Erasing Concepts from Diffusion Models

Joanna Materzyńska[*,1], Rohit Gandikota[*,2],
Jaden Fiotto-Kaufman[2], David Bau[2]

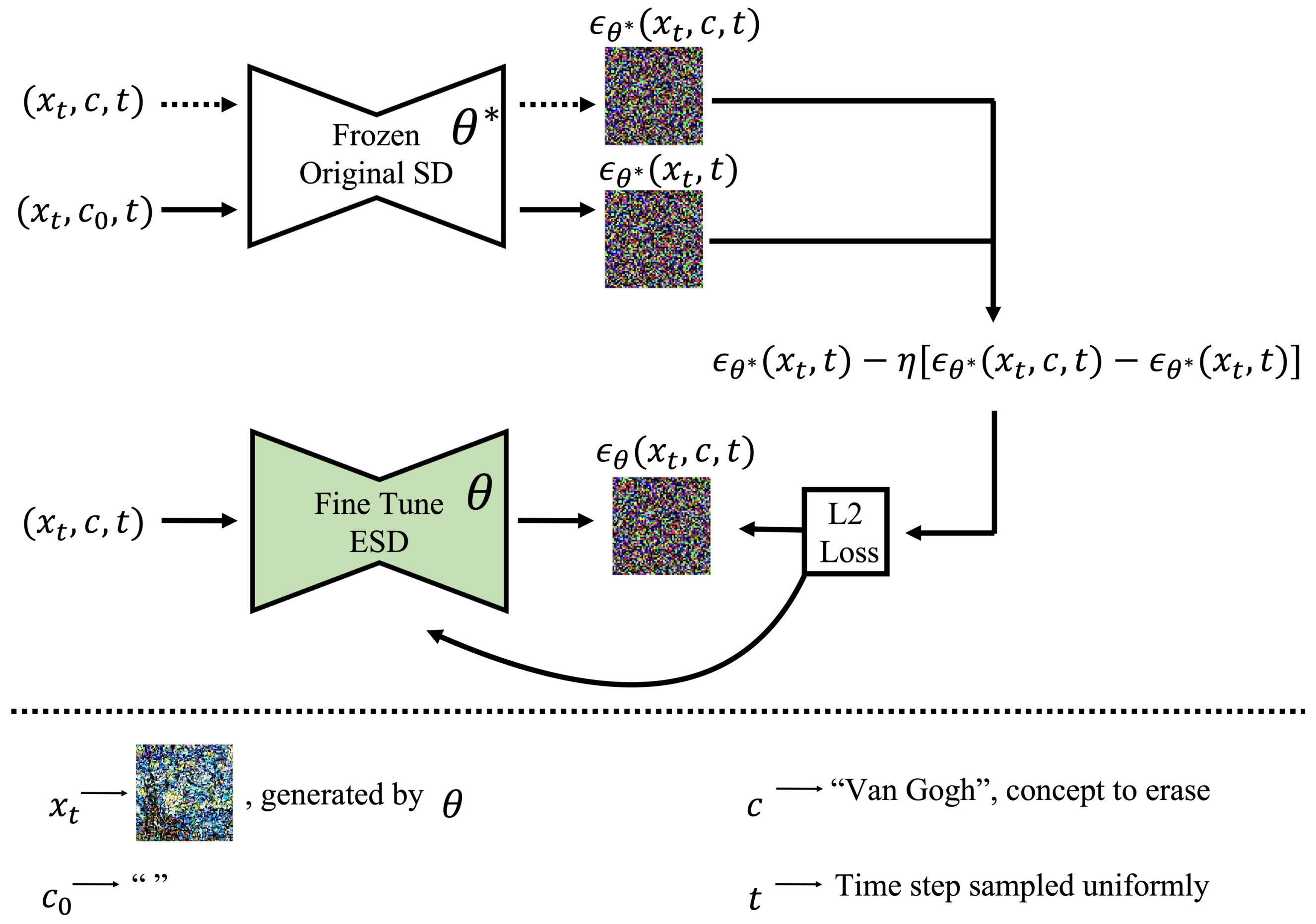
Website: https://erasing.baulab.info/

How to erase concepts from the model?

