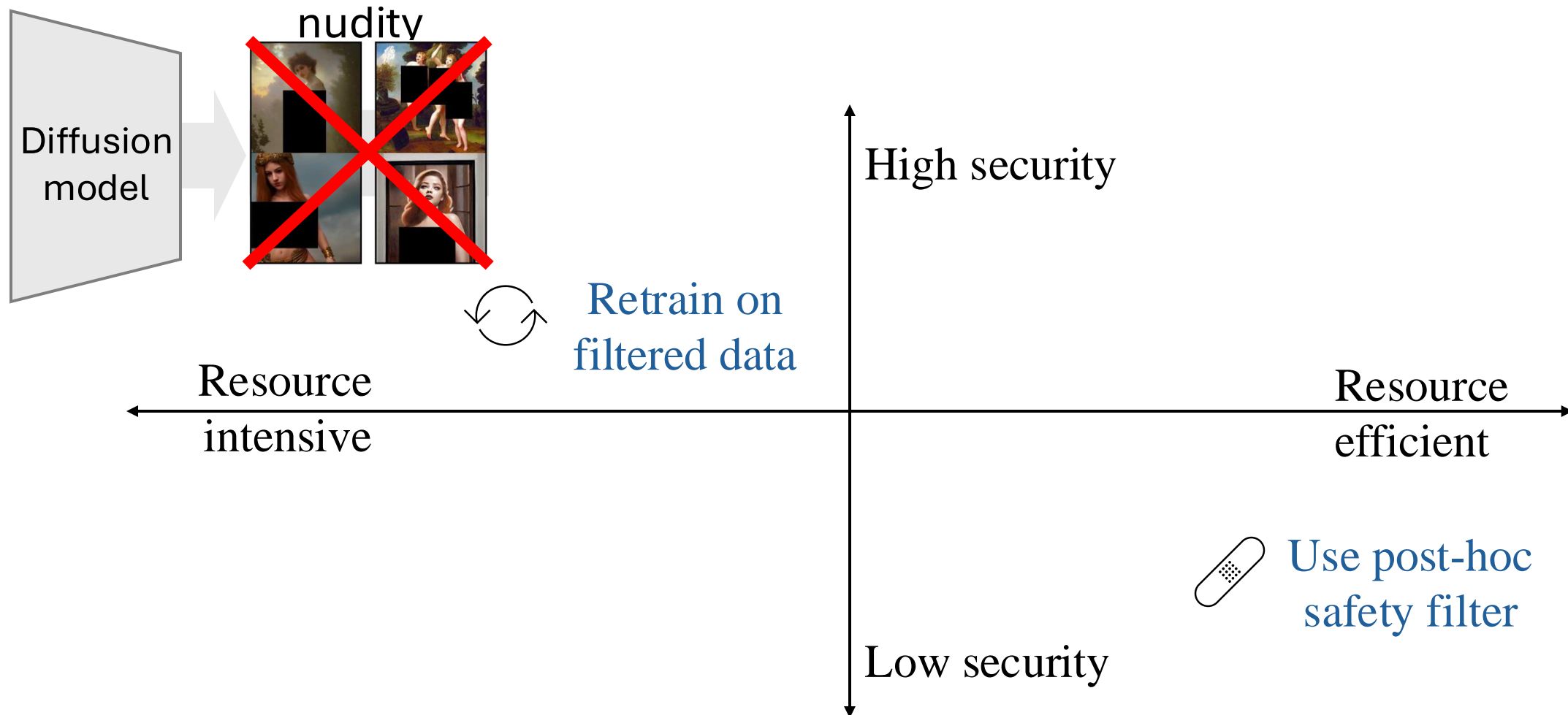


...hate, harassment,
violence, self-harm,
porn, shocking, illegal
activity...

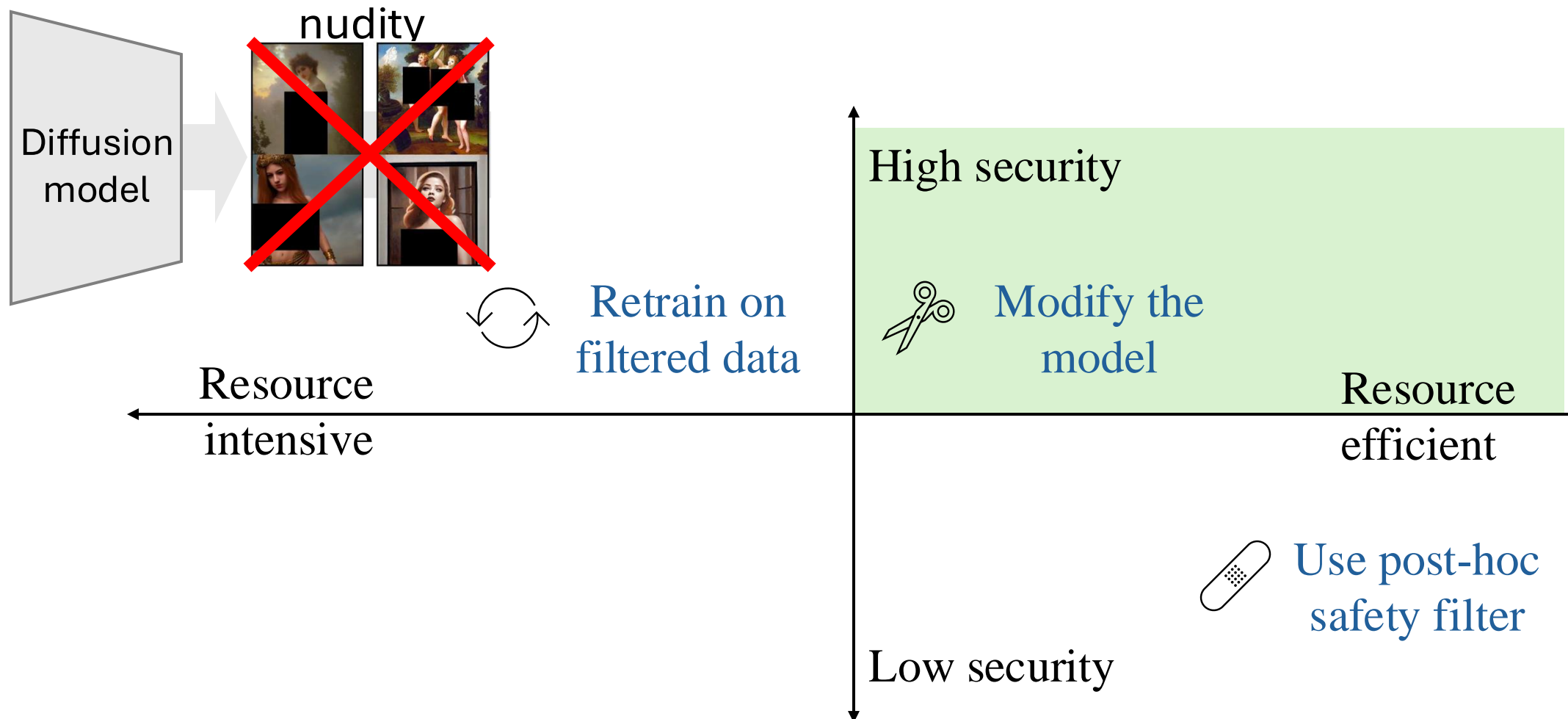
How to limit unethical image generation?



