CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

Quantifying Privacy Risks of Prompts in Visual Prompt Learning

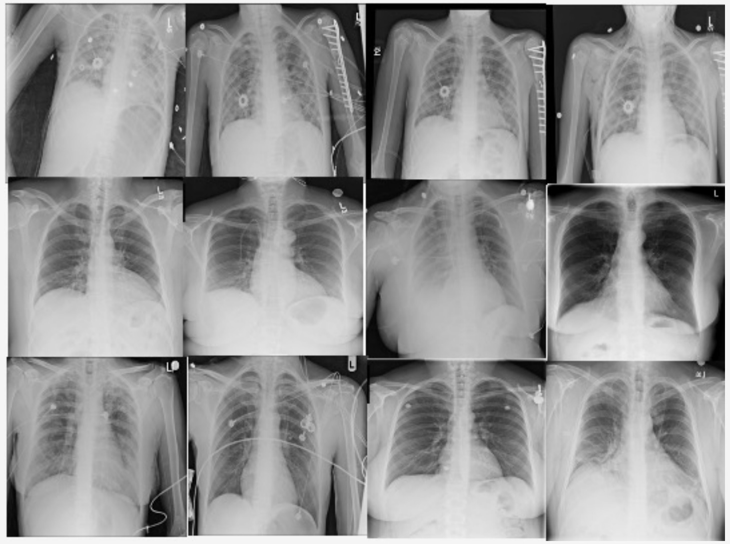
Yixin Wu,[1] *Rui Wen*,[1] *Michael Backes*,[1] *Pascal Berrang*,[2] *Mathias Humbert*,[3] *Yun Shen*,[4] *Yang Zhang*[1]

[1] CISPA Helmholtz Center for Information Security

[2] University of Birmingham [3] University of Lausanne [4] NetApp

Traditional Fine-Tuning Paradigm

What is a common method to adapt pre-trained models to specific tasks?

