Training-Free Safe Text Embedding Guidance for Text-to-Image Diffusion Models

Byeonghu Na[1], Mina Kang[1], Jiseok Kwak[1], Minsang Park[1], Jiwoo Shin[1], SeJoon Jun[1], Gayoung Lee[2], Jin-Hwa Kim[2,3], Il-Chul Moon[1,4]

1 KAIST

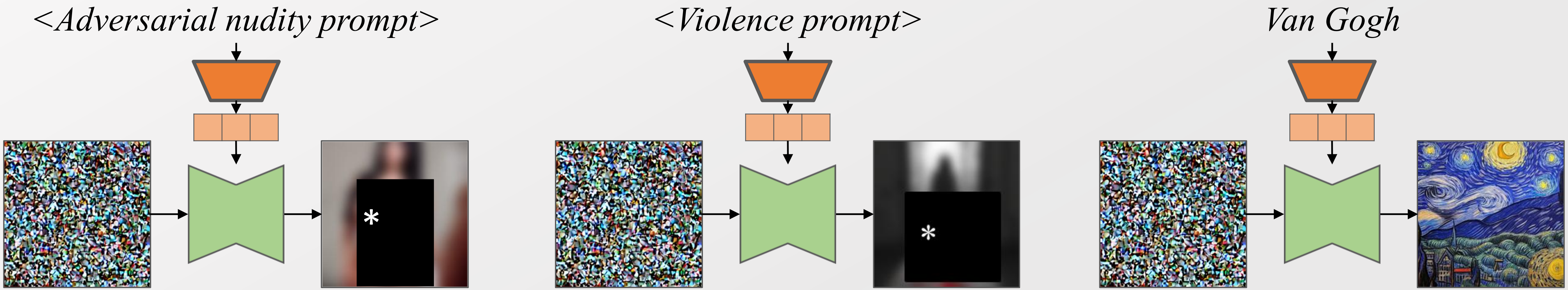
2 NAVER AI LAB

3 AIIS Artificial Intelligence Institute Seoul National University

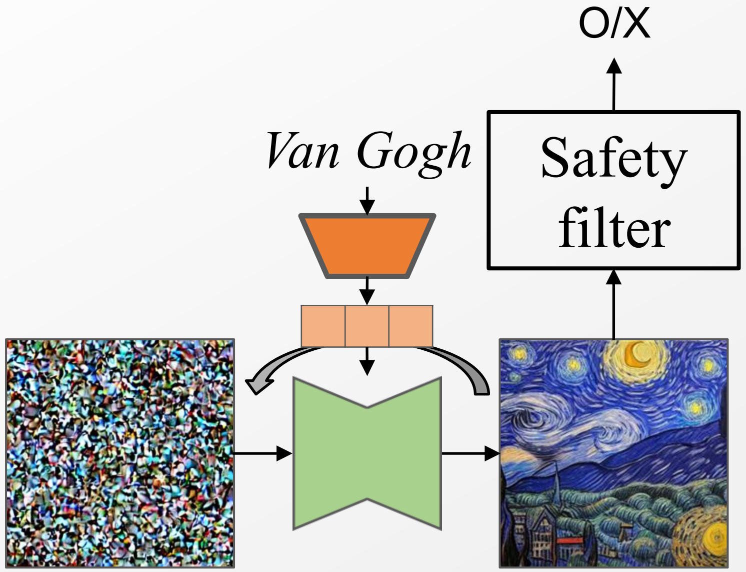
4 summary.ai

Safe Text-to-Image Generation

- Recent advances in text-to-image models raise concerns about unsafe or biased content.
- What is considered *safe* can vary widely depending on individual sensitivities, cultural contexts, and social norms, making it challenging to define a universally safe model.
- ➔ Need for safe generation methods that can adapt to diverse perspectives.