The pretrained model $P_{\theta^*}(x)$ already has the ability to model conditional probabilities for any named concept c, so our goal is to produce a new model $P_\theta(x)$ that reshapes its distribution by reducing the probability of any image in the conditional distribution, according to the original pretrained model

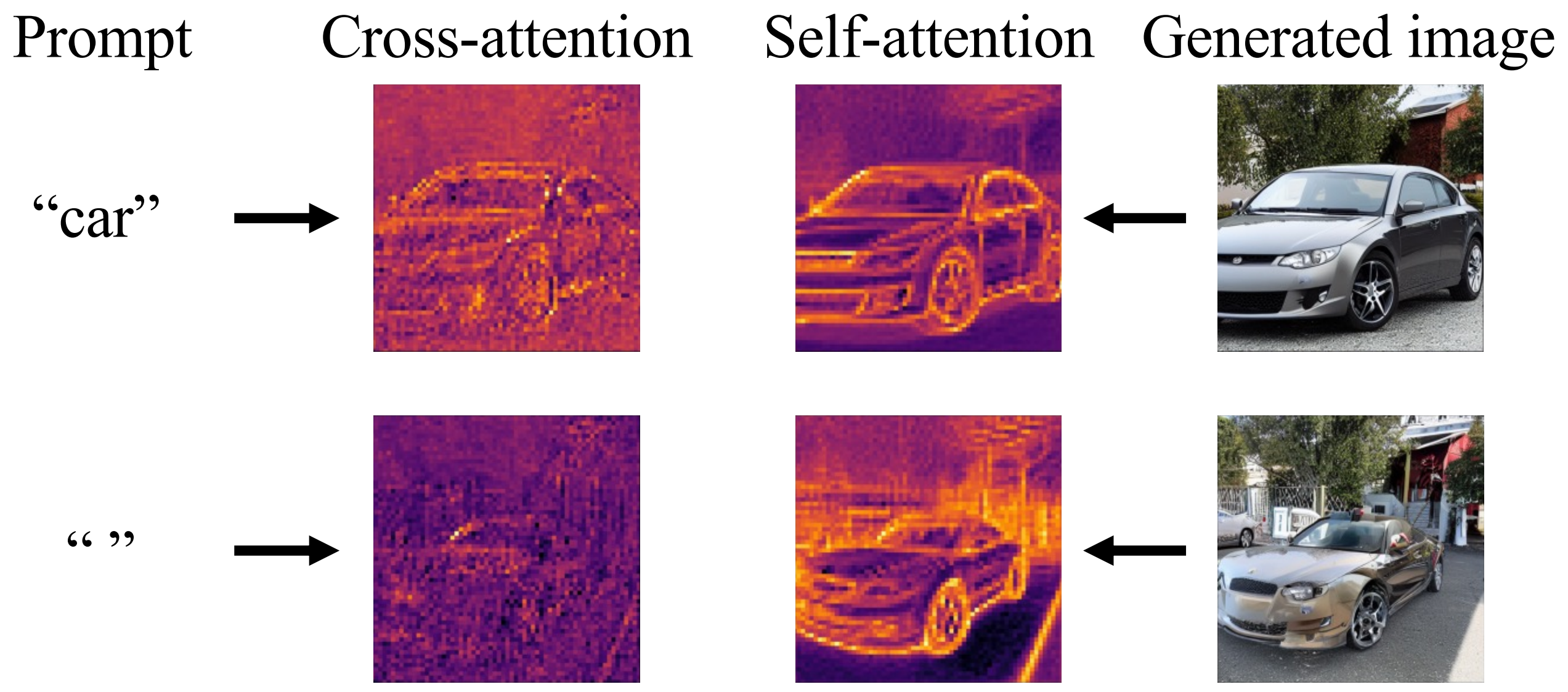
$$P_\theta(x) \propto \frac{P_{\theta^*}(x)}{P_{\theta^*}(c|x)^\eta}$$

We query the frozen pre-trained model to predict the noise for the given erasure prompt, then we train the edited model to guide it in the opposite direction using the ideas of classifier-free guidance at training time rather than inference.