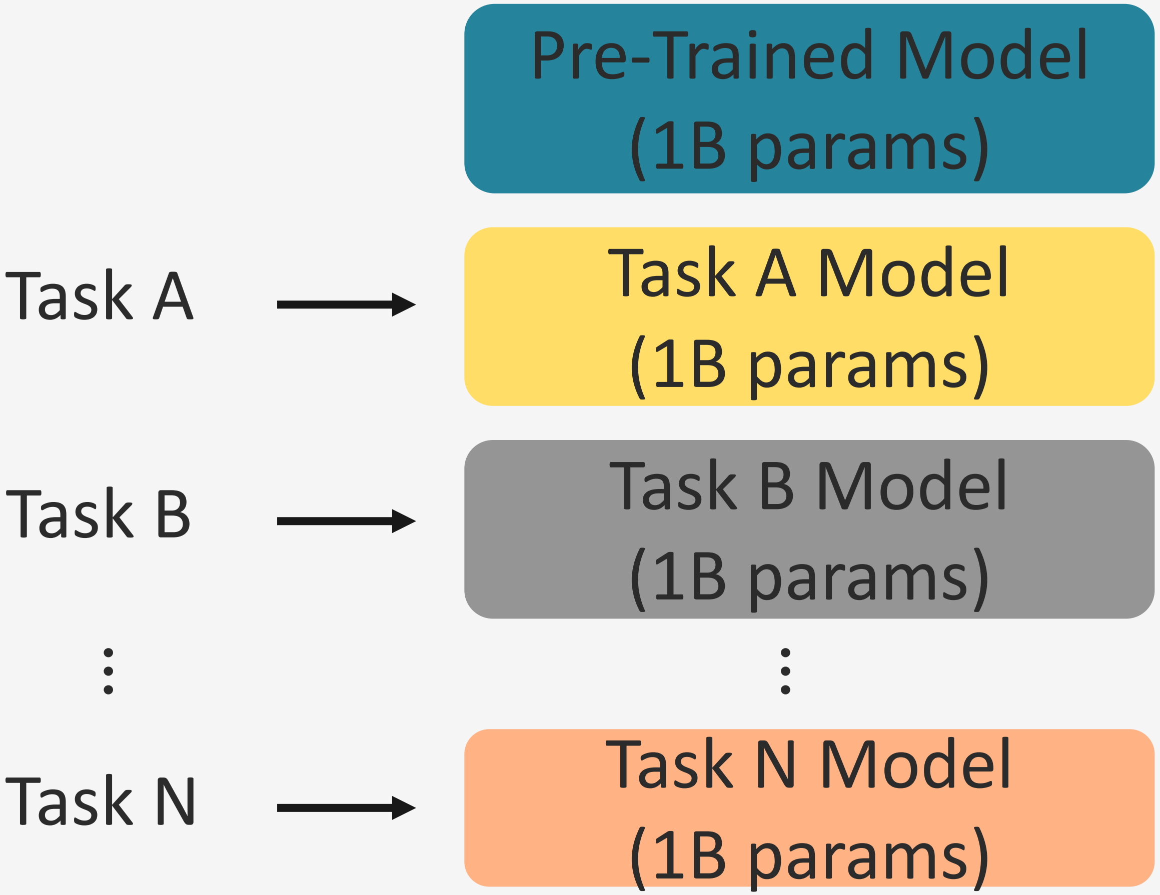


Fine-Tune → Pre-Trained Model → Task-Specific Model

Infer → Task-Specific Model → Diagnosis Results