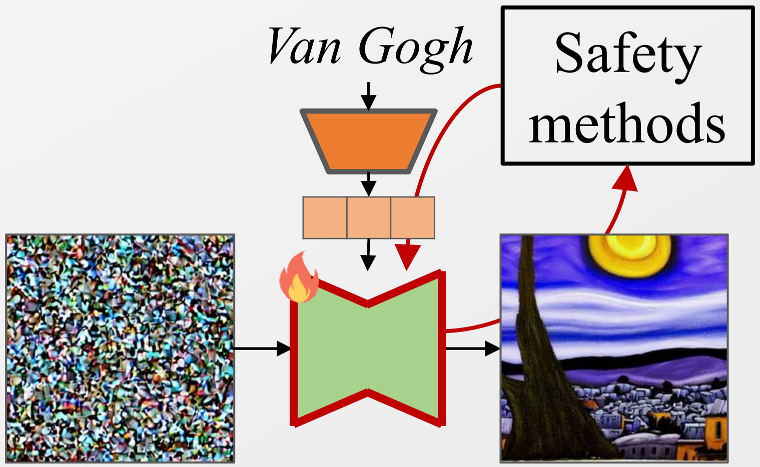


Previous Work



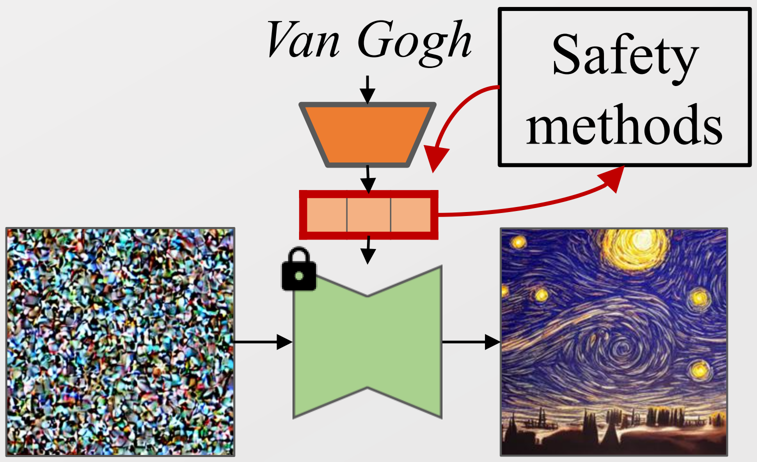
Post-hoc filtering

- Exclude unsafe images after final generation.
- Significantly increases generation time due to filtering.



Training-based approach

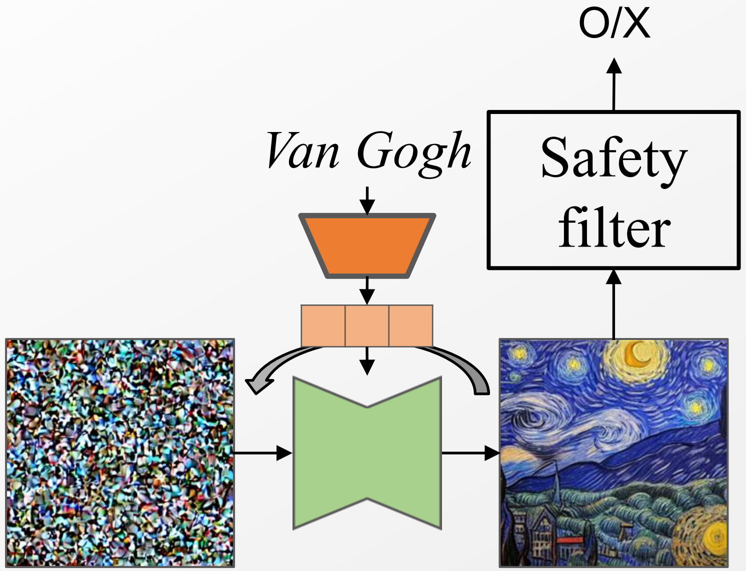
- Fine-tune the diffusion model.
- Difficult to preserve the original generative capability.
- Require curated safety-annotated datasets and high computational cost.



Training-free approach (previous)

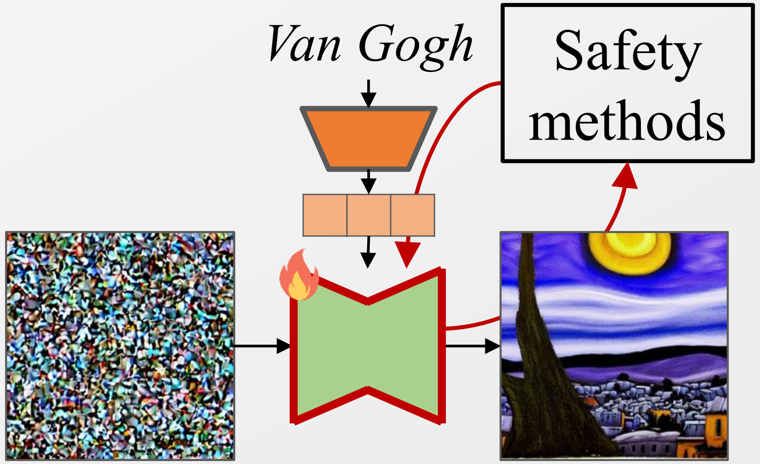
- Manipulate inputs or intermediate representations during inference.
- Do not directly utilize intermediate samples from diffusion models.