How to limit unethical image generation?



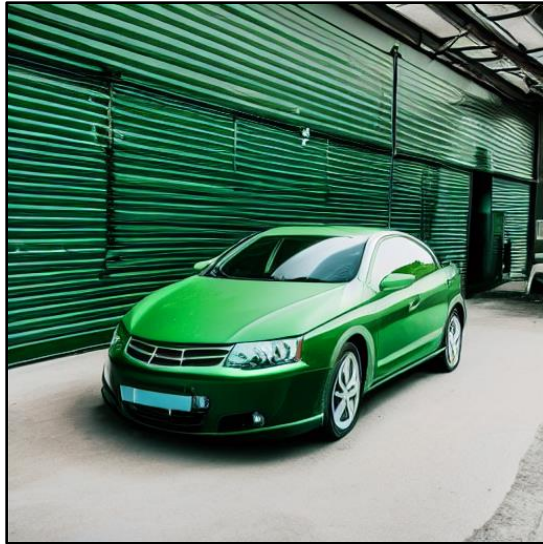
How to limit unethical image generation?




✂ Prior work: modify the model to erase concepts

Original
model

Prompt: car

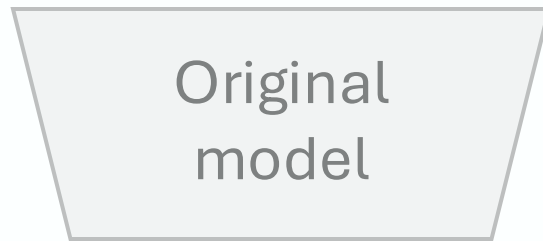


 Modified
model

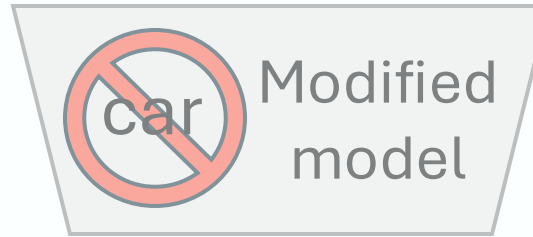
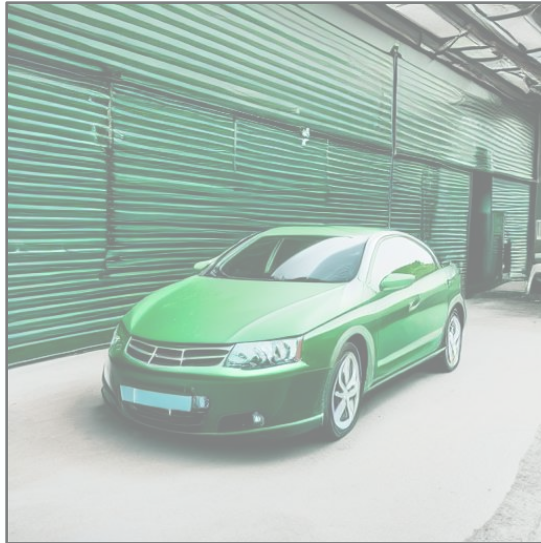
Prompt: car



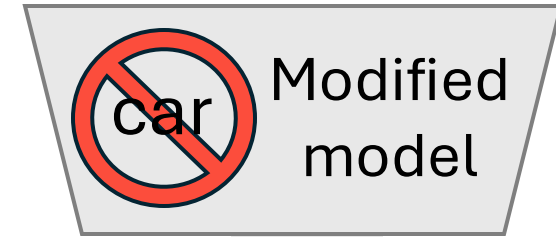
This work: concepts are not really erased!