Traditional Fine-Tuning Paradigm

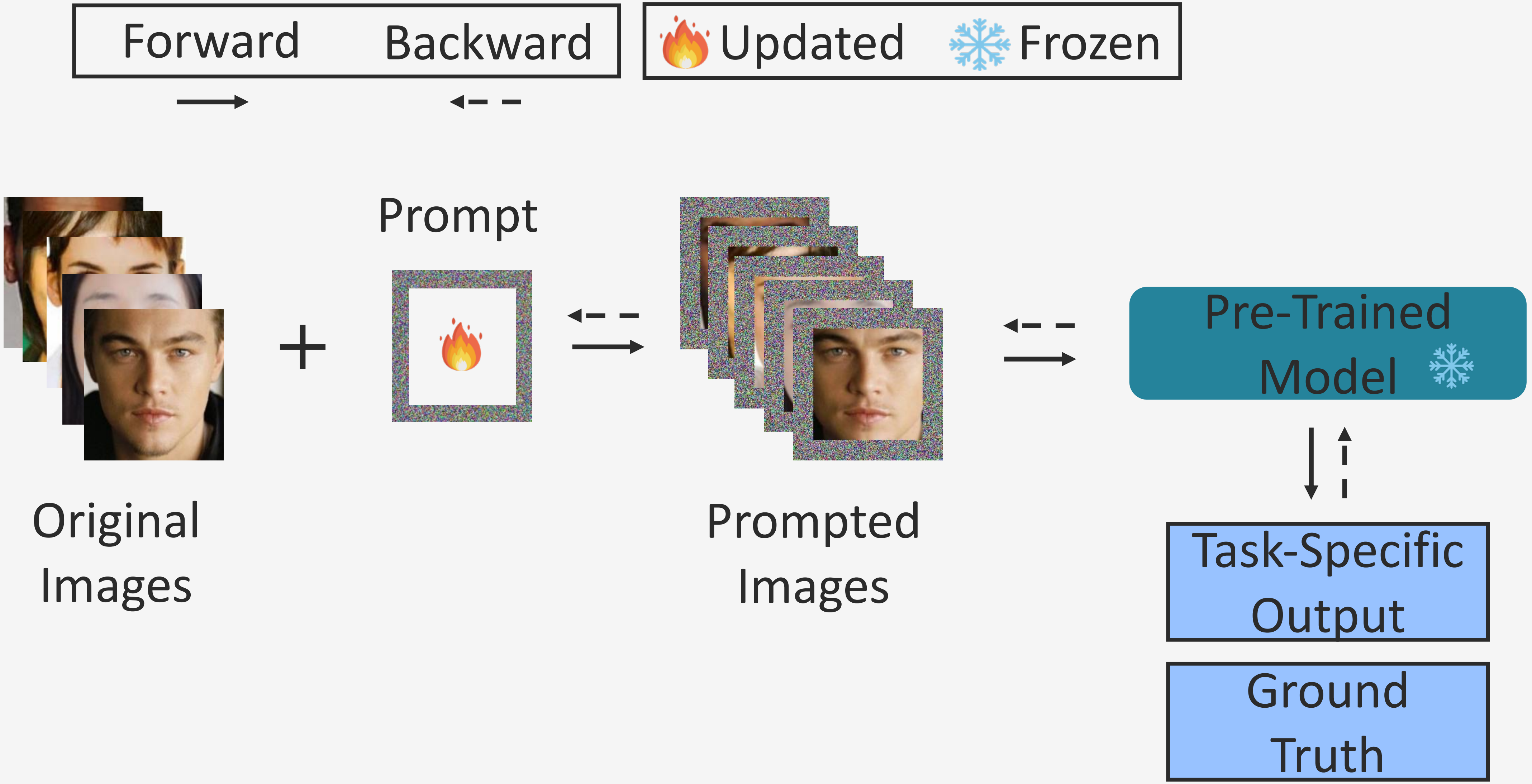
As the number of specific tasks gradually increases…