Safe Text Embedding Guidance



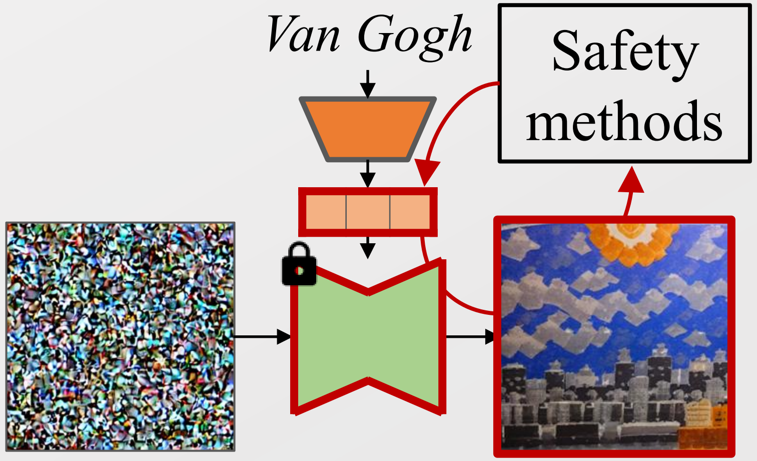
Post-hoc filtering

- Exclude unsafe images after final generation.
- Significantly increases generation time due to filtering.



Training-based approach

- Fine-tune the diffusion model.
- Difficult to preserve the original generative capability.
- Require curated safety-annotated datasets and high computational cost.

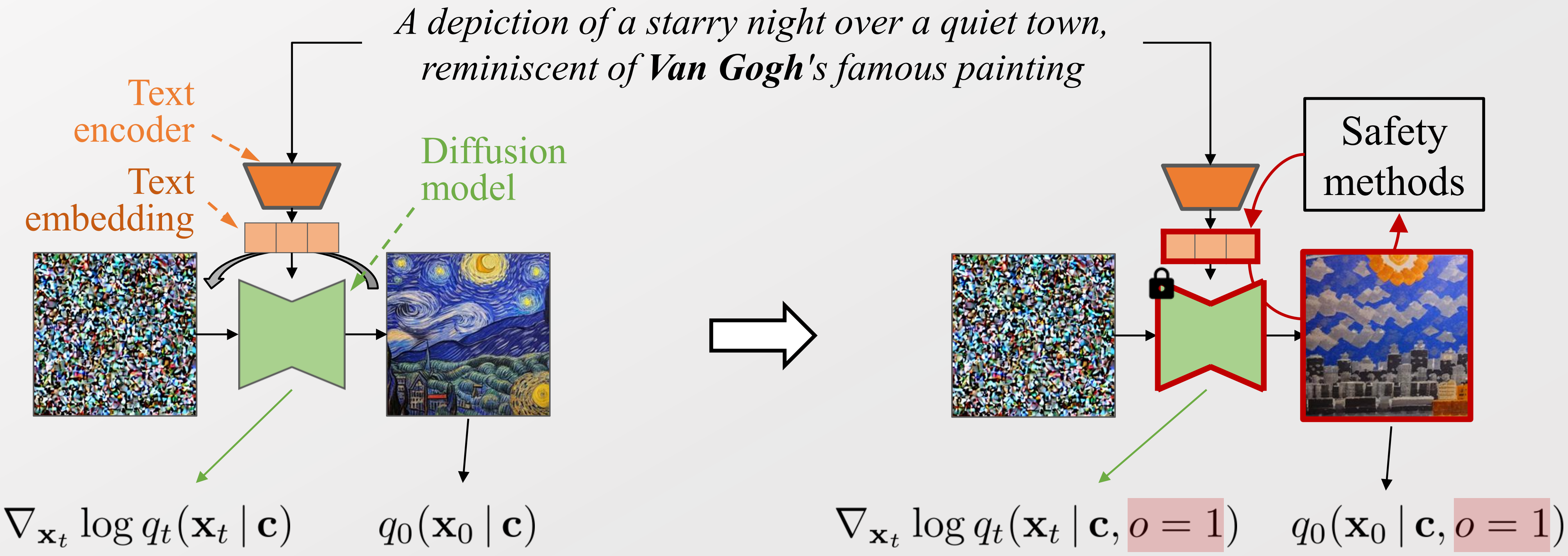


Training-free approach (ours)

- Guide text embeddings in safer directions during inference.
- Directly incorporates intermediate samples for guidance

Safe Text-to-Image Diffusion Models

- Generate samples from the safe text-conditional distribution $q_0(x_0|c, o=1)$
  - Samples that both align with the text condition c and satisfy the safety criterion $o=1$.
  - Define a binary random variable $o \in \{0,1\}$ as a safety indicator, where $o=1$ denotes a safe sample.
- To sample from this safe distribution using a diffusion model, we need the safe text-conditional score function $\nabla_{x_t} \log q_t(x_t|c, o=1)$.