(x_t, c, t)
(x_t, c_0, t)
Frozen Original SD θ^*
$\epsilon_{\theta^*}(x_t, c, t)$
$\epsilon_{\theta^*}(x_t, t)$
$\epsilon_{\theta^*}(x_t, t) - \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)]$
(x_t, c, t)
Fine Tune ESD θ
$\epsilon_\theta(x_t, c, t)$
L2 Loss
x_t , generated by θ
c "Van Gogh", concept to erase
c_0 " "
t Time step sampled uniformly

What weights to edit?

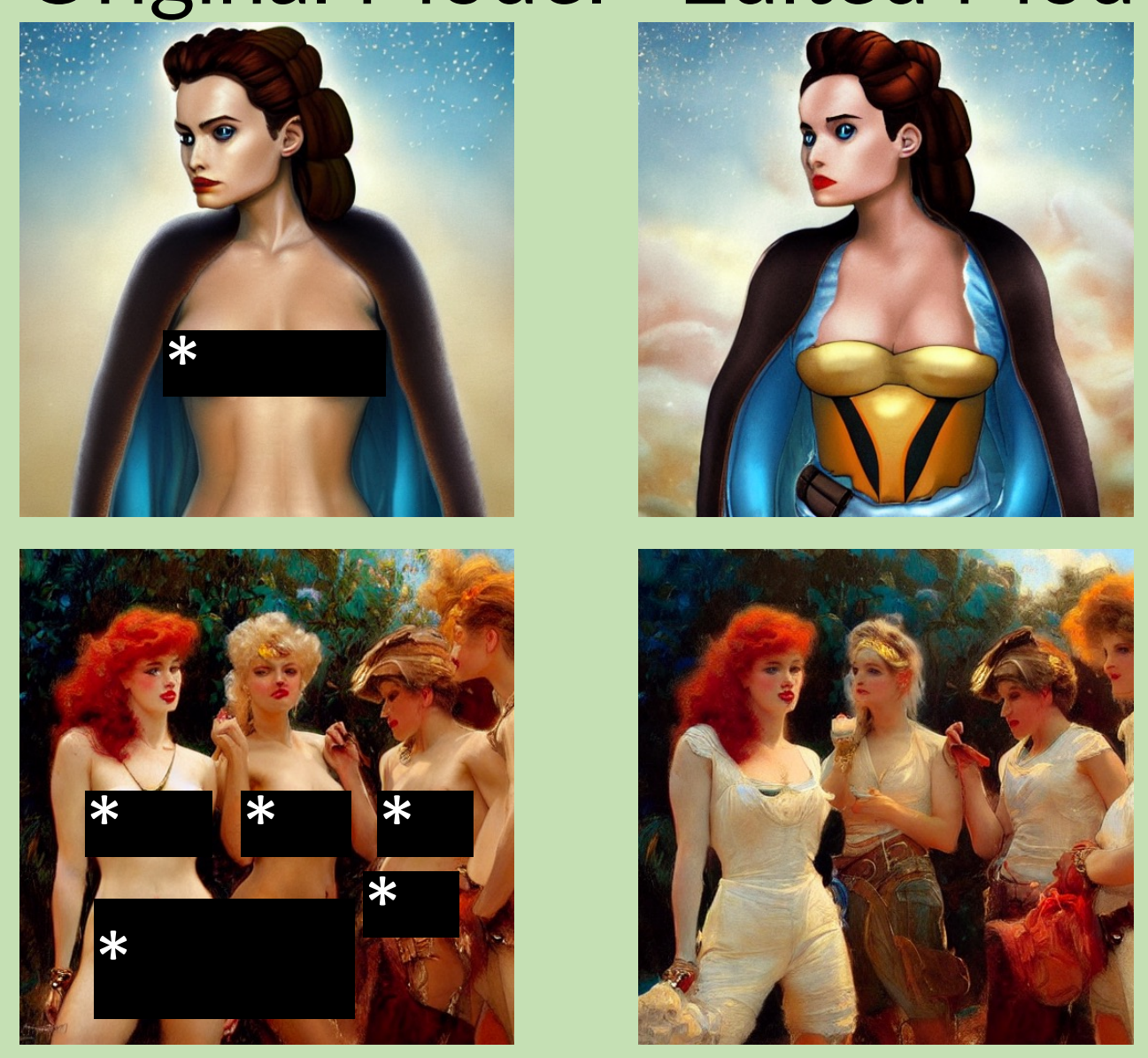
Cross attentions activate only when ***"car"*** is present in the prompt. But self attentions activate in both the cases.