Large-scale pre-trained models are costly to share and serve in the fine-tuning paradigm

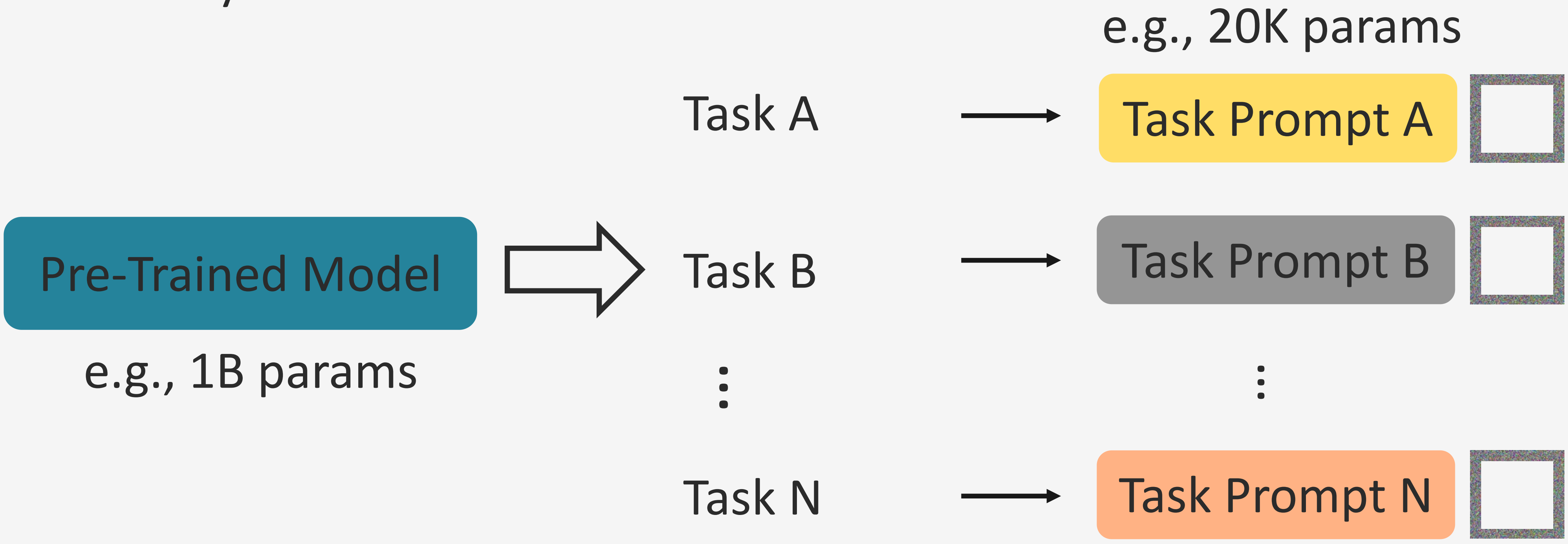
Visual Prompt Learning and Its Workflow

A new paradigm is introduced to solve such limitations

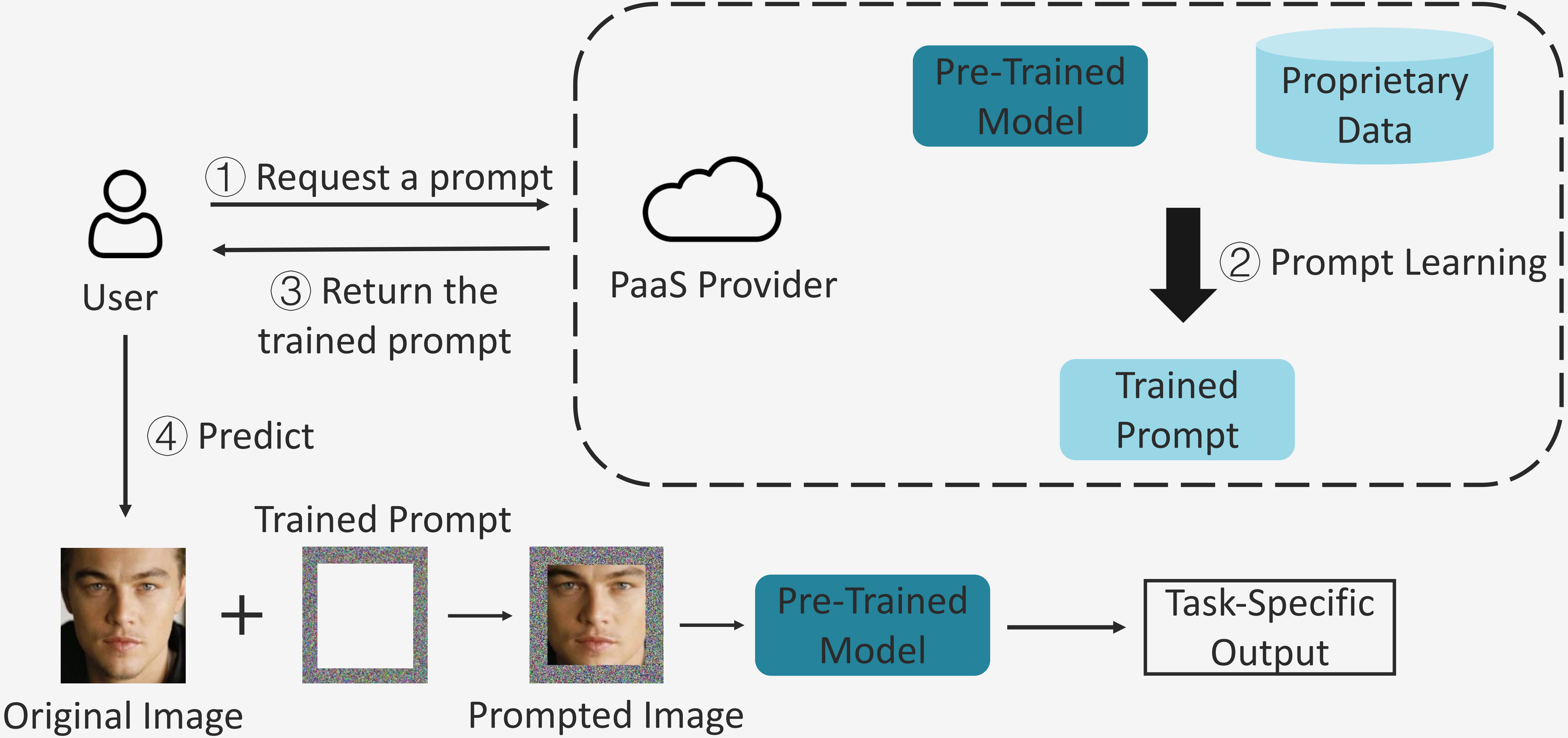


Pros of Prompt Learning

- Remain the pre-trained model frozen
- Far fewer parameters are updated
- Easy to share and serve to users



Prompt as a Service (PaaS)