A depiction of a starry night over a quiet town,
*reminiscent of **Van Gogh**'s famous painting*
Text encoder
Text embedding
Diffusion model
Safety methods
$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c})$
$q_0(\mathbf{x}_0 \mid \mathbf{c})$
$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c}, o=1)$
$q_0(\mathbf{x}_0 \mid \mathbf{c}, o=1)$

Guidance Methods in Diffusion Models

- We formulate a framework that extends the existing guidance methods to safe generation.

Method	Guidance framework	Guidance target	Guidance module
SLD (Schramowski et al., 2023)	Classifier-Free Guidance (Ho and Salimans, 2021)	Perturbed data	Unsafe-conditional score network
Safe Guidance (SG)	Classifier Guidance (Dhariwal and Nichol, 2021)	Perturbed data	Time-dependent classifier
Safe Data Guidance (SDG)	Universal Guidance (Bansal et al., 2023)	Perturbed data	Time-independent classifier
Safe Text embedding Guidance (STG, ours)	Diffusion Adaptive Text Embedding (Na et al., 2025)	Text embedding	Time-independent classifier

(Schramowski et al., 2023) Schramowski, P., Brack, M., Deiseroth, B., & Kersting, K. (2023). Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
(Ho and Salimans, 2021) Ho, J., & Salimans, T. Classifier-Free Diffusion Guidance. In NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.
(Dhariwal and Nichol, 2021) Dhariwal, P., & Nichol, A. (2021). Diffusion models beat gans on image synthesis. Advances in neural information processing systems, 34.
(Bansal et al., 2023) Bansal, A., Chu, H. M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., & Goldstein, T. (2023). Universal guidance for diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.
(Na et al., 2025) Na, B., Park, M., Sim, G., Shin, D., Bae, H., Kang, M., … Moon, I.-C. (2025). Diffusion Adaptive Text Embedding for Text-to-Image Diffusion Models. The Thirty-Ninth Annual Conference on Neural Information Processing Systems.

Safe Guidance

- Sample from the safe text-conditional distribution without modifying the model parameters
  ➔ by using guidance methods!

- Safety function $g: \mathbb{R}^d \rightarrow \mathbb{R}$
  - Evaluate whether a clean image x_0 is safe or not.
  - This function can be implemented using open-source classifiers (e.g., NudeNet for nudity), or using pre-trained vision-language models (e.g., CLIP).

- Safe Guidance (SG)
  - Use an external time-dependent safety function.
  - Cannot directly apply the safety function g which operates only on fully denoised images.

$$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, \mathbf{c}, o = 1) = \underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \,|\, \mathbf{c})}_{\text{original text-conditional score}} + \underbrace{\nabla_{\mathbf{x}_t} \log q_t(o = 1 | \, \mathbf{x}_t, \mathbf{c})}_{\text{safe guidance}}$$

Safe Data Guidance

Safe Guidance (SG)

$$\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t\,|\,\mathbf{c},o=1)=\underbrace{\nabla_{\mathbf{x}_t}\log q_t(\mathbf{x}_t\,|\,\mathbf{c})}_{\text{original text-conditional score}}+\underbrace{\nabla_{\mathbf{x}_t}\log q_t(o=1|\,\mathbf{x}_t,\mathbf{c})}_{\text{safe guidance}}$$

- Safe Data Guidance (SDG)
  - Assume that the safety function g is proportional to the safe probability distribution $q(o=1|x_0)$.
  - Then, the safe guidance term can be derived as follows:

$$\begin{aligned}
\nabla_{\mathbf{x}_t}\log q_t(o=1|\,\mathbf{x}_t,\mathbf{c}) &= \nabla_{\mathbf{x}_t}\log \mathbb{E}_{\mathbf{x}_0\sim q(\mathbf{x}_0\,|\,\mathbf{x}_t,\mathbf{c})}[q(o=1|\,\mathbf{x}_0)] \\
\textit{Assumption}\longrightarrow\quad &= \nabla_{\mathbf{x}_t}\log \mathbb{E}_{\mathbf{x}_0\sim q(\mathbf{x}_0\,|\,\mathbf{x}_t,\mathbf{c})}[g(\mathbf{x}_0)] =: g_t(\mathbf{x}_t,\mathbf{c}) \quad \textit{time-dependent safety function} \\
\textit{First-order Taylor approx.}\longrightarrow\quad &\approx \nabla_{\mathbf{x}_t}\log g(\mathbb{E}_{\mathbf{x}_0\sim q(\mathbf{x}_0\,|\,\mathbf{x}_t,\mathbf{c})}[\mathbf{x}_0]) \\
\textit{Tweedie's formula}\longrightarrow\quad &\approx \nabla_{\mathbf{x}_t}\log g\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t+(1-\bar{\alpha}_t)\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{c},t)\right)\right)
\end{aligned}$$

Safe Data Guidance (SDG)

$$\mathbf{s}_{\mathrm{SDG}}(\mathbf{x}_t,\mathbf{c},t):=\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{c},t)+\nabla_{\mathbf{x}_t}\log g\left(\frac{1}{\sqrt{\bar{\alpha}_t}}\left(\mathbf{x}_t+(1-\bar{\alpha}_t)\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t,\mathbf{c},t)\right)\right)$$

