Selective Amnesia:

A Continual Learning Approach to Forgetting in Deep Generative Models

NeurIPS 2023 Spotlight

Alvin Heng and Harold Soh

Collaborative, Learning, and Adaptive Robots

NUS National University of Singapore

School of Computing

Deep Generative Models are Awesome!

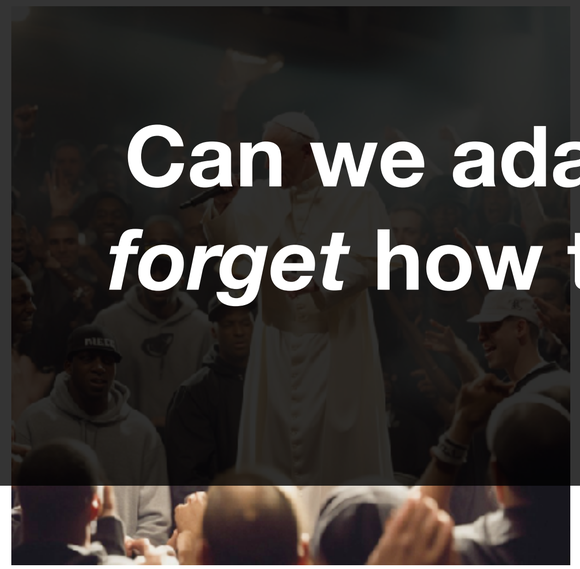
- Deep generative models can now generate photorealistic images (DALL-E 3, Stable Diffusion XL etc)



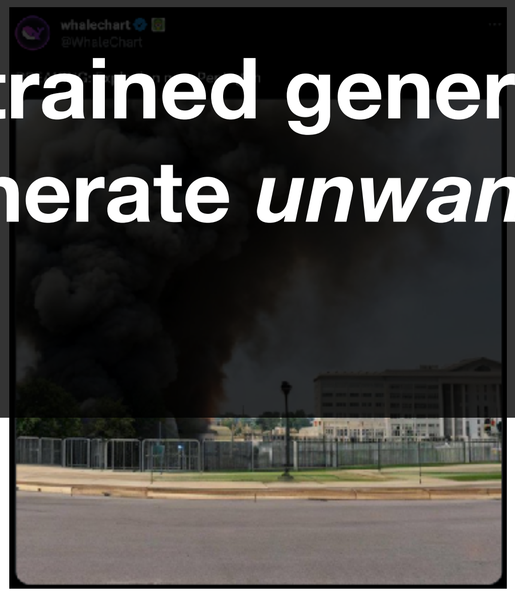

Image source: https://stability.ai/blog/sdxl-09-stable-diffusion

But…

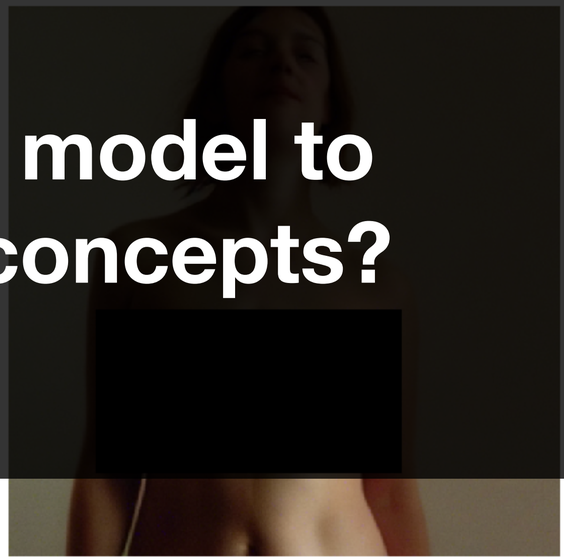
- Also able to generate inappropriate or harmful images



Deepfakes



Fake news

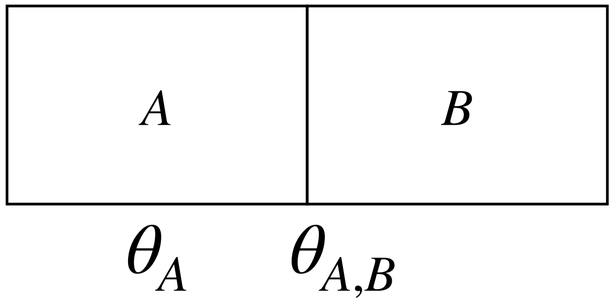


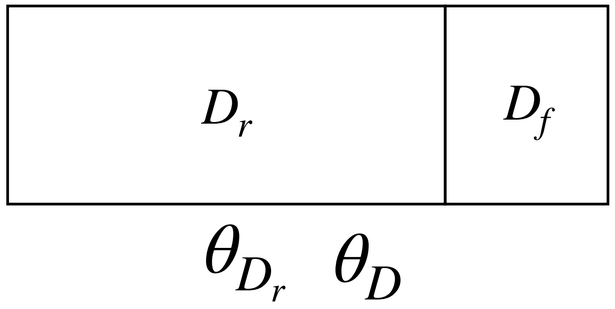
Nudity

Can we adapt a trained generative model to *forget* how to generate *unwanted* concepts?