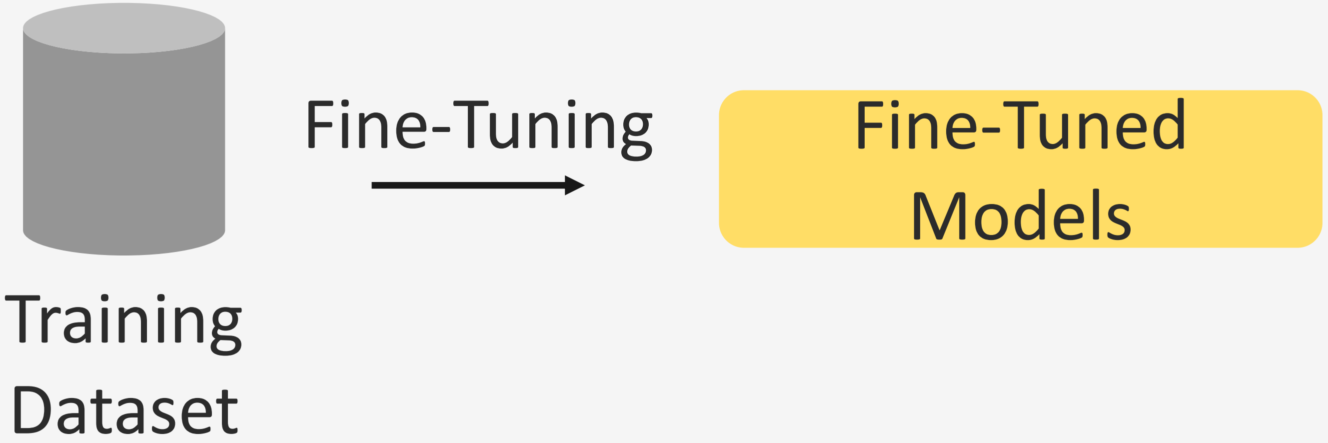


Pros of PaaS

- For users
  - Minimize their effort in developing a prompt
  - Keep their data on premise
  - Easily adapt to different downstream tasks
- For providers
  - Reuse a single pre-trained model to support multiple downstream tasks
  - Less computational resource for training
  - Less storage space
- A well-generalized prompt becomes a valuable asset for PaaS providers

Privacy Risks of ML Models

- Most previous research about privacy risks has focused on ML models at the model level

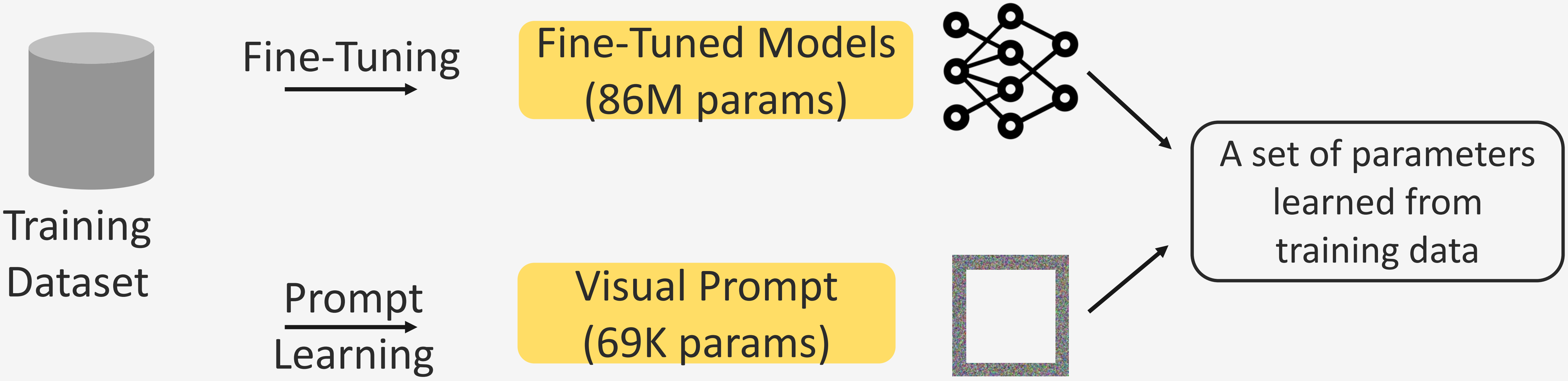


Privacy information is leaked through fine-tuned models[1][2]

[1] Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In International Conference on Machine Learning (ICML), pages 19641974. PMLR, 2021.

[2] Fatemehsadat Mireshghallah, Archit Uniyal, Tianhao Wang, David Evans, and Taylor Berg-Kirkpatrick. An Empirical Analysis of Memorization in Fine-tuned Autoregressive Language Models. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1816–1826. ACL, 2022.