Prompt: car



Prompt: car

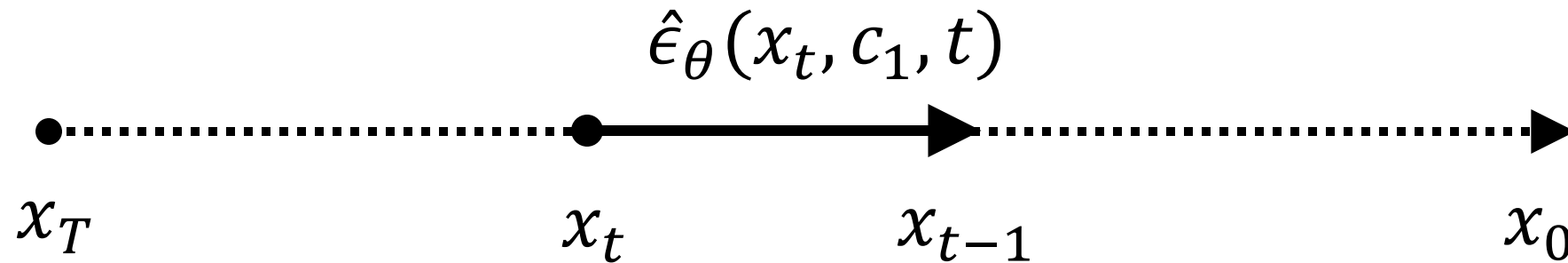


Our prompt



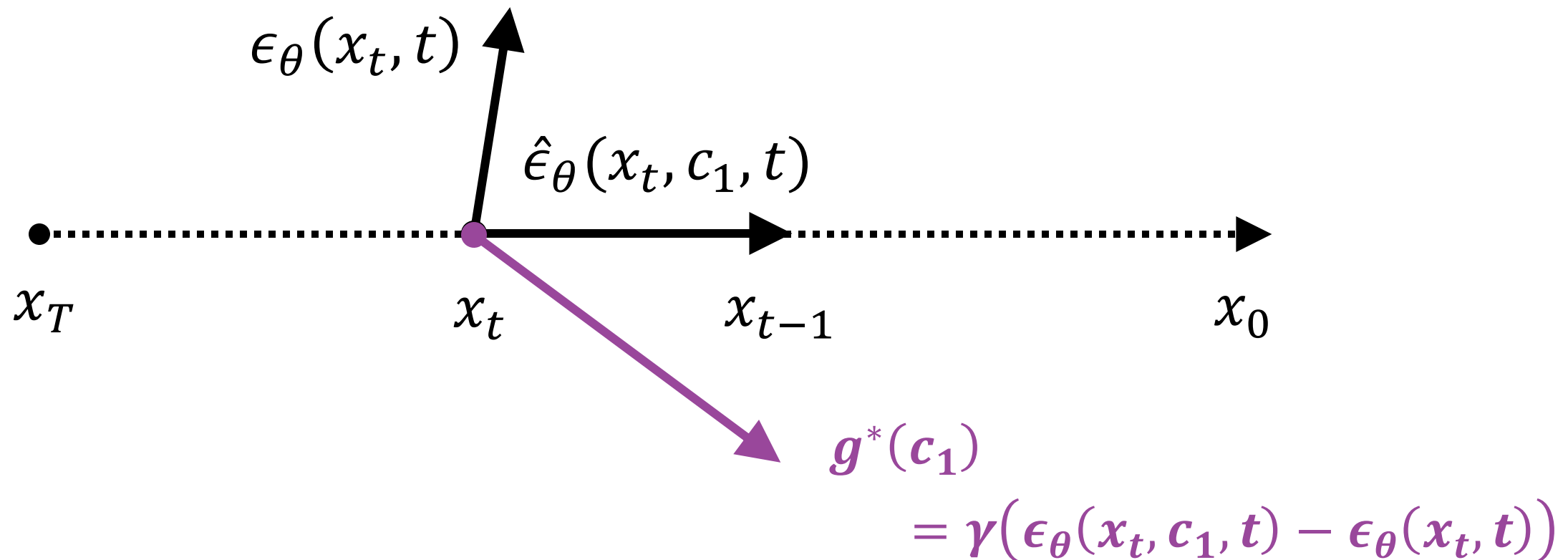
Background: Denoising Diffusion Step

$c_1 =$ A horse standing in the field



Background: Denoising Diffusion Step

$c_1 =$ A horse standing in the field



Background: Concept Inhibition

Optimize the weights to enforce some predefined output as guidance for target.

$$g(c_\tau) \leftarrow y_0 \quad \text{via optimization} \quad \theta = \arg \min_{\theta} \mathcal{L}(g_{\theta}(c_\tau), y_0)$$

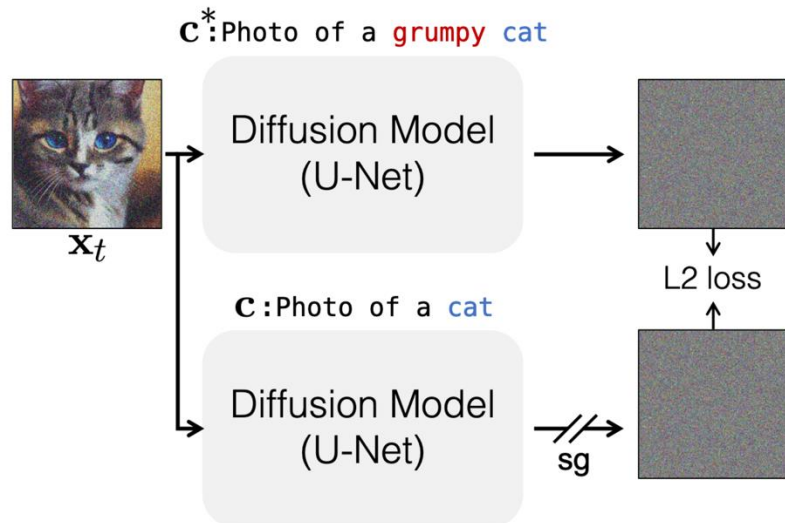
Background: Concept Inhibition

Optimize the weights to enforce some predefined output as guidance for target.

