Leveraging Model Guidance to Extract Training Data from Personalized Diffusion Models

Xiaoyu Wu*, Jiaru Zhang, Steven Wu
CMU

* Work done during internship at CMU

Few-Shot Fine-Tuning of Diffusion Models

Base




FT



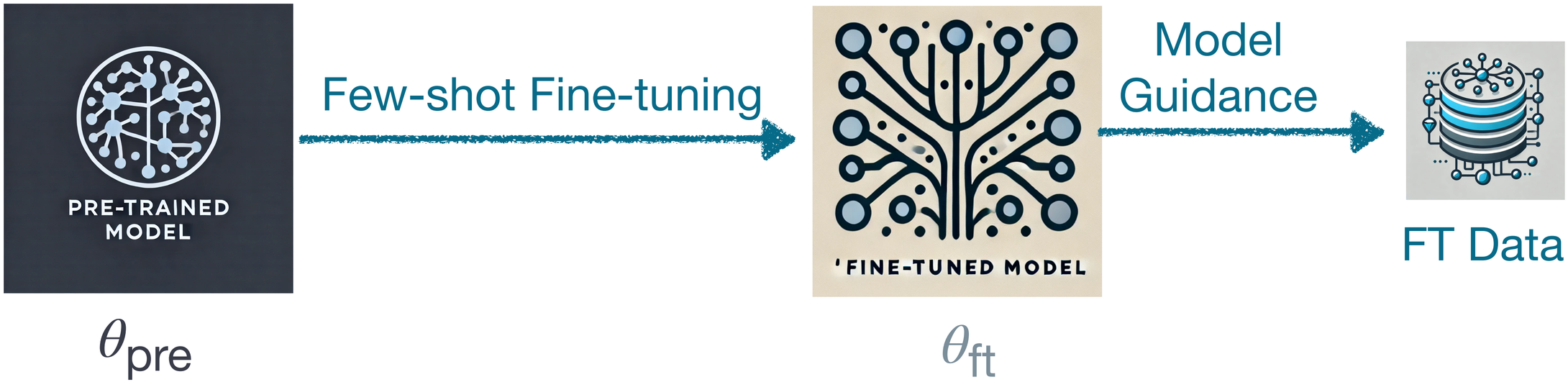

Online Communities of Shared Checkpoints

Hugging Face

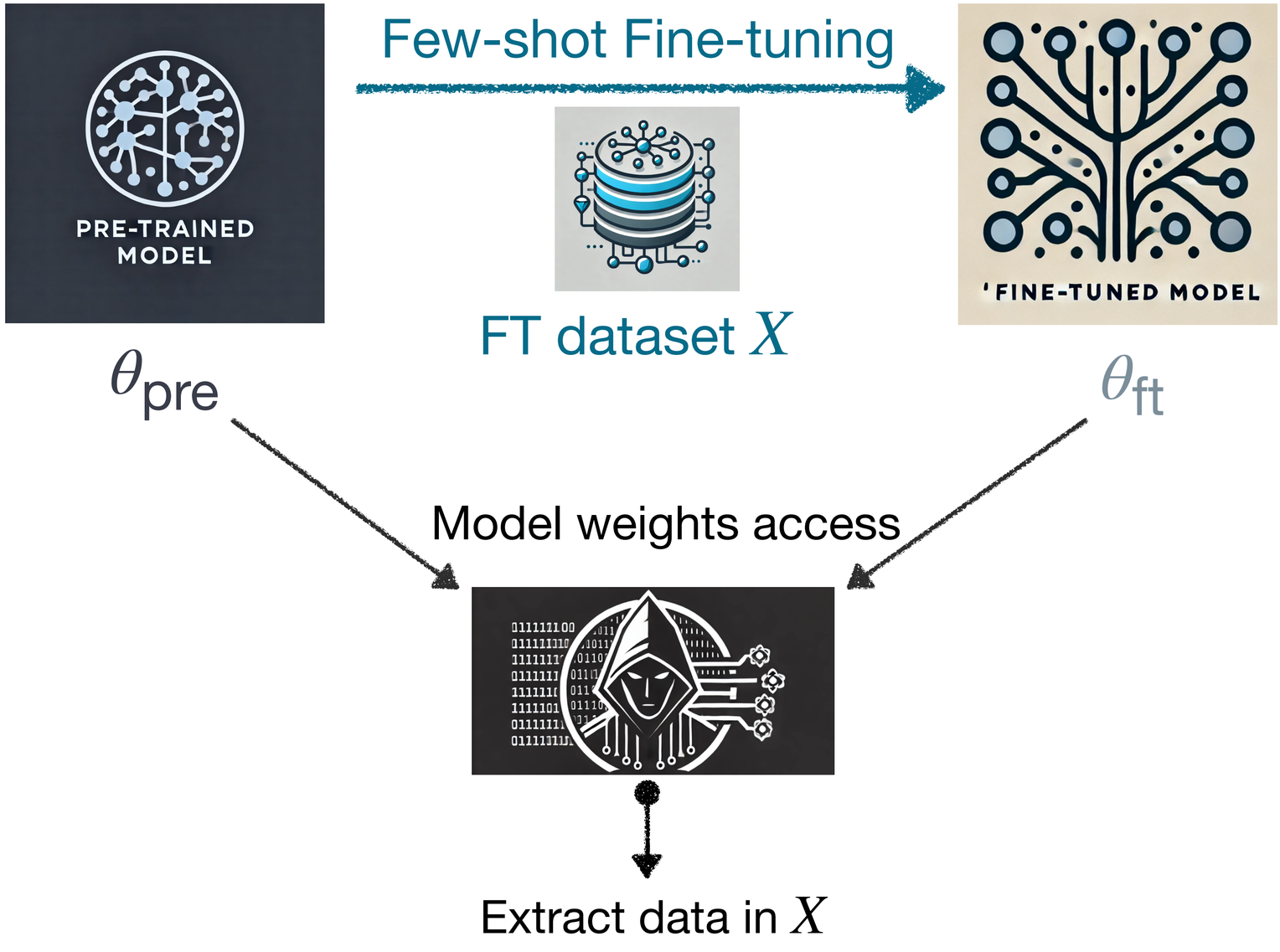
- Millions of personalized checkpoints shared publicly (Civitai, HuggingFace)
- Copyright risks: Unauthorized use of artists' work in fine-tuning generative models
- Privacy risks: Fine-tuning may involve sensitive data, such as human faces