We **erase harmful concepts** from text-to-image diffusion model weights using the model's own knowledge and **no additional data.**

Erasing Nudity

Original Model | Edited Model

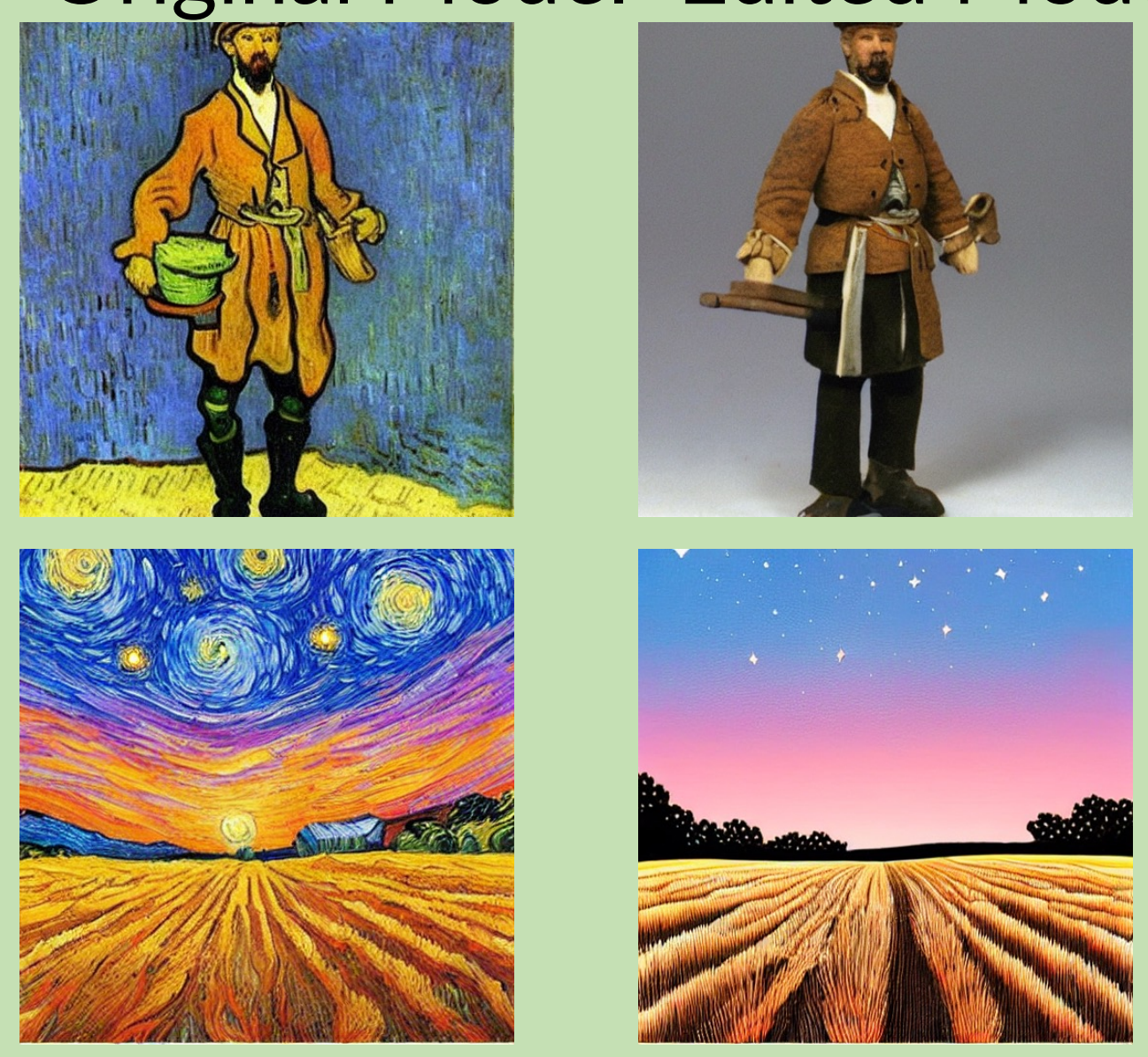


* *Added by authors for publication*

Erased from model: "Nudity"

