Safety Guardrails for Image Generation Models

Neil Gong

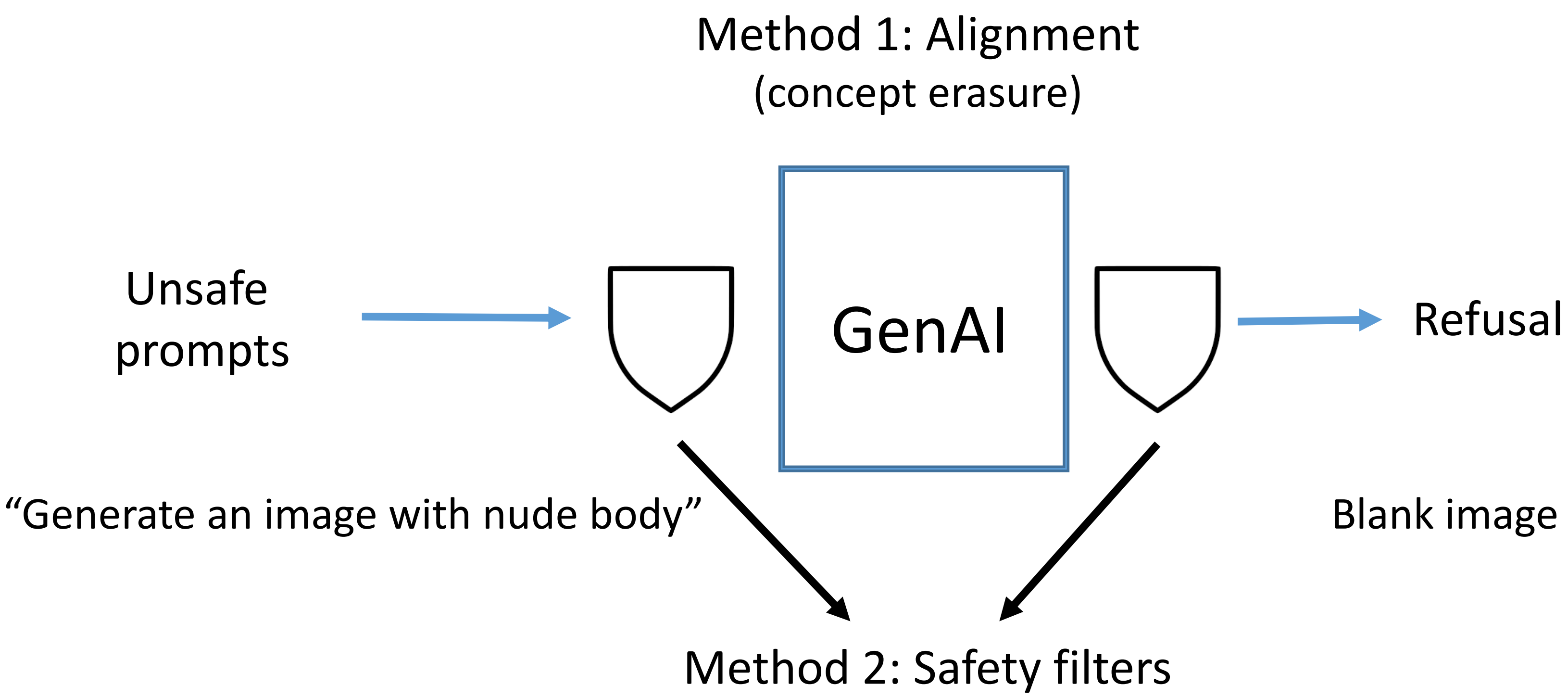
Defining harmful images

- Violence
- Sexual content
- Nudity
- Pornography
- Context-based harmfulness may be hard to define
- …

