Towards Safe Self-Distillation of Internet-Scale Text-to-Image Diffusion Models

Sanghyun Kim[1], Seohyeon Jung[1], Balhae Kim[1], Moonseok Choi[1], Jinwoo Shin[1], Juho Lee[1,2]
[1]Kim Jaechul Graduate School of AI, KAIST [2]AITRICS

Motivation *"Internet-trained models have internet-scale biases."* (Brown et al., 2020)

Pre-processing
Censoring training data (Stable Diffusion v2.0)

Impossible to remove all images
Lower image generation quality

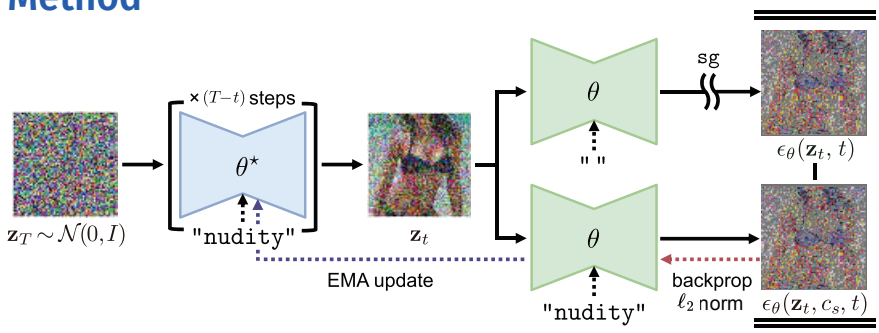
Processing
Inference-time techniques (SD+NEG, SLD, SEGA, ...)

Memory/computational costs
Cannot handle multi-concept

Post-processing
Built-in safety checker (CLIP zero-shot classification)

Can easily be circumvented
High false-negative rate

Method



Algorithm 1 SDD with multiple concepts

$\boldsymbol{\theta}^\star \leftarrow \boldsymbol{\theta}$, $\mathbf{c}_s = \texttt{CLIP}_{\text{text}}([c_1; \ldots; c_K])$
for $i = 1$ **to** N **do**
  $\mathbf{z}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $t \sim \mathcal{U}(\{0, \ldots, T-1\})$
  $\mathbf{c}_p \leftarrow \mathcal{U}(\{\texttt{CLIP}_{\text{text}}(c_1), \ldots, \texttt{CLIP}_{\text{text}}(c_K)\})$
  for $\tau = T$ **to** $t+1$ **do**
    $\tilde{\epsilon} \leftarrow \epsilon_{\boldsymbol{\theta}^\star}(\mathbf{z}_\tau, \tau) + s_g(\epsilon_{\boldsymbol{\theta}^\star}(\mathbf{z}_\tau, \mathbf{c}_p, \tau) - \epsilon_{\boldsymbol{\theta}^\star}(\mathbf{z}_\tau, \tau))$
    $\mathbf{z}_{\tau-1} \leftarrow \texttt{sampler}(\mathbf{z}_\tau, \tilde{\epsilon}, \tau)$
  end for
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \|\epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t, \mathbf{c}_s, t) - \texttt{sg}(\epsilon_{\boldsymbol{\theta}}(\mathbf{z}_t, t))\|_2^2$
  $\boldsymbol{\theta}^\star \leftarrow m\boldsymbol{\theta}^\star + (1-m)\boldsymbol{\theta}$
end for

Results

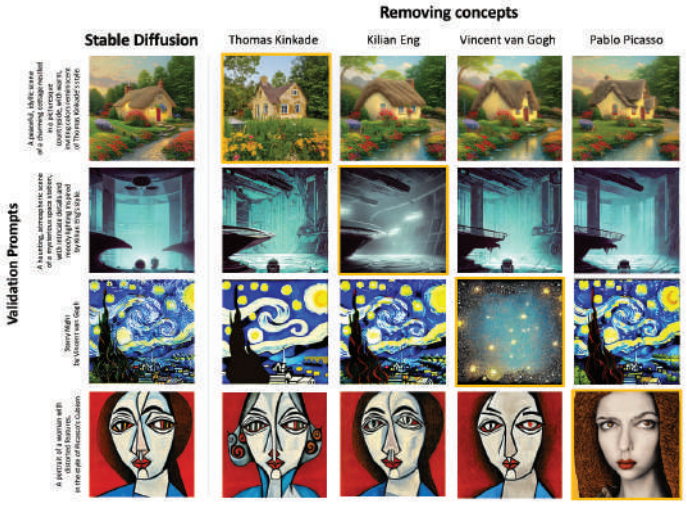
NSFW Content Removal

Method	"body" % NUDE ↓	COCO30k FID ↓	COCO30k LPIPS ↓	COCO30k CLIP ↑
SD	74.18	21.348	N/A	0.2771
SD + NEG	20.44	14.278	0.1954	0.2706
SLD medium	70.02	17.201	0.1015	0.2689
SLD max	4.30	13.634	0.1574	0.2709
SEGA	72.04	–	–	–
ESD -u-3	43.30	–	–	–
ESD -x-3	14.32	13.808	0.1587	0.2690
SDD (ours)	1.68	15.423	0.1797	0.2673
COCO ref.				0.2693

I2P Multi-Concept Removal

Method	"body" % NUDE ↓	I2P % HARM ↓	COCO30k FID ↓	COCO30k LPIPS ↓	COCO30k CLIP ↑
SD	74.18	24.42	21.348	N/A	0.2771
SD + NEG	63.78	9.51	18.021	0.1925	0.2659
SLD medium	74.16	7.42	14.794	0.4216	0.2720
SLD max	56.78	5.19	21.729	0.4377	0.2572
SEGA	74.10	16.84	–	–	–
ESD -x-3	47.38	13.04	16.411	0.2036	0.2631
SDD (ours)	12.62	5.03	15.142	0.2443	0.2560

Artist Concept Removal



EMA Teacher vs. Student