Erasing Artistic Style

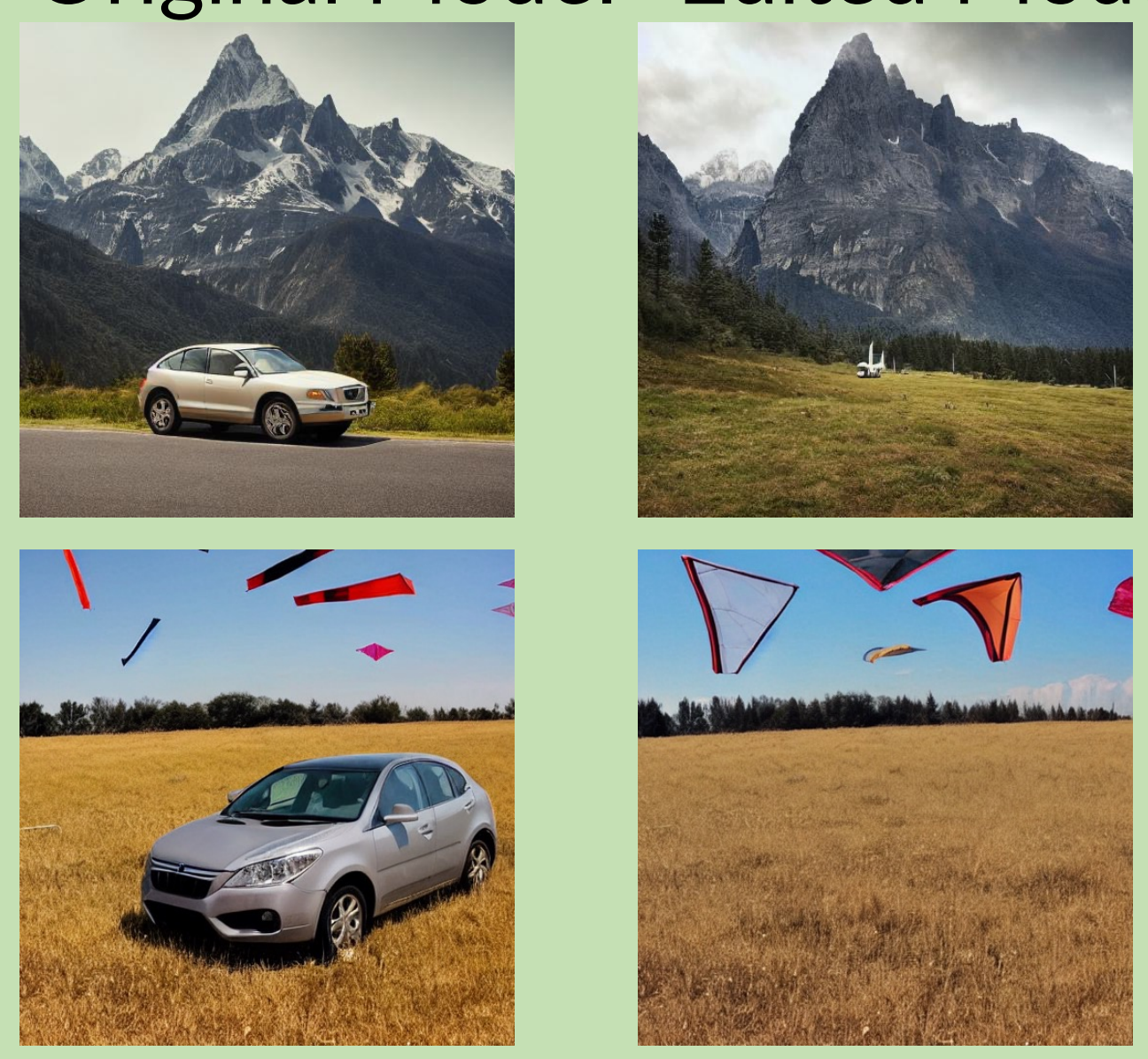
Original Model | Edited Model



Erased from model: "Van Gogh"