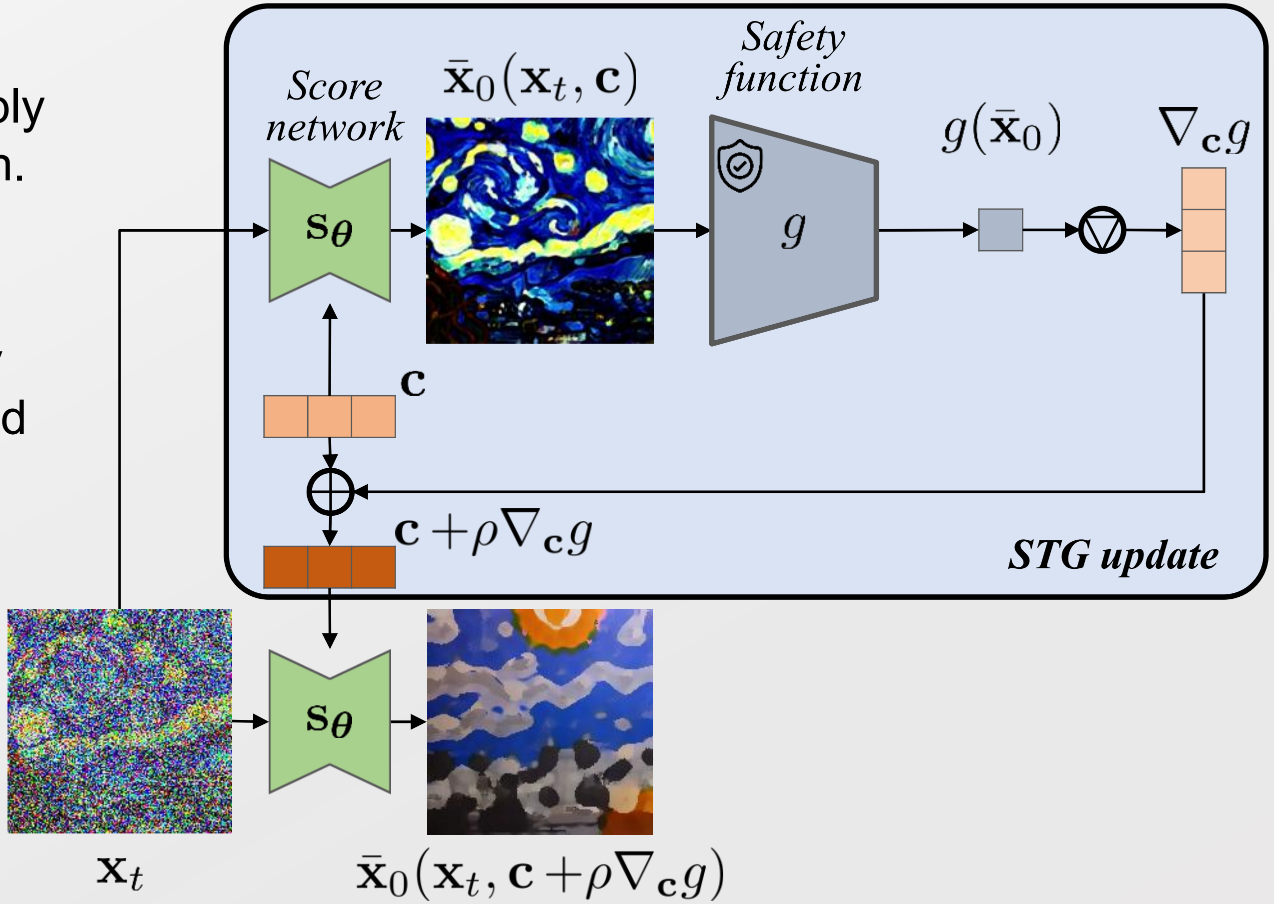
Safe Text Embedding Guidance

- Safe Text embedding Guidance (STG)
  - Alternative to direct guidance in the data space, apply guidance to text embedding using the safey function.
  - Adjust the text embedding c toward a safer representation.
  - Apply gradient ascent on the time-dependent safety function g_t, then the updated text embedding is used in the score network to perform diffusion sampling.

$$\mathbf{c} \leftarrow \mathbf{c} + \rho \nabla_{\mathbf{c}} g_t(\mathbf{x}_t, \mathbf{c})$$

Same logic of the previous derivation

$$\approx \nabla_{\mathbf{c}} g\left(\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t + (1-\bar{\alpha}_t)\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{c}, t))\right)$$



Score network
$\bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c})$
Safety function
$\mathbf{s}_{\boldsymbol{\theta}}$
g
$g(\bar{\mathbf{x}}_0)$
$\nabla_{\mathbf{c}} g$
$\mathbf{c}$
$\mathbf{c} + \rho \nabla_{\mathbf{c}} g$
STG update
$\mathbf{x}_t$
$\mathbf{s}_{\boldsymbol{\theta}}$
$\bar{\mathbf{x}}_0(\mathbf{x}_t, \mathbf{c} + \rho \nabla_{\mathbf{c}} g)$

Safe Text embedding Guidance (STG)

$$\mathbf{s}_{\mathrm{STG}}(\mathbf{x}_t, \mathbf{c}, t) := \mathbf{s}_{\boldsymbol{\theta}}\left(\mathbf{x}_t, \mathbf{c} + \rho \nabla_{\mathbf{c}} g\left(\frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t + (1-\bar{\alpha}_t)\,\mathbf{s}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{c}, t))\right), t\right)$$

Analysis of STG on Data Space

- STG can be interpreted as a form of SG!
  - Set the safe probability for intermediate samples by aligning the underlying model likelihood with the desired safety objective.
  - Simultaneously preserve the original distribution of the base model and guides the generation toward safer outputs.

Safe Guidance (SG)

$$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c}, o = 1) = \underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c})}_{\text{original text-conditional score}} + \underbrace{\nabla_{\mathbf{x}_t} \log q_t(o = 1 \mid \mathbf{x}_t, \mathbf{c})}_{\text{safe guidance}}$$

Safe Text embedding Guidance (STG)

$$\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c} + \rho \nabla_{\mathbf{c}} g_t(\mathbf{x}_t, \mathbf{c})) = \underbrace{\nabla_{\mathbf{x}_t} \log q_t(\mathbf{x}_t \mid \mathbf{c})}_{\text{original text-conditional score}} + \underbrace{\nabla_{\mathbf{x}_t} \{\rho \nabla_{\mathbf{c}} g_t(\mathbf{x}_t, \mathbf{c})^T \nabla_{\mathbf{c}} \log q_t(\mathbf{x}_t \mid \mathbf{c})\}}_{\text{safe guidance}} + O(\rho^2)$$

