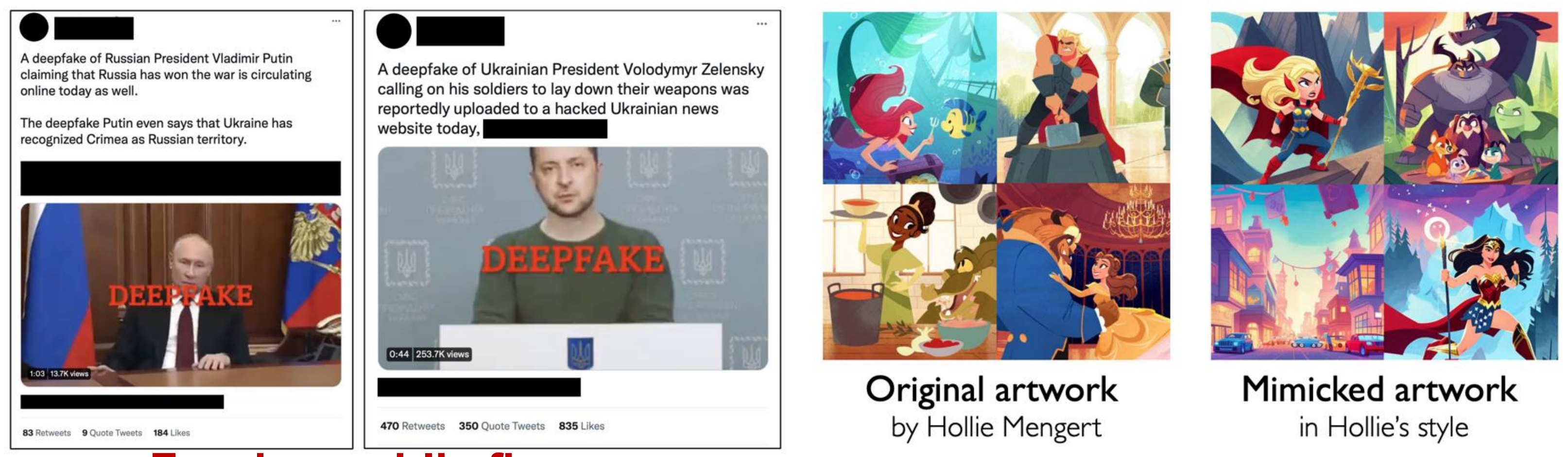
FreezeAsGuard: Mitigating Illegal Adaptation of Diffusion Models via Selective Tensor Freezing

Kai Huang, Haoming Wang and Wei Gao
University of Pittsburgh

[Lab Homepage]

Background and Motivation

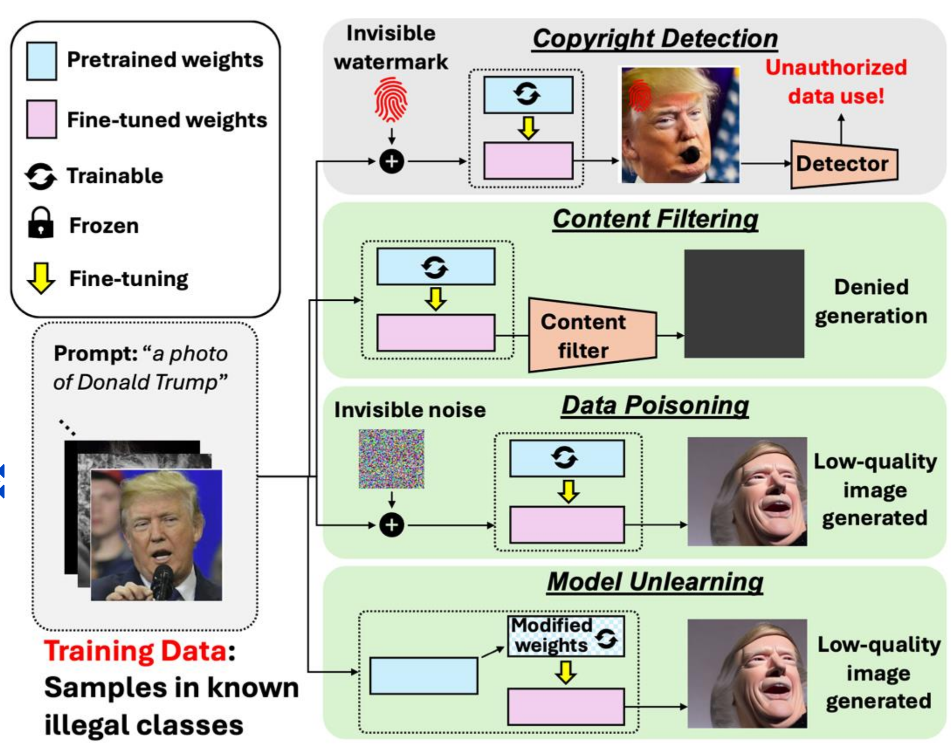
❖ Illegal Adaptation of Diffusion Models