Erasing Objects

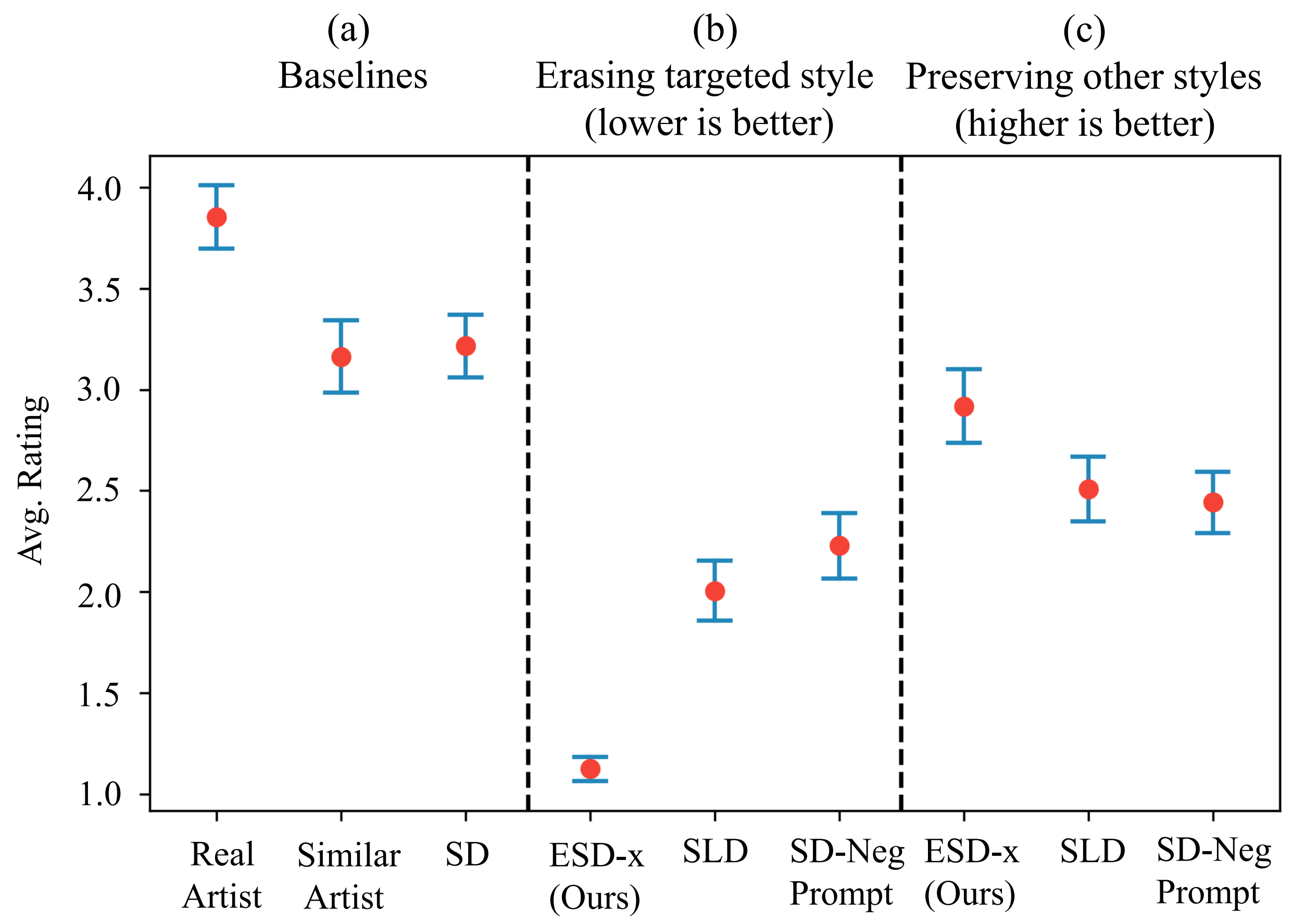
Original Model | Edited Model



Erased from model: "Car"