⇩

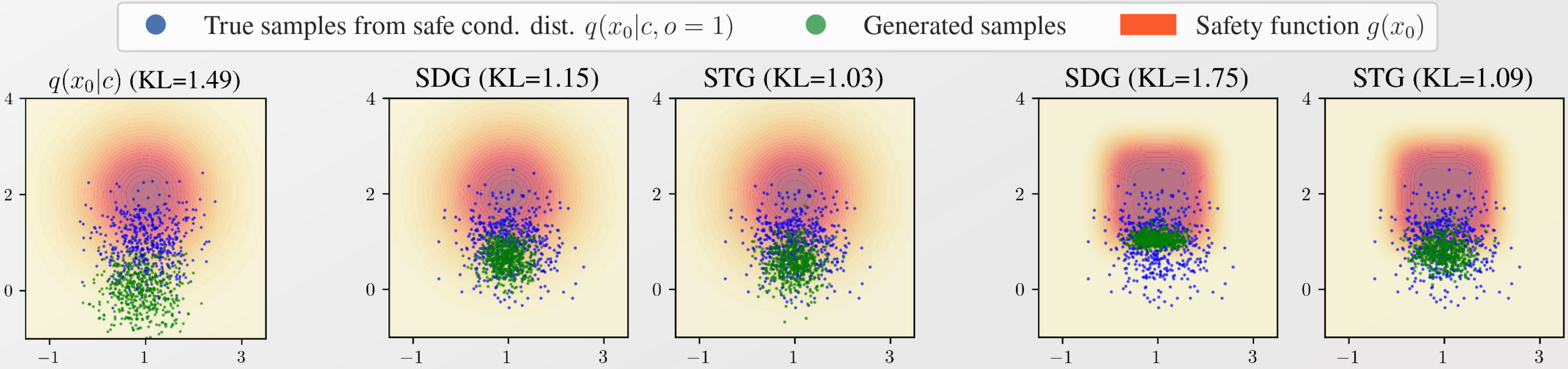
$$q_t^{\mathrm{STG}}(o = 1 \mid \mathbf{x}_t, \mathbf{c}) \underset{\sim}{\propto} \exp\left(\rho \nabla_{\mathbf{c}} g_t(\mathbf{x}_t, \mathbf{c})^T \nabla_{\mathbf{c}} \log q_t(\mathbf{x}_t \mid \mathbf{c})\right)$$

Safety function *Text-conditional likelihood*

Comparison between SDG and STG

- Using the ***ideal*** safety function (proportional to the true safe distribution)
  - Both SDG and STG effectively guide samples to the correct safe region without bias.
- Using the ***approximate*** safety function (preserve the relative ordering but deviate in its shape)
  - SDG produces biased samples, while STG shows more robust performance.

➔ STG generates samples that better preserve the underlying model distribution, reducing mode collapse and improving overall sample quality.