$$g(c_\tau) \leftarrow y_0 \quad \text{via optimization} \quad \theta = \arg \min_{\theta} \mathcal{L}(g_{\theta}(c_\tau), y_0)$$

$$\epsilon_{\theta}(x_t, c_\tau, t) \leftarrow \epsilon_{\theta^*}(x_t, c_\alpha, t)$$

$$g(c_\tau) \leftarrow g^*(c_\alpha)$$



Ablating Concepts (AC) [Kumari et al., 2023]

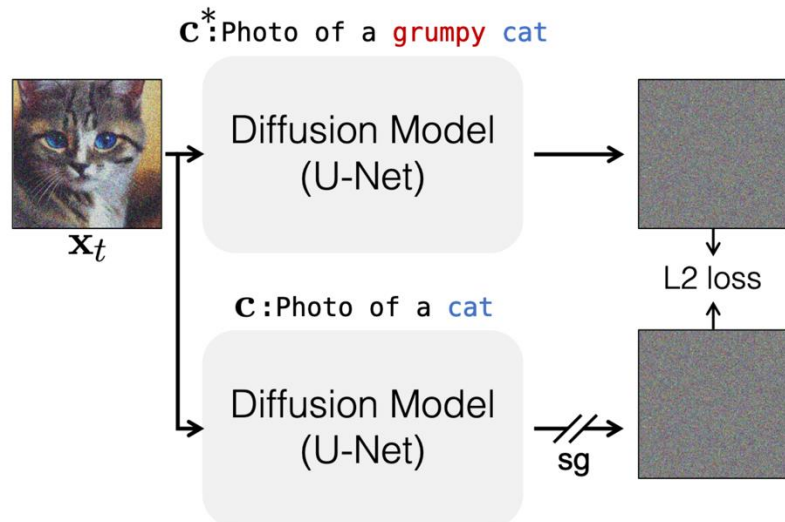
Background: Concept Inhibition

Optimize the weights to enforce some predefined output as guidance for target.

$$g(c_\tau) \leftarrow y_0 \quad \text{via optimization} \quad \theta = \arg \min_{\theta} \mathcal{L}(g_{\theta}(c_\tau), y_0)$$

$$\epsilon_{\theta}(x_t, c_\tau, t) \leftarrow \epsilon_{\theta^*}(x_t, c_\alpha, t)$$

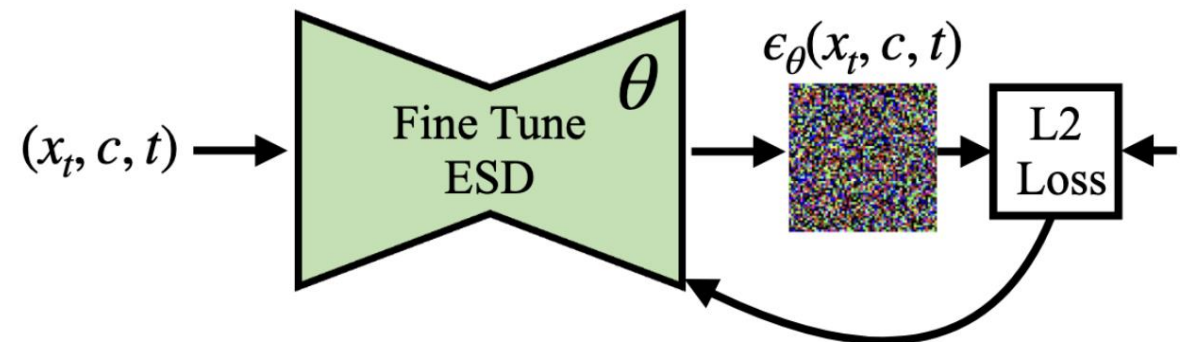
$$g(c_t) \leftarrow g^*(c_\alpha)$$



Ablating Concepts (AC) [Kumari et al., 2023]

$$\epsilon_{\theta}(x_t, c_\tau, t) \leftarrow \epsilon_{\theta^*}(x_t, t) - \eta(\epsilon_{\theta^*}(x_t, c_\tau, t) - \epsilon_{\theta^*}(x_t, t))$$

$$g(c_t) \leftarrow g^*(\emptyset) - g^*(c_t)$$



Erasing Concepts (ESD) [Gandikota et al., 2023]

Analyzing Concept Inhibition

Optimize the weights to enforce some predefined output as guidance for target.

$$g(c_\tau) \leftarrow y_0 \quad \text{via optimization} \quad \theta = \arg \min_{\theta} \mathcal{L}(g_{\theta}(c_\tau), y_0)$$



$$g(c_\tau) \leftarrow y_0$$

Modification is local (at c_τ):
+ requires only local update
– its effect can be limited