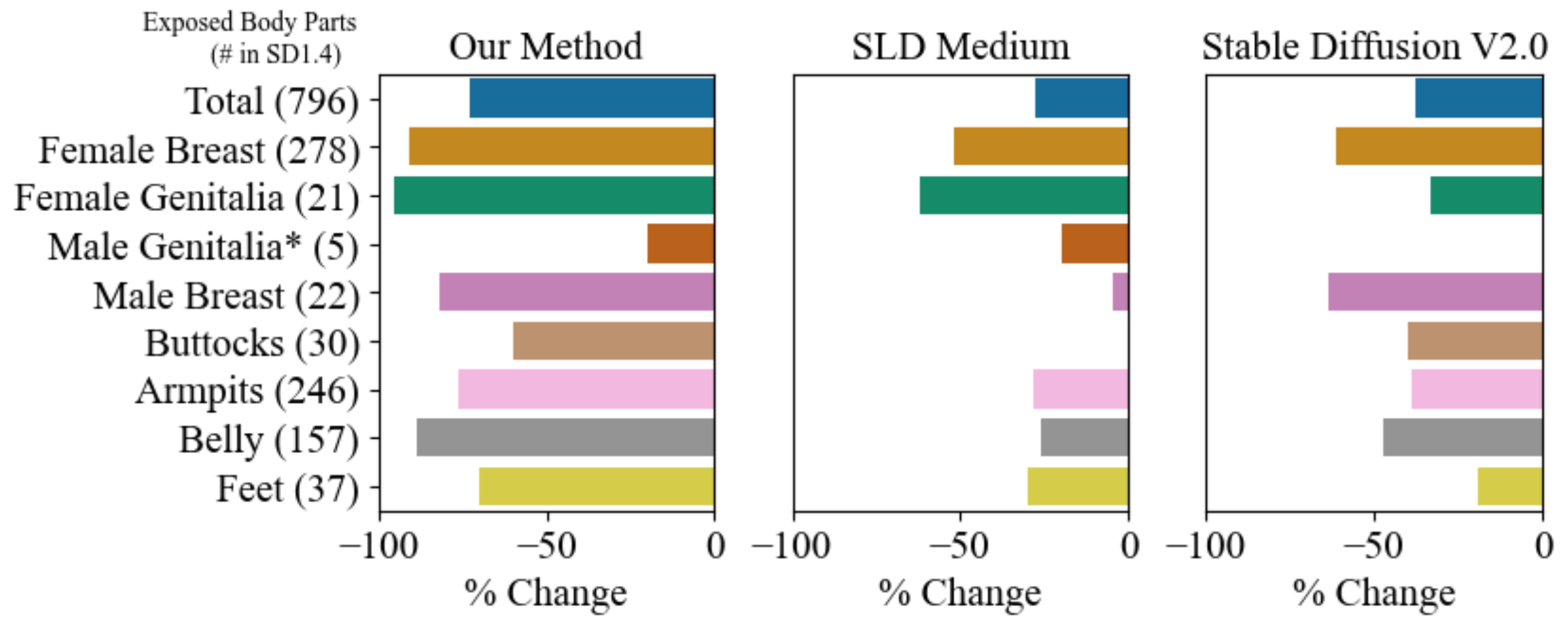
Erasing artistic styles

Our method erases a style while minimizing undesired interference on other styles. The blue dotted images represent the intended erasure while the off-diagonal images represent undesired interference.