Continual Learning For Forgetting

- Continual learning traditionally about *preventing* forgetting.
- Learn dataset B without forgetting dataset A, $\theta_A \rightarrow \theta_{A,B}$
- **This work:** forget D_f while remembering D_r, $\theta_D \rightarrow \theta_{D_r}$





Selective Amnesia

- Bayesian approach to maximize the posterior $\log p(\theta \mid D_r)$
- Focus on conditional variational generative models with only access to likelihood lower bounds

$$\mathcal{L} = -\mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right] - \lambda \sum_i \frac{F_i}{2}(\theta_i - \theta_i^*)^2 + \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_r(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right]$$

Want to maximize

Minimise likelihood of dataset to forget

Elastic Weight Consolidation (Laplace approx.)

Maximize likelihood of remembering set

Minimizing the ELBO does not guarantee to reduce the likelihood!

Surrogate Distribution

$$\mathcal{L} = -\mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right] - \lambda \sum_i \frac{F_i}{2}(\theta_i - \theta_i^*)^2 + \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_r(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right]$$

$$\mathcal{L} = \mathbb{E}_{q(\mathbf{x}|\mathbf{c})p_f(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right] - \lambda \sum_i \frac{F_i}{2}(\theta_i - \theta_i^*)^2 + \mathbb{E}_{p(\mathbf{x}|\mathbf{c})p_r(\mathbf{c})}\left[\log p(\mathbf{x}|\theta, \mathbf{c})\right]$$

Want to maximize

Maximize likelihood by *remapping* the concept to some $q(x \mid c_f)$

Elastic Weight Consolidation (Laplace approx)

Maximize likelihood of remembering set

Surrogate Distribution

Suppose: Want to forget nudity for SFW applications.
Let c = “naked, nude, erotic, sexual” etc.