Forging public figures **Copying artistic styles**

- Open-source text-to-image diffusion models can be efficiently adapted to new identities, artistic styles, and application domains using small custom datasets.
- The same flexibility can enable unauthorized uses, including forging public figures' portraits, duplicating copyrighted artistic styles, and generating explicit content.

❖ Limitation of Existing works



Only detects watermarked data; cannot prevent adaptation.

Easy to bypass; high false-positive rate.

Requires control of training data; fails on unprotected data.

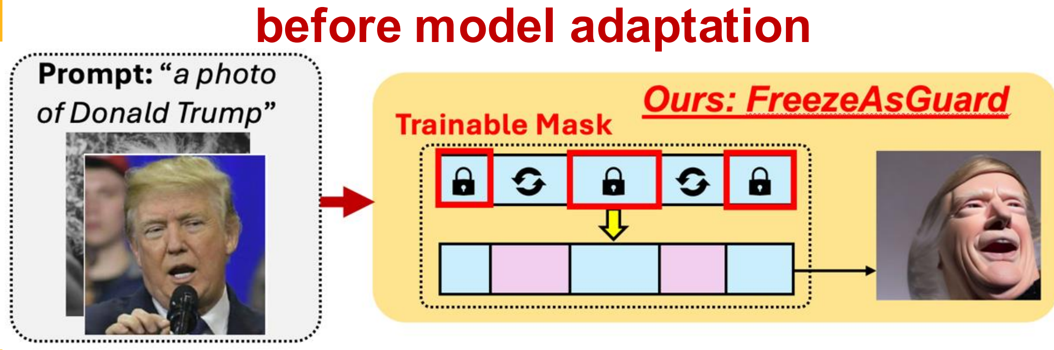
Knowledge can be relearned; may harm legal adaptation.

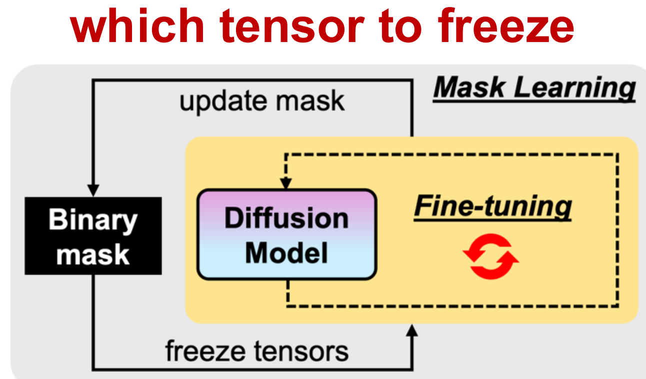
Method Overview

- FreezeAsGuard constrains what the model can change by selectively freezing tensors that are critical to illegal adaptations while leaving tensors needed for legal adaptations trainable.