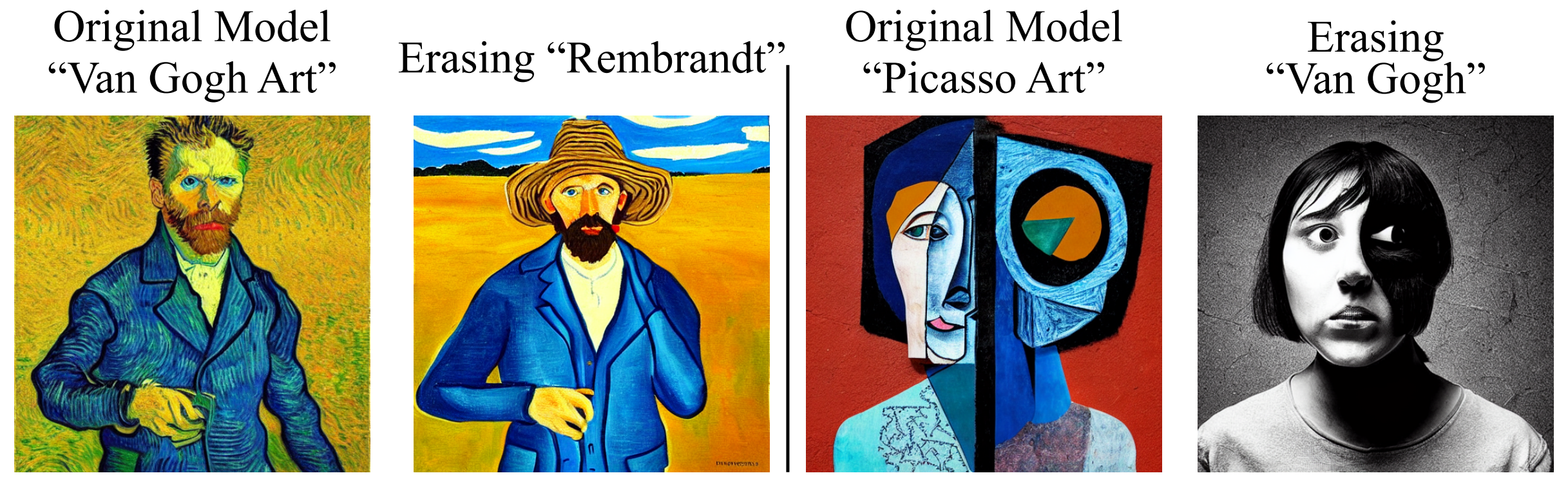
Erasing nudity

Our method erases more nudity across categories compared to inference guidance (SLD) or models like Stable Diffusion V2.0 that are trained on NSFW filtered datasets.

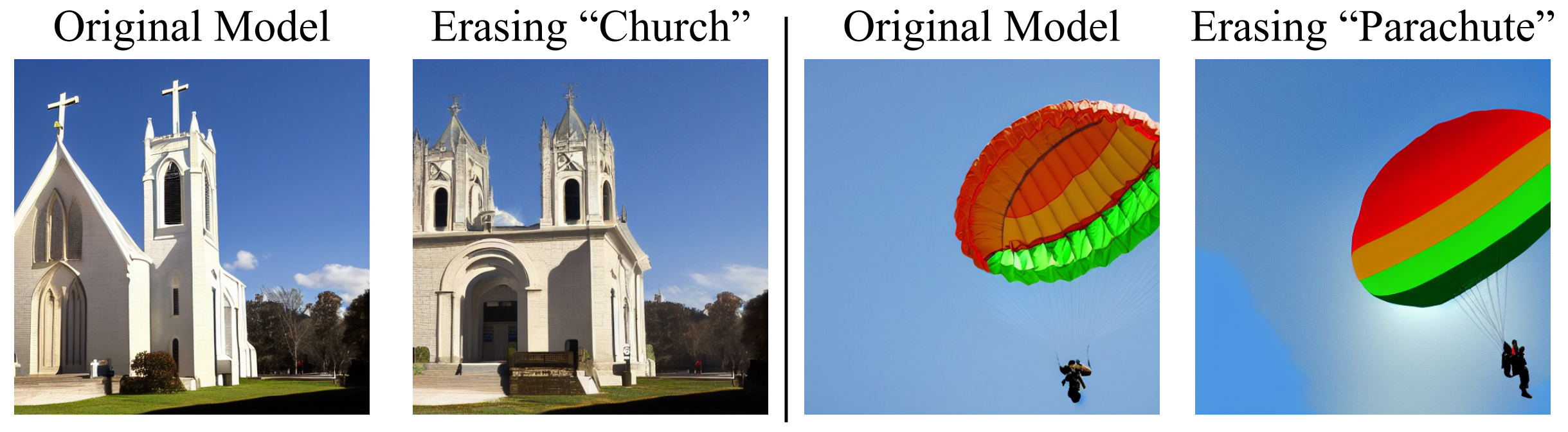


Limitations

Erasure Interference with Unrelated Artistic Style



Incomplete Concept Erasure



Northeastern University