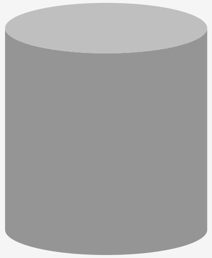
Privacy Risks of Prompt Learning



- **Assumption:** will prompt learning heavily compress the training dataset information, thus leading to less effective privacy attacks?
  - Compared to the fine-tune paradigm, only 0.08% params are updated

Privacy Risks of Prompt Learning

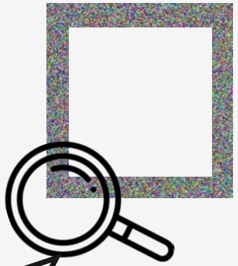
- **Assumption:** Will prompt learning heavily compress the training dataset information, thus leading to less effective privacy attacks?



Training Dataset

Prompt Learning

Visual Prompt (69K params)



Membership Inference

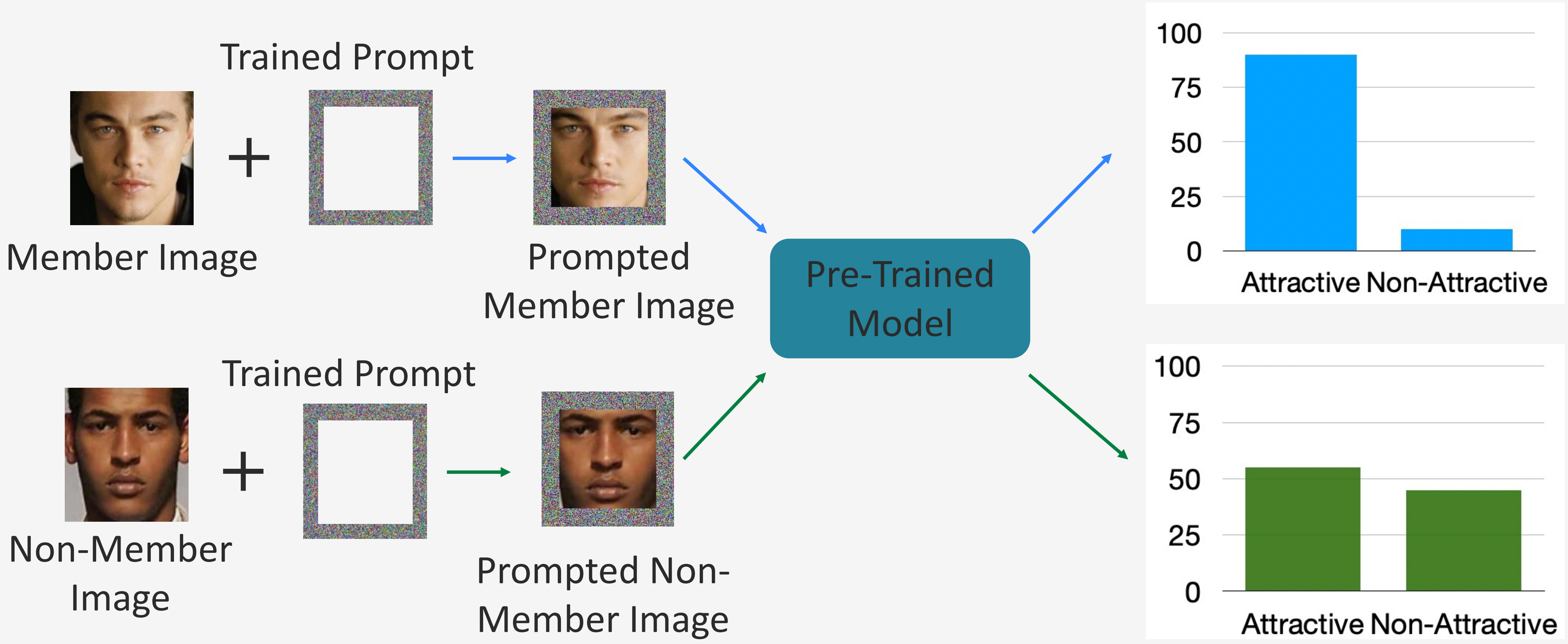
Member Image

Skewed Towards Females

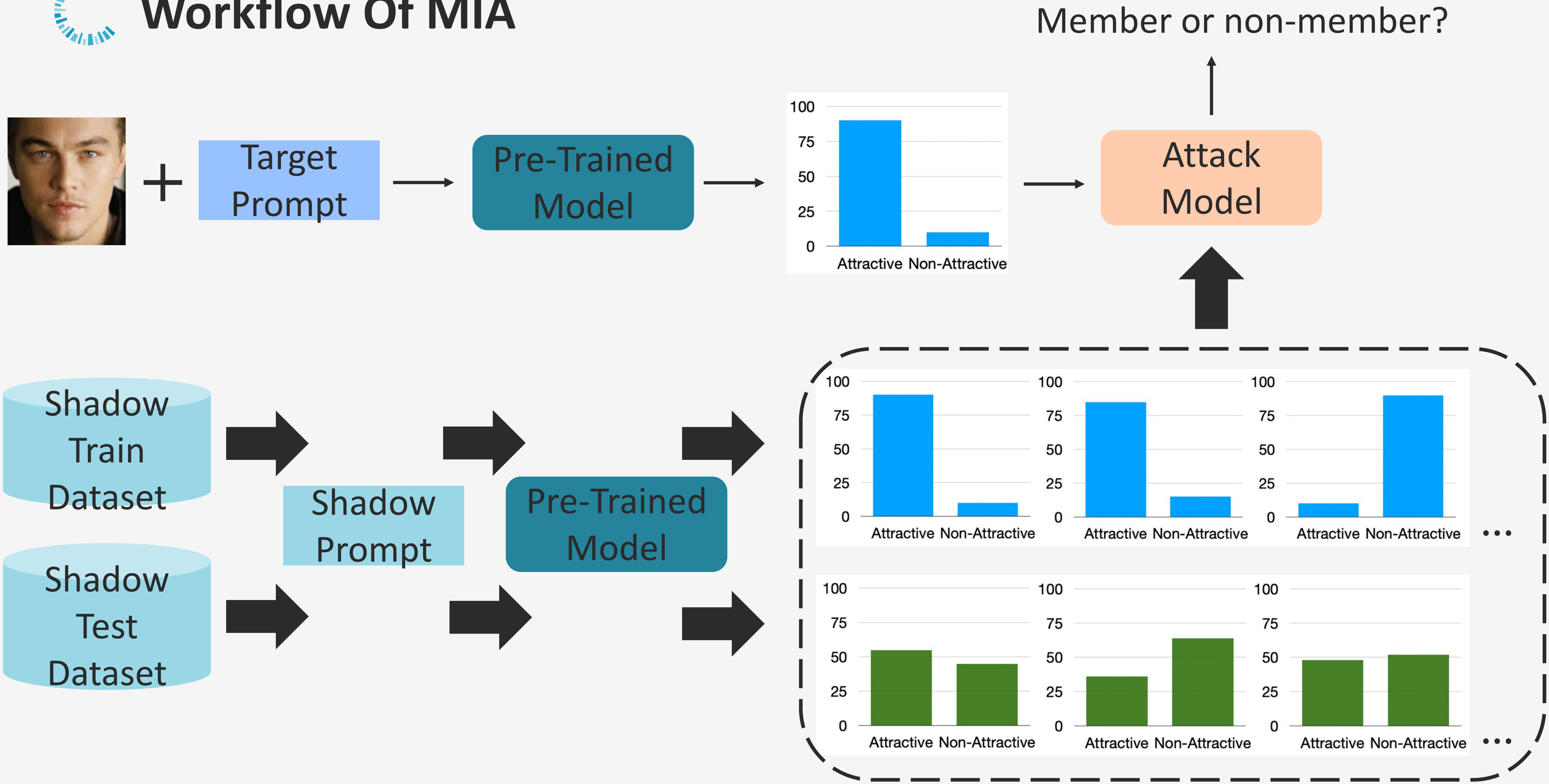
Property Inference

Membership Inference Attacks (MIAs)

- Membership inference attacks (MIAs): infer whether a given data sample x was in the training dataset of the target prompt