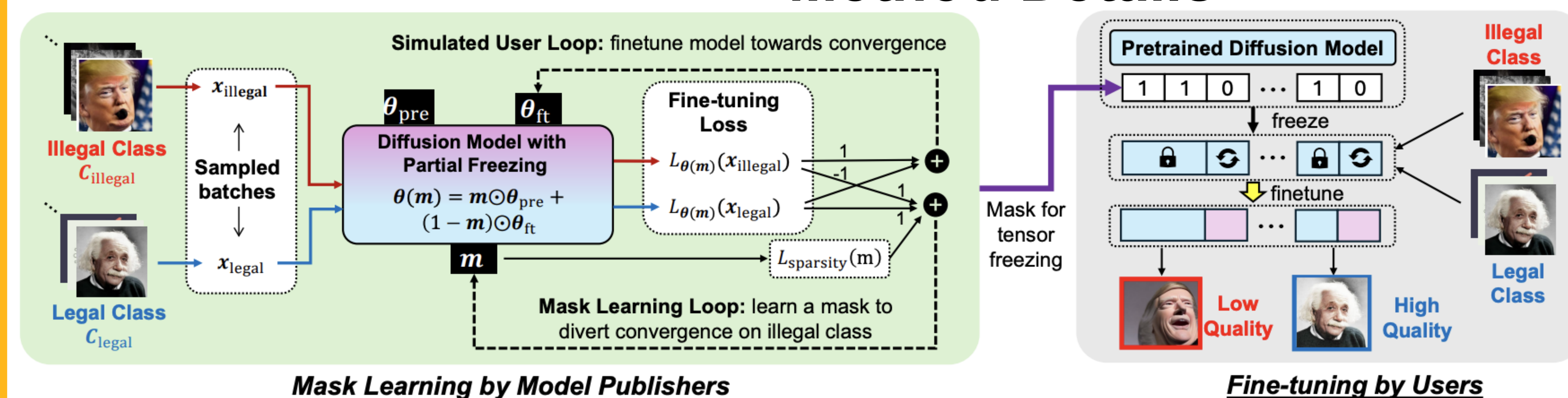
Selectively Freezing Tensors before model adaptation

Bilevel optimization to identify which tensor to freeze





Method Details



m: Binary mask .
m^*: Optimized mask.
$\theta(m)$: Model tensors controlled by mask m.
$\theta^*(m)$: Converged model tensors after fine-tuning with mask m.
$x_{illegal}$: Training samples from illegal classes .
x_{legal} : Training samples from legal classes.

❖ Mask Learning via Multi-objective Bi-level Optimization

Mask learning Loop **Simulated User Loop**

$$\mathbf{m}^* = \arg\min_{\mathbf{m}} \underbrace{\left(-\mathcal{L}_{\theta^*(\mathbf{m})}(\mathbf{x}_{illegal}), \mathcal{L}_{\theta^*(\mathbf{m})}(\mathbf{x}_{legal})\right)}_{-\lambda_1\mathcal{L}_{\theta^*(\mathbf{m})}(\mathbf{x}_{illegal}) + \lambda_2\mathcal{L}_{\theta^*(\mathbf{m})}(\mathbf{x}_{legal})} \quad \text{s.t.} \quad \theta^*(\mathbf{m}) = \arg\min_{\theta(\mathbf{m})} \underbrace{\left(\mathcal{L}_{\theta(\mathbf{m})}(\mathbf{x}_{illegal}), \mathcal{L}_{\theta(\mathbf{m})}(\mathbf{x}_{legal})\right)}_{\mathcal{L}_{\theta^*(\mathbf{m})}(\mathbf{x}_{illegal}) + \mathcal{L}_{\theta^*(\mathbf{m})}(\mathbf{x}_{legal})}$$

linear scalarization

- Estimate $\theta(m)$ to apply gradient solver

$$\theta(\mathbf{m}) = \mathbf{m} \odot \theta_{pre} + (\mathbf{1} - \mathbf{m}) \odot \theta_{ft}$$