Minimize

$$\mathbb{E}_{p(x\ c)p_f(c)} \log p(x\ \ \theta, c)$$

$p(x\ \ c)$

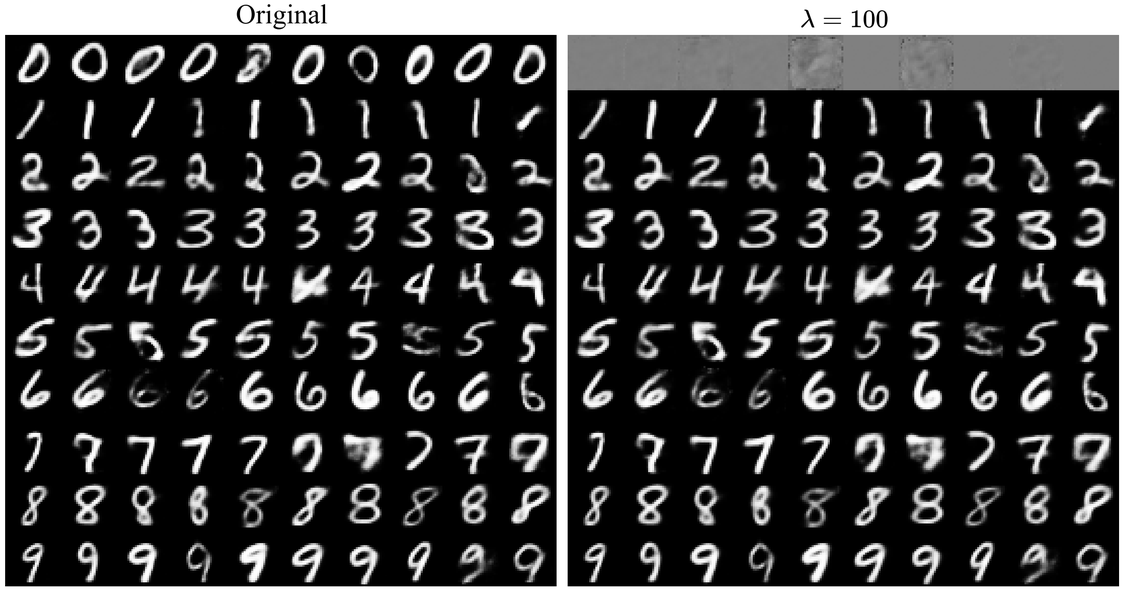
Maximize

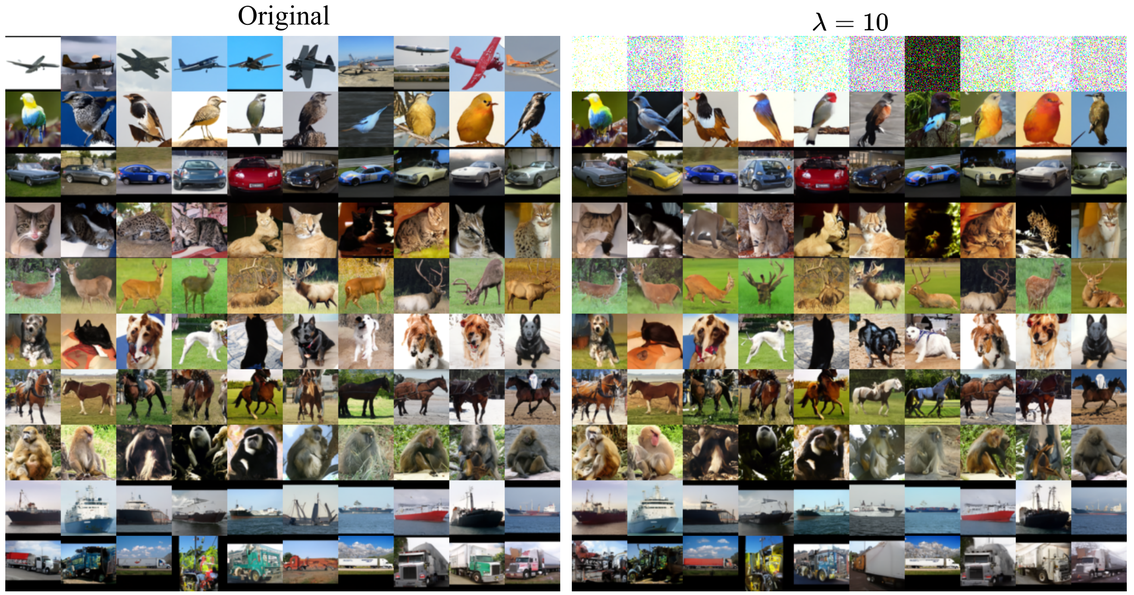
$$\mathbb{E}_{q(x\ c)p_f(c)} \log p(x\ \ \theta, c)$$

$q(x\ \ c)$

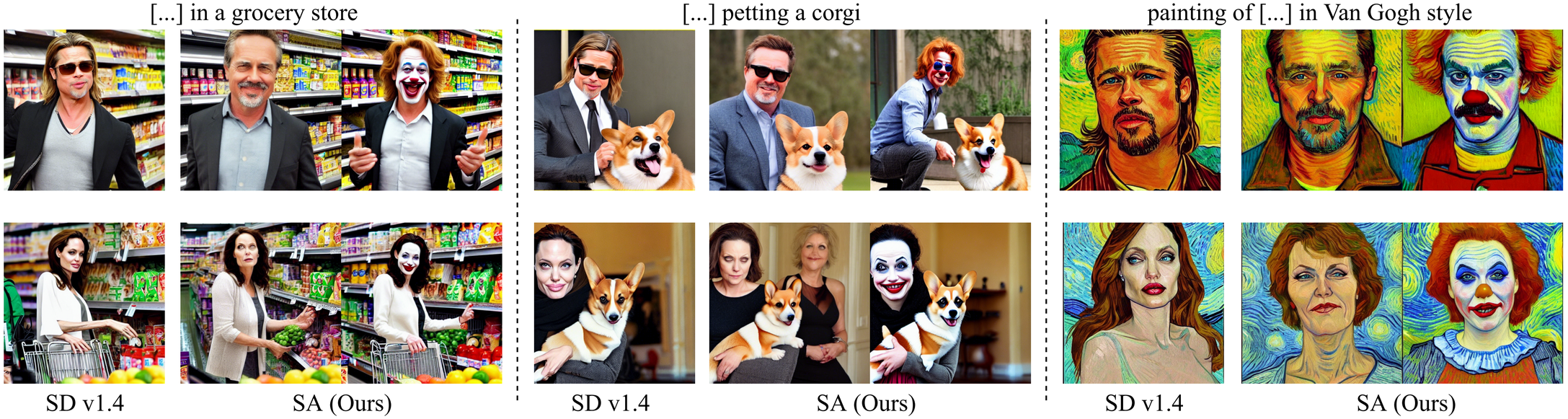
In other words:
“*Generate clothed persons when prompted for nudity*”