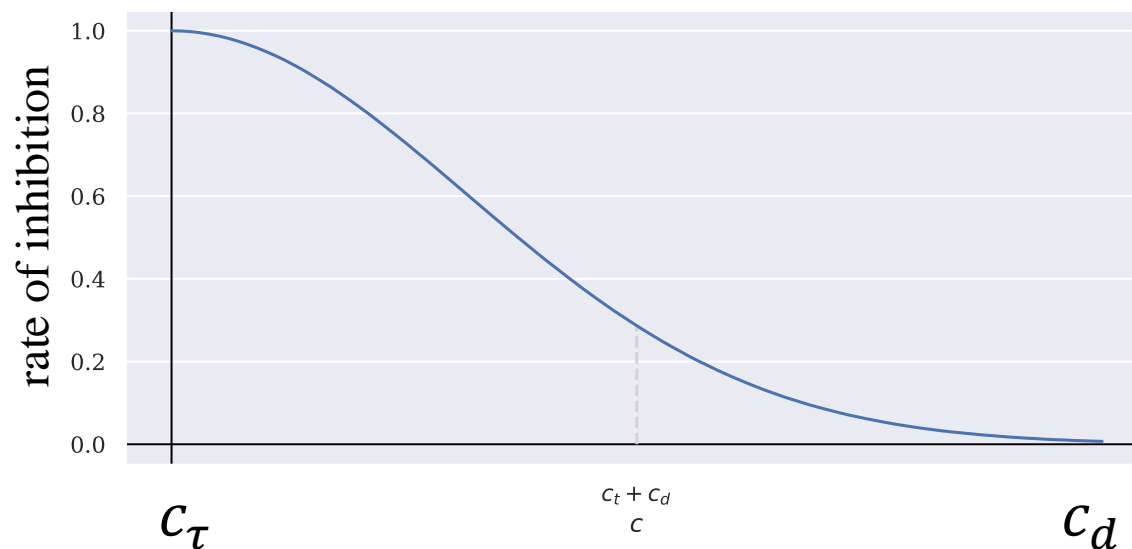
Analyzing Concept Inhibition

Optimize the weights to enforce some predefined output as guidance for target.

$$g(c_\tau) \leftarrow y_0 \quad \text{via optimization} \quad \theta = \arg \min_{\theta} \mathcal{L}(g_{\theta}(c_\tau), y_0)$$

$$g(c_\tau) \leftarrow y_0$$

Modification is local (at c_τ):
+ requires only local update
– **its effect can be limited**



Hypothesis: the rate of inhibition effect decays as the distance from target concept increases.

Background: Concept Composition

Original



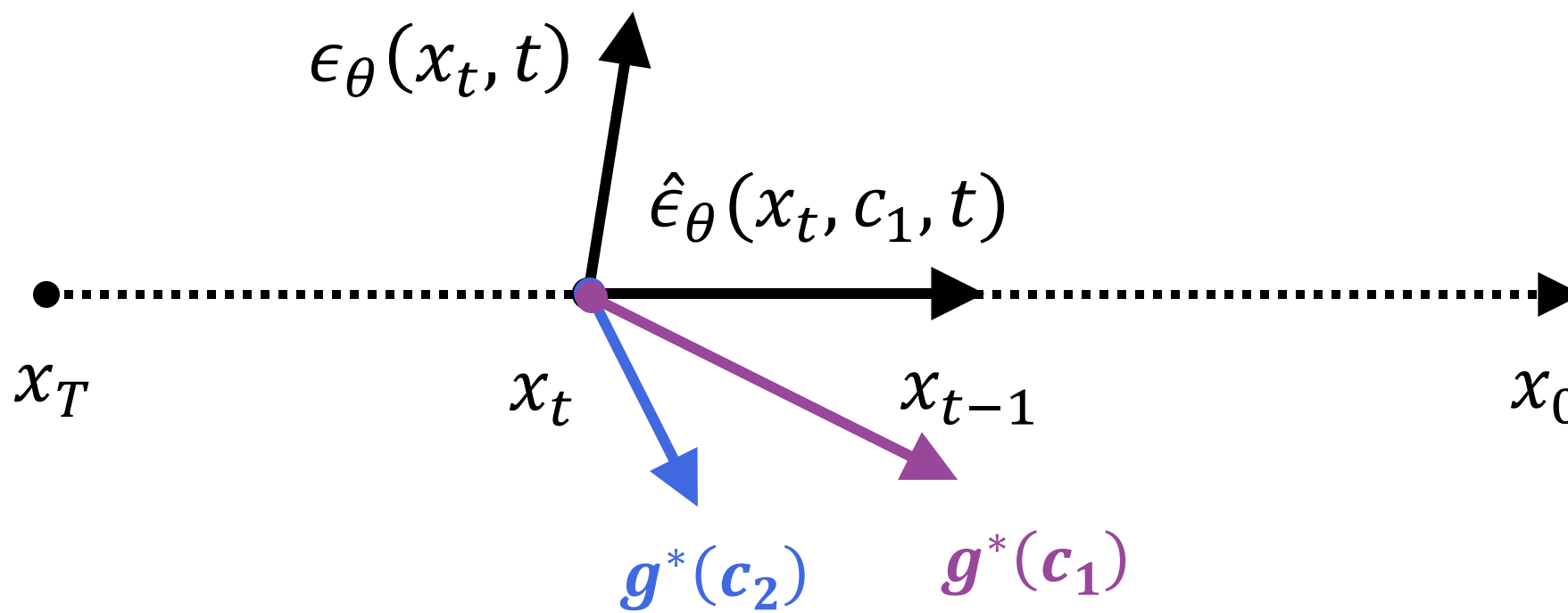
king — 'male' + 'female'