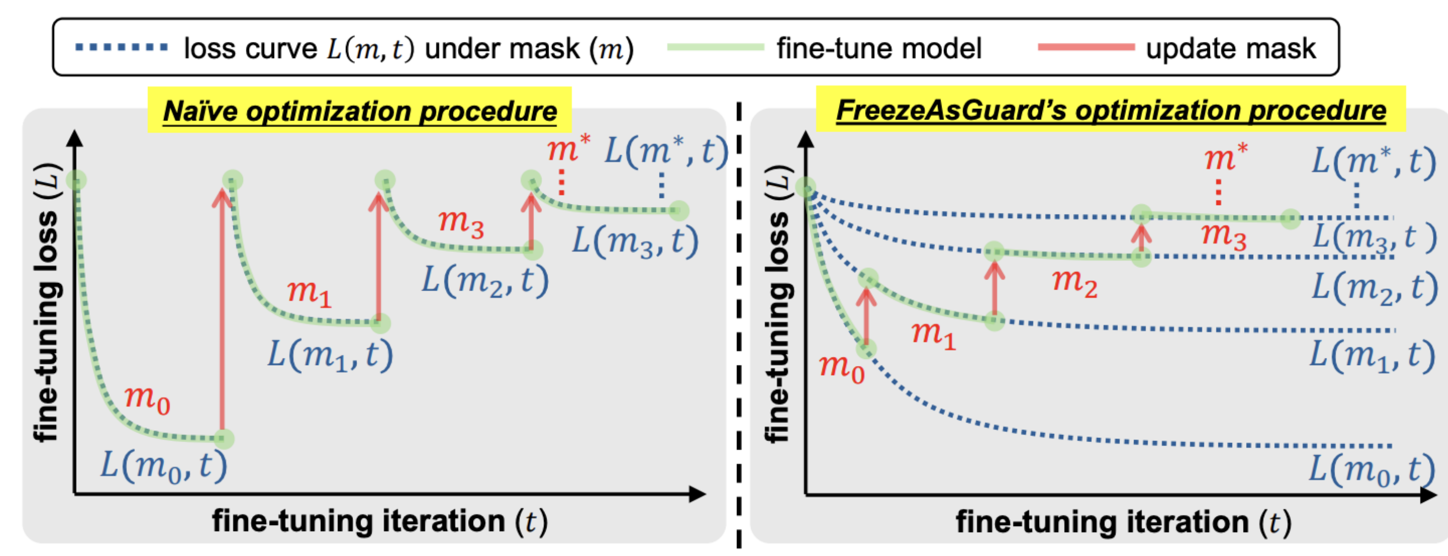
- Controlling mitigation power via mask ratio

$$\mathcal{L}_{sparsity} = \|\mathbf{1}^\top \mathbf{m}/N - \rho\|_2^2 \longrightarrow$$

Added to the loss of mask learning loop

❖ Efficient Optimization

Early Stop strategy



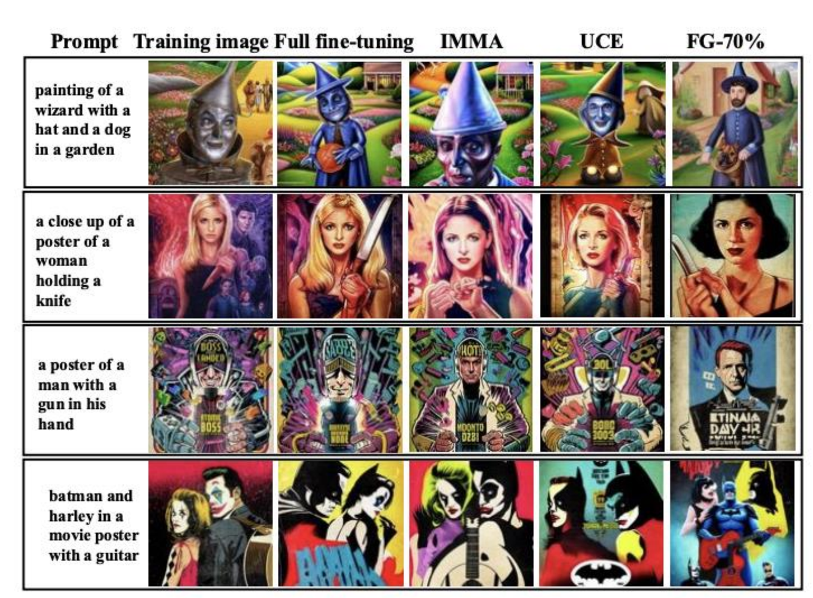
- fine-tuning loss typically drops fast in the first few iterations and then violently fluctuates
- do not wait for the loss to converge to reduce cost

Reuse model states across iterations

Stores only the masked weights $\theta(\mathbf{m})$ and the difference $\theta_d = \theta_{pre} - \theta_{ft}$, instead of three full model copies.

Experiment results

❖ Comparison with baselines



❖ Different Frozen ratio