Workflow Of MIA



MIA Evaluation

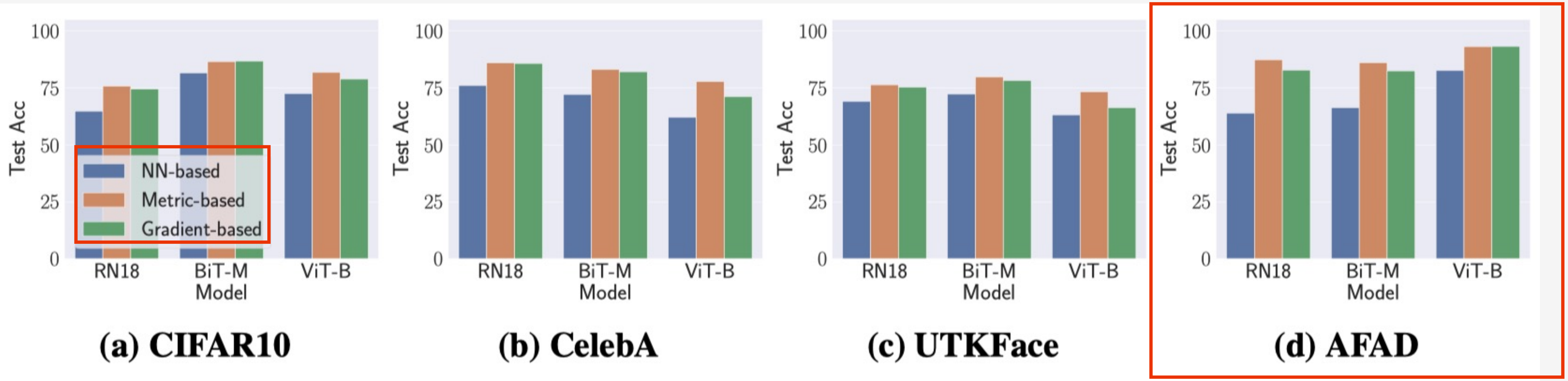
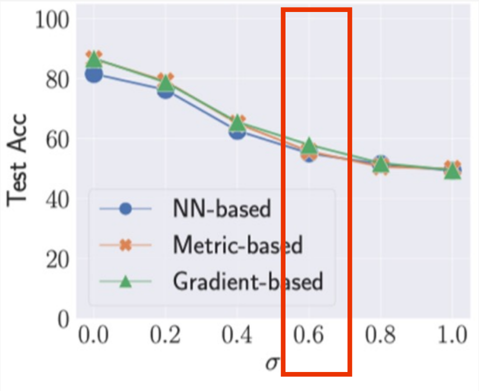
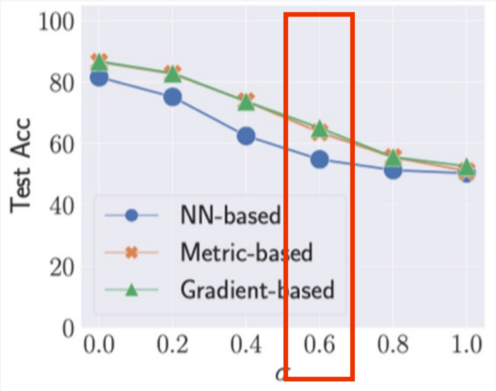


Figure 6: Attack performance of three membership inference attacks on four datasets.

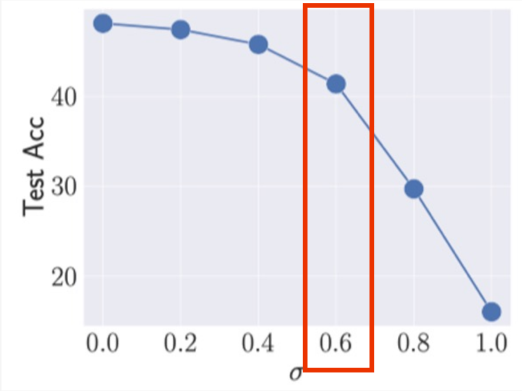
- Prompts are vulnerable to the membership inference attacks
- Metric-based attacks achieve the best performance in most cases, e.g., 93.20% on AFAD

MIA Defense





Naïve Attack Adaptive Attack



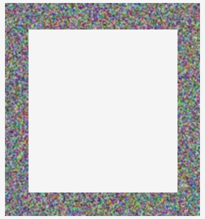
Prompt Utility

- Adding Gaussian noise to the prompts
- This defense mechanism can achieve a decent utility-defense trade-off when setting $\sigma = 0.6$

Property Inference Attacks (PIAs)

- Property inference attacks (PIAs): infer confidential properties of the training dataset that the PaaS provider does not intend to share

Trained Prompt