Q: Is it possible to extract fine-tuning data from these fine-tuned Diffusion Model checkpoints released online?

Our work: data extraction attacks by leveraging *model guidance* techniques.

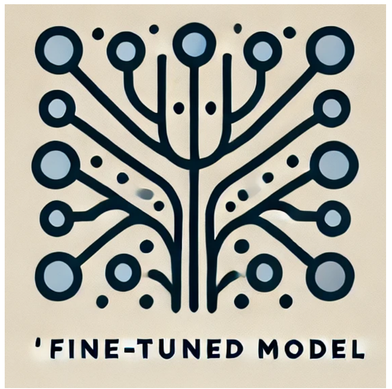


Threat Model



Model Guidance

Few-shot Fine-tuning



θ FT distribution q θ'

Parametric approximation:

$$P_{\theta'}(x) \propto P_{\theta}(x)^{1-\lambda} \cdot q(x)^{\lambda}$$

Guidance towards q using *scores* of θ and θ'

$$\nabla_x \log q(x) = \nabla_x \log P_{\theta'}(x) + \frac{1-\lambda}{\lambda}(\nabla_x \log P_{\theta'}(x) - \nabla_x \log P_{\theta}(x)) .$$

Denoising function

Extension to caption conditional guidance

Our Method: FineXtract

Training Images X_0

Pretrained θ

few-shot fine-tuning

Fine-tuned θ'

Accessible to Attacker

Guided Generation

$(\times N)$

Generated Image X_i

Training Caption c

Generated Image Set X

Clustering

Extracted Images $\widehat{X}$

Extraction from FT Checkpoints on HuggingFace

Extracted:

Training images:

Evaluation Metrics

- Fine-tuning dataset X_0
- Extracted dataset $\hat{X}$
- sim: SSCD similarity function
- Metric 1: Average similarity (with the closest extracted image)

$$\text{AS}(X_0, \hat{X}) = \frac{1}{X_0} \sum_{i=1}^{X_0} \max_j \text{sim}(X_0^{(i)}, \hat{X}^{(j)})$$

- Metric 2: Average Extraction Success Rate (A-ESR)

$$\text{A-ESR}_\tau = \frac{1}{X_0} \sum_{i=1}^{X_0} \mathbf{1}\left(\max_j \text{sim}(X_0^{(i)}, \hat{X}^{(j)}) > \tau \right)$$

Comparing FineXtract w/ Baelines

Object-Driven Generation: DreamBooth Dataset

Metrics and Settings	DreamBooth			LoRA		
	AS↑	A-ESR$_{0.7}$↑	A-ESR$_{0.6}$↑	AS↑	A-ESR$_{0.7}$↑	A-ESR$_{0.6}$↑
Direct Text2img+Clustering	0.418	0.03	0.11	0.347	0.00	0.02
CFG+Clustering	0.528	0.15	0.36	0.379	0.01	0.05
FineXtract	**0.557**	**0.25**	**0.45**	**0.466**	**0.04**	**0.18**

Leveraging Model Guidance to Extract Training Data from Personalized Diffusion Models

Xiaoyu Wu (nicholaswu2022@gmail.com), Jiaru Zhang,
Steven Wu (zstevenwu@cmu.edu)