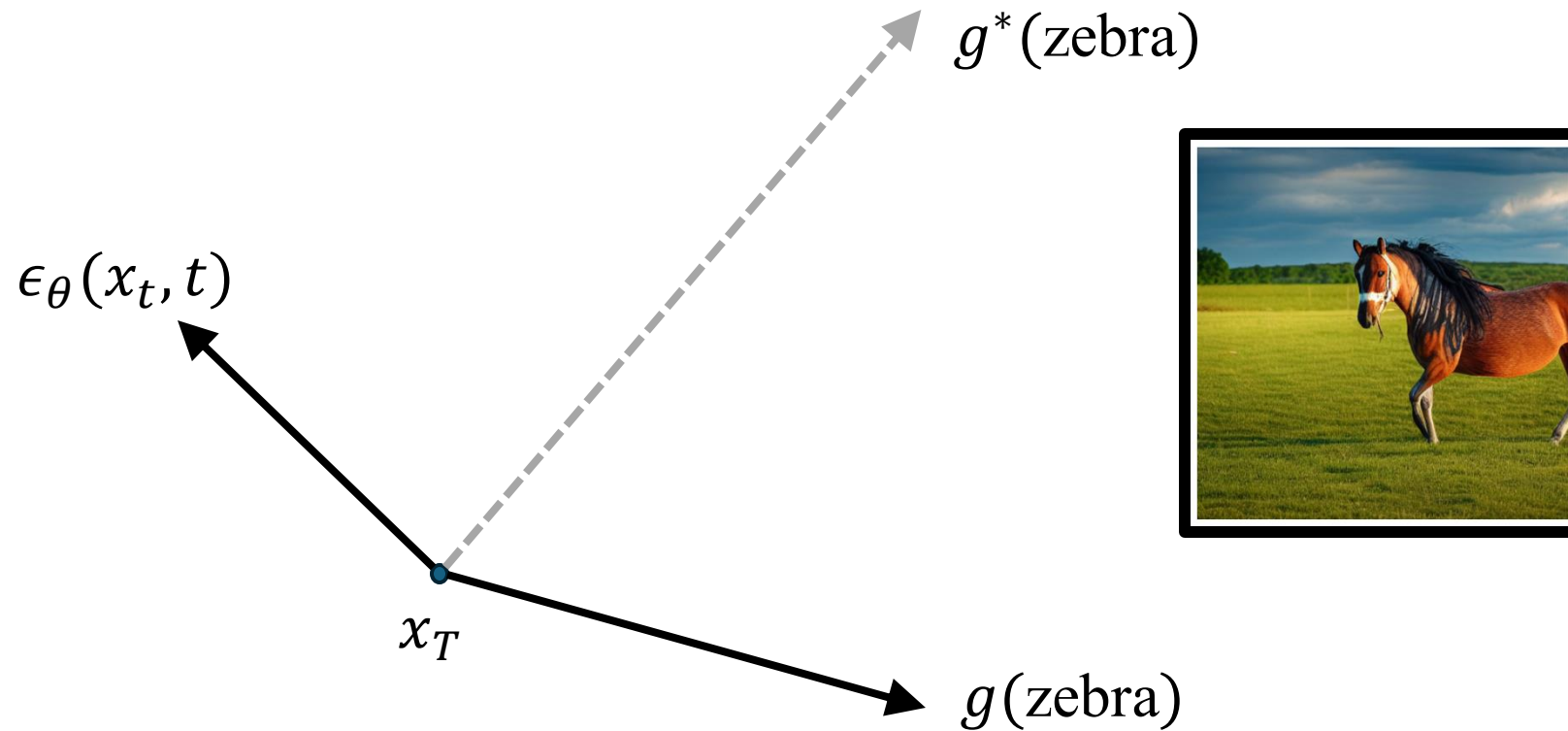
'a portrait of a king'

Background: Concept Composition

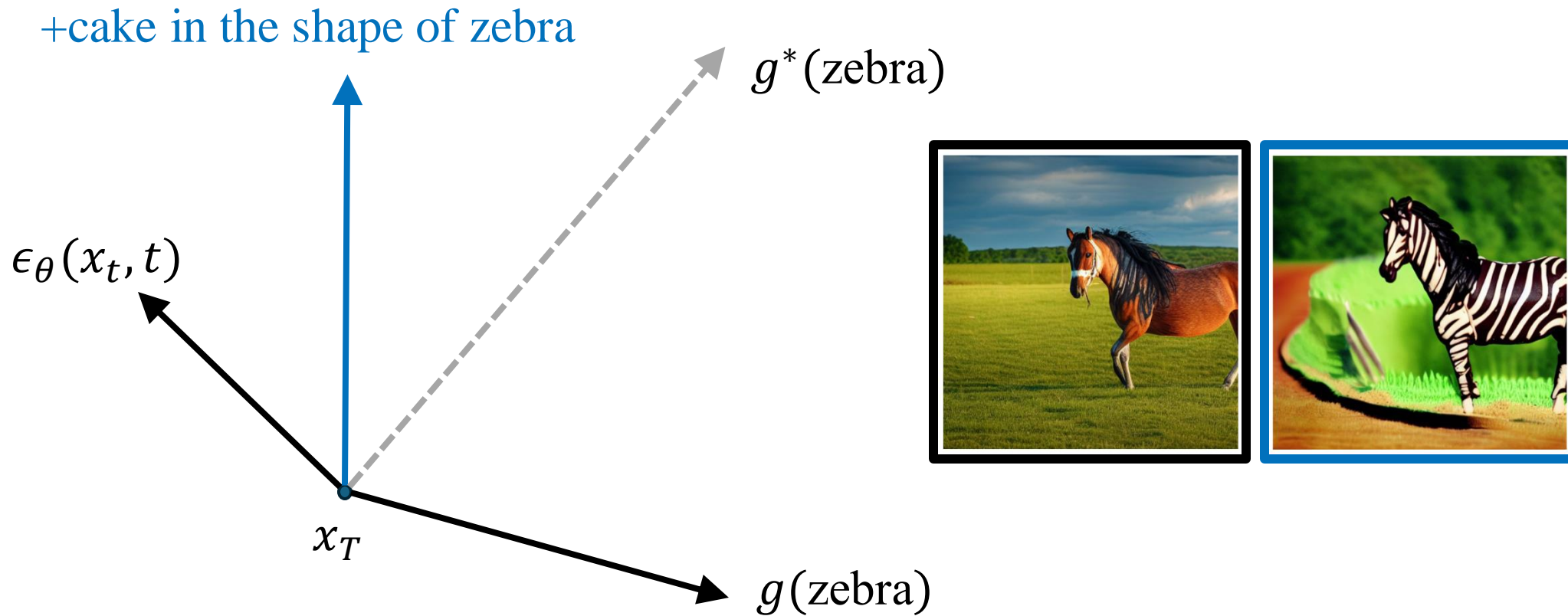
$c_1 =$ A horse standing in the field $c_2 =$ A striped animal