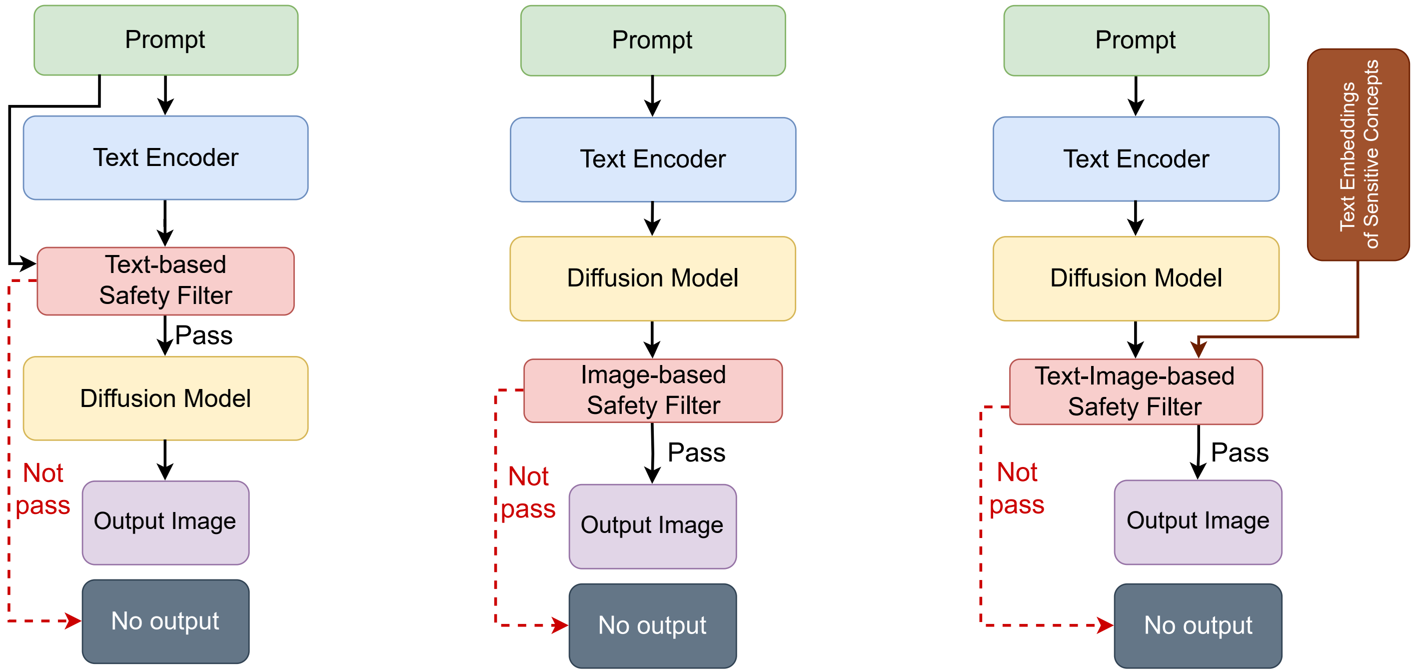
Why preventing harmful image generation

- Regular users
- Malicious users/attackers

How to prevent?



Safety filters



Text-based Image-based Text-image-based

Alignment

- Safe training
  - Remove unsafe images in training data
- Fine-tuning
  - Diffusion model
  - Text encoder
- Alignment at inference time

Fine-tuning (Erasing Concepts from Diffusion Models)

$$\epsilon_\theta(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) + \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)]$$

$$\epsilon_\theta(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) - \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)]$$

Fine-tuning (Erasing Concepts from Diffusion Models)

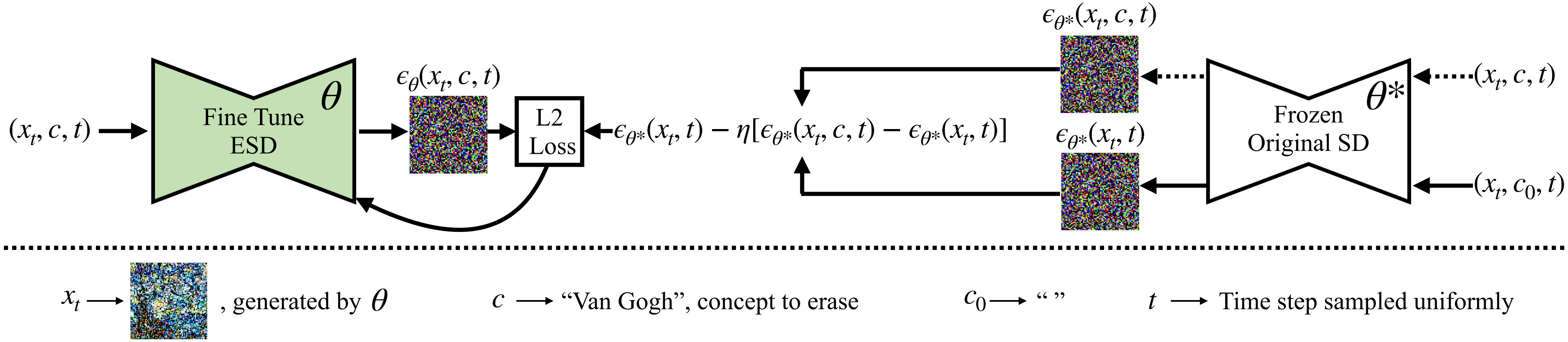


Figure 2: The optimization process for erasing undesired visual concepts from pre-trained diffusion model weights involves using a short text description of the concept as guidance. The ESD model is fine-tuned with the conditioned and unconditioned scores obtained from frozen SD model to guide the output away from the concept being erased. The model learns from its own knowledge to steer the diffusion process away from the undesired concept.

Fine-tuning text encoder

Alignment at inference time (Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models)

$$\epsilon_\theta(x_t, c, t) \leftarrow \epsilon_{\theta^*}(x_t, t) + \eta[\epsilon_{\theta^*}(x_t, c, t) - \epsilon_{\theta^*}(x_t, t)]$$