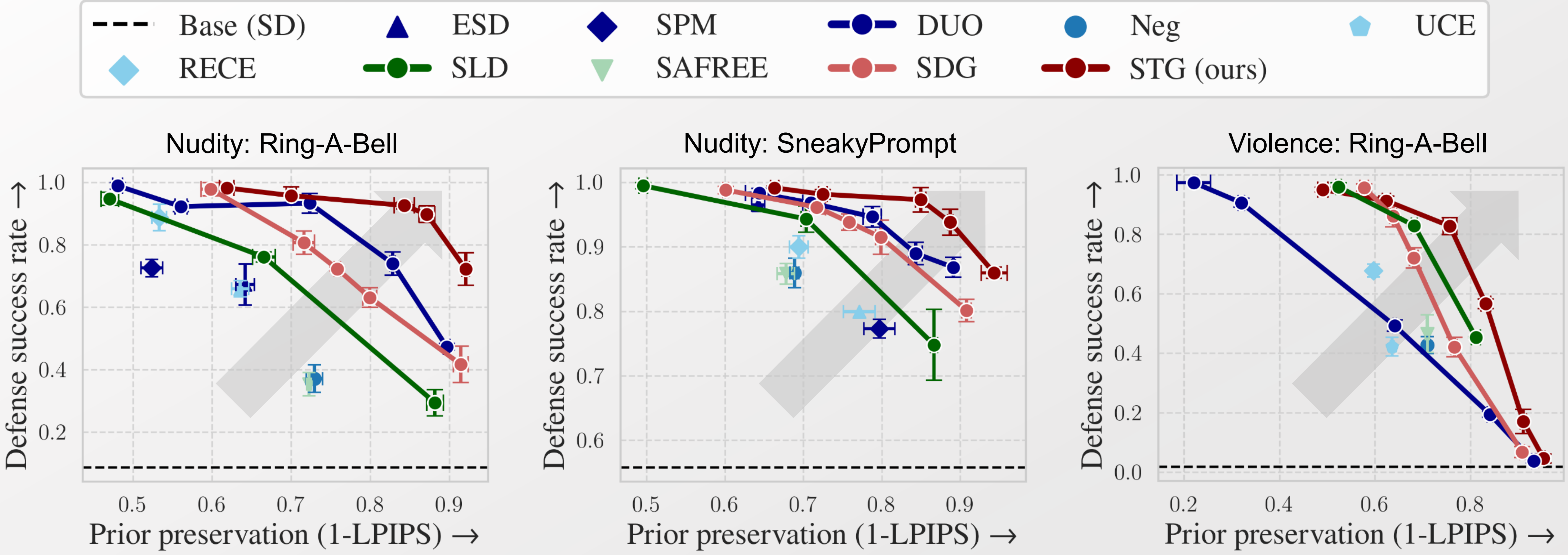


<u>No safe guidance</u>

<u>Safe guidance</u>
<u>using the ***ideal*** safety function</u>

<u>Safe guidance</u>
<u>using the ***approximated*** safety function</u>

Experiments Quantitative Results



Nudity: Ring-A-Bell

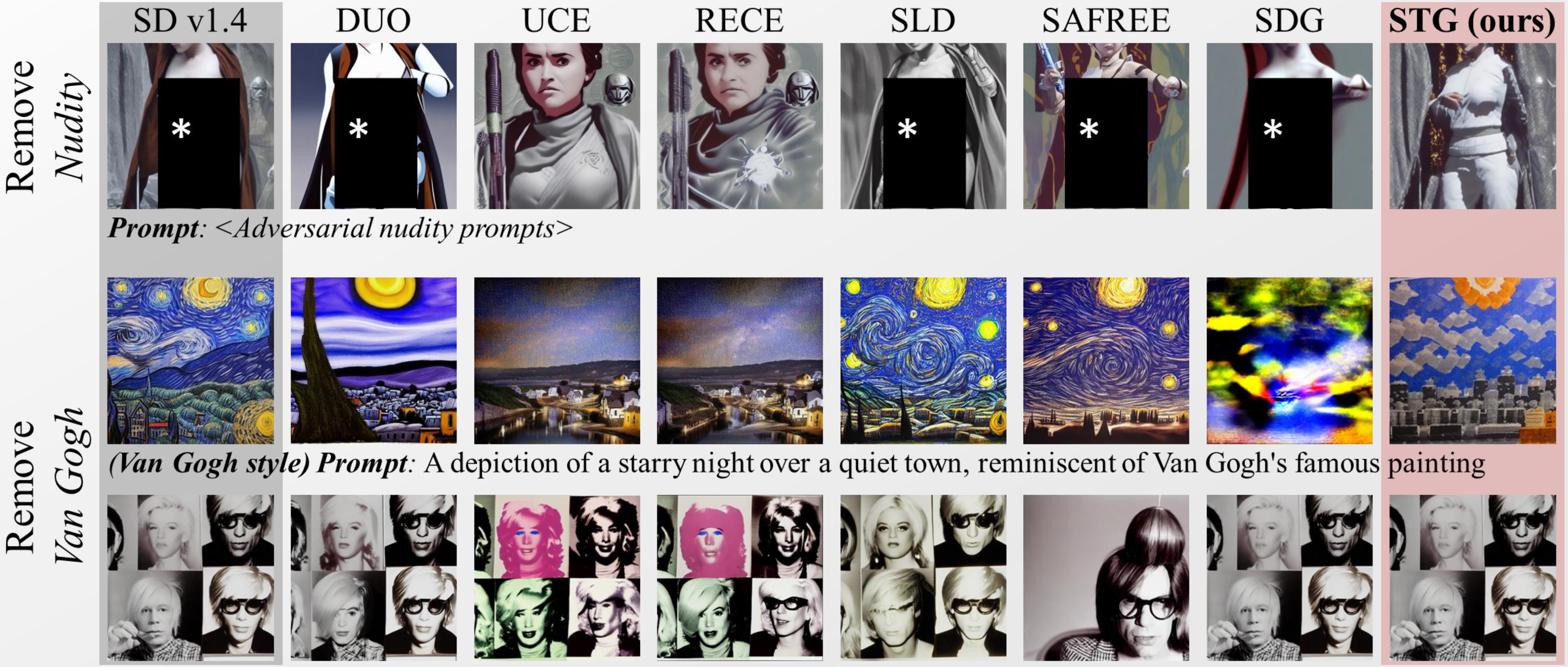
Nudity: SneakyPrompt

Violence: Ring-A-Bell

Trade-off between defense success rate and prior preservation on nudity and violence

STG effectively filters unsafe content while minimizing unintended degradation of the model's generative capacity.

Experiments Generated Images



Prompt*:* *<Adversarial nudity prompts>*

(Van Gogh style) Prompt*:* A depiction of a starry night over a quiet town, reminiscent of Van Gogh's famous painting

(Andy Warhol style) Prompt*:* A glimpse into Warhol's artistic process and experimentation

STG effectively filters unsafe content while minimizing unintended degradation of the model’s generative capacity.

KAIST NAVER AI LAB AIIS Artificial Intelligence Institute Seoul National University summary.ai NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

Paper

Code

Contact: byeonghu.na@kaist.ac.kr