Compositional Inference Attacks

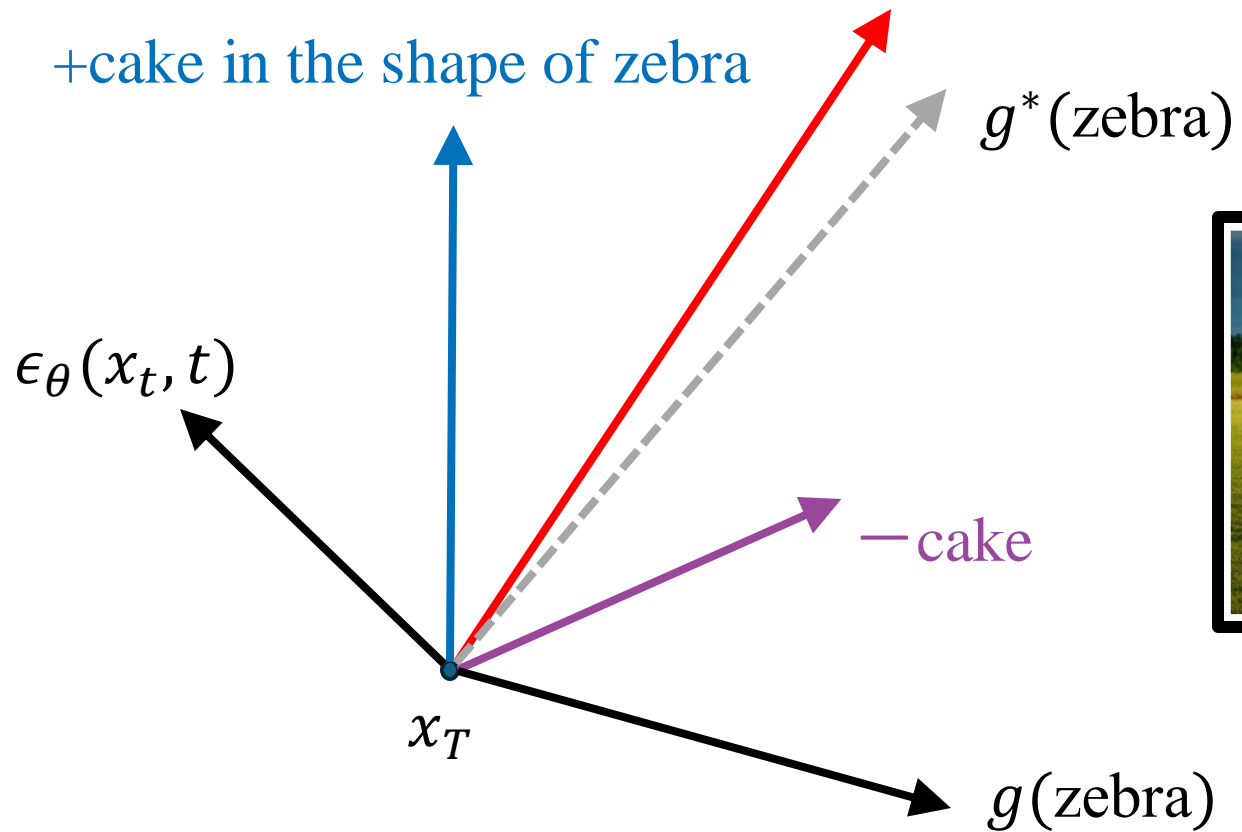


Compositional Inference Attacks



Compositional Inference Attacks

+cake in the shape of zebra - cake



Compositional Inference Attacks

We show that as the distance between some arbitrary distractor concept c_d and inhibited concept c_τ increases, the linear combination(s) of $g()$ can be used to compute a vector colinear with $g^*(c_\tau)$

Proposition P1. If $|c_d - c_\tau| \rightarrow +\infty$ and $g^*(c_\tau \pm c_d) = g^*(c_\tau) \pm g^*(c_d)$, then

$$g(c_\tau \pm c_d) \mp g(c_d) \rightarrow g^*(c_\tau),$$

where \rightarrow denotes convergence in the limit.

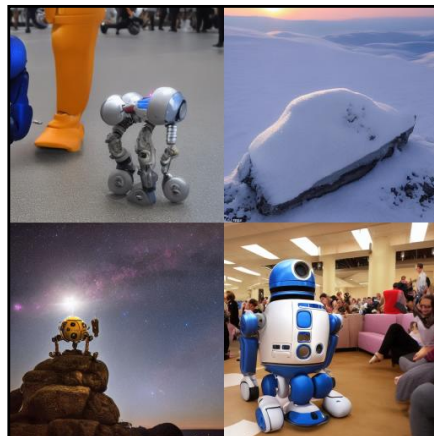
Qualitative Results

Inhibited model

zebra



R2D2



car



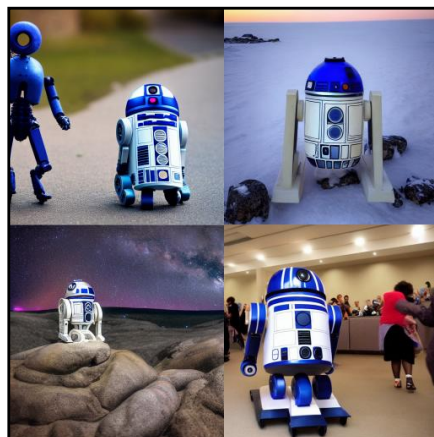
golf ball



Our attack



Inhibited "zebra"
(AC with anchor "horse")



Inhibited "r2d2"
(AC with anchor "robot")



Inhibited "car"
(ESD-u)



Inhibited "golf ball"
(UCE)

Quantitative Results: Nudity Inhibition

Inhibition Methods.

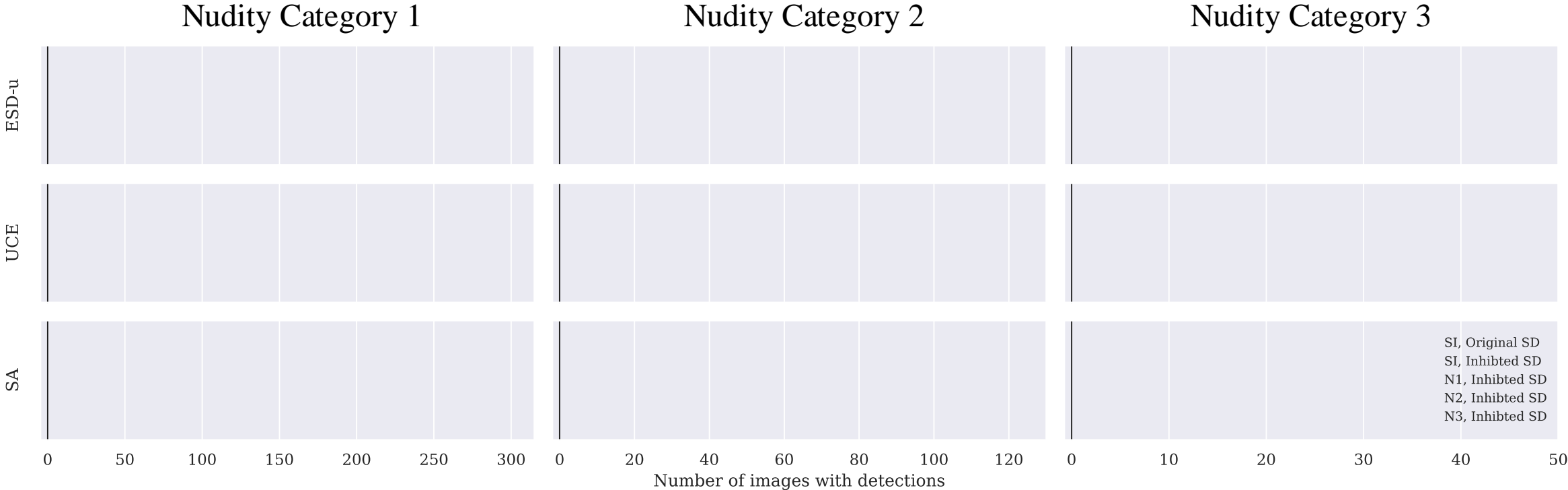
- ESD, UCE, SA

Concept Presence Metric

Number of images with detected nudity category (NudeNet)

Prompts

Inappropriate Image Prompts (I2P)



Quantitative Results: Object Inhibition

Inhibition Methods.

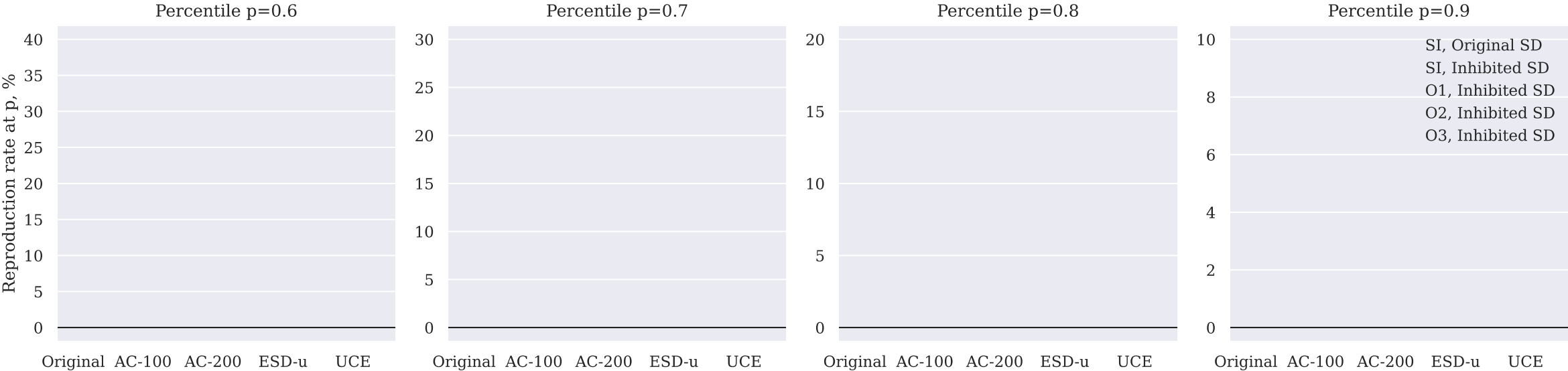
- ESD, UCE, AC

Concept Presence Metric

- CLIP Score based.
- Avg. over 15 concepts**

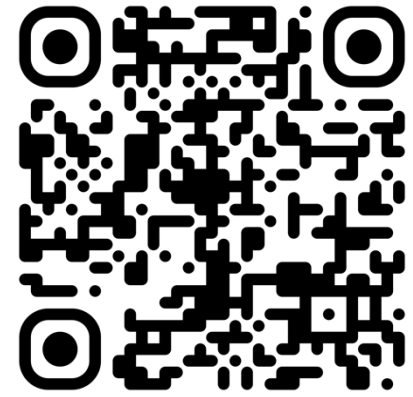
Prompts

Generated with Chat-GPT.



Conclusion

- Localized nature of existing approaches does not erase the information about the concept fully.
- Proposed compositional inference attacks are an efficient way of extracting this information.
- Inhibited models risk being circumvented in this way not only in an open-source scenario, but also via multi-prompt API calls.
- Any editing of diffusion models should take the compositional property into consideration.



Project webpage

<https://cs-people.bu.edu/vpetsiuk/arc>



On the job market!