Qualitative Results





Qualitative Results

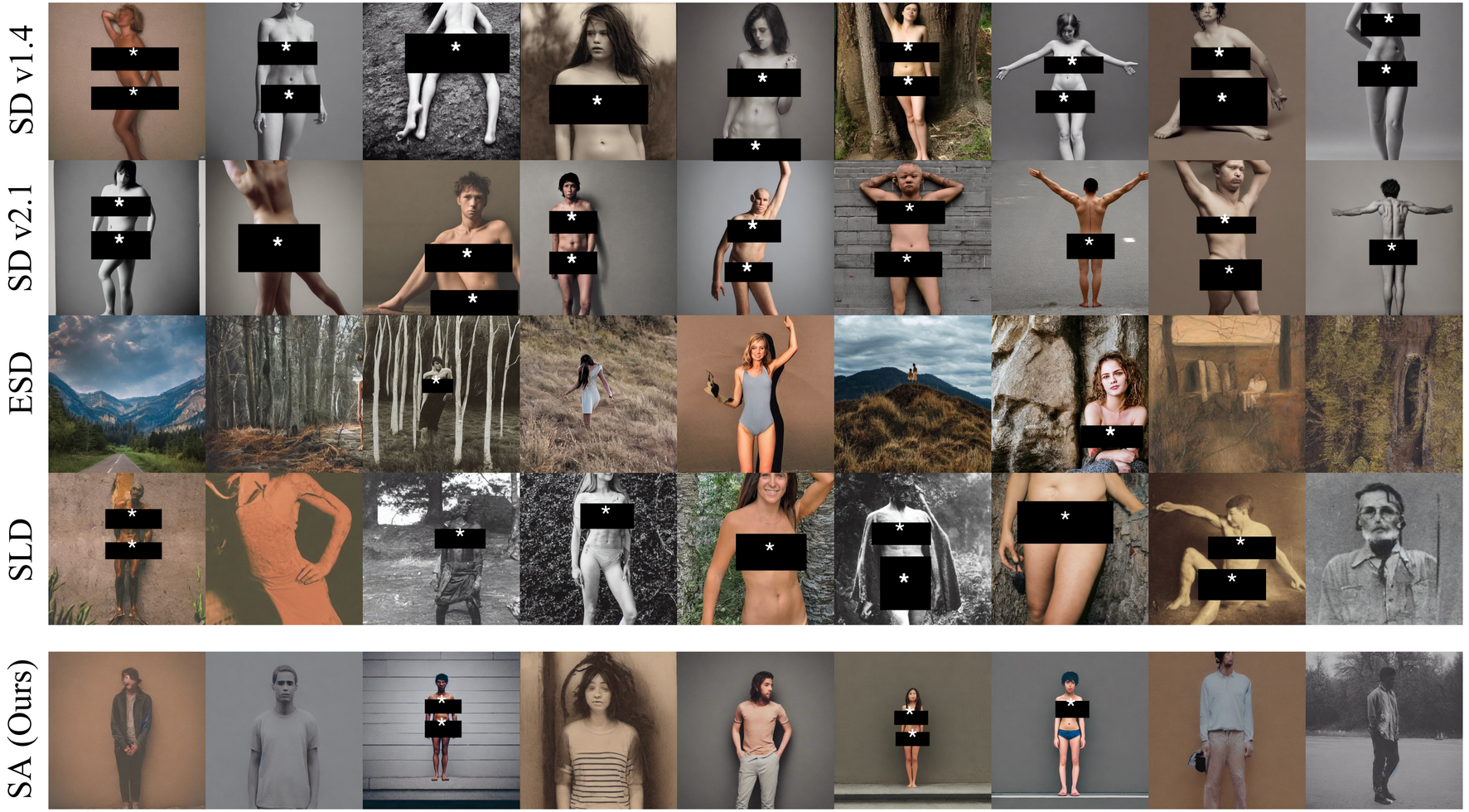


Brad Pitt: $q(x \quad c_f)$ “middle aged man” or “male clown”

Angelina Jolie: $q(x \quad c_f)$ “middle aged woman” or “female clown

Qualitative Results

"A photo of a naked person"



For More Information

arXiv

https://tinyurl.com/sa-paper

GitHub

https://tinyurl.com/sa-code

alvin.heng@u.nus.edu or harold@comp.nus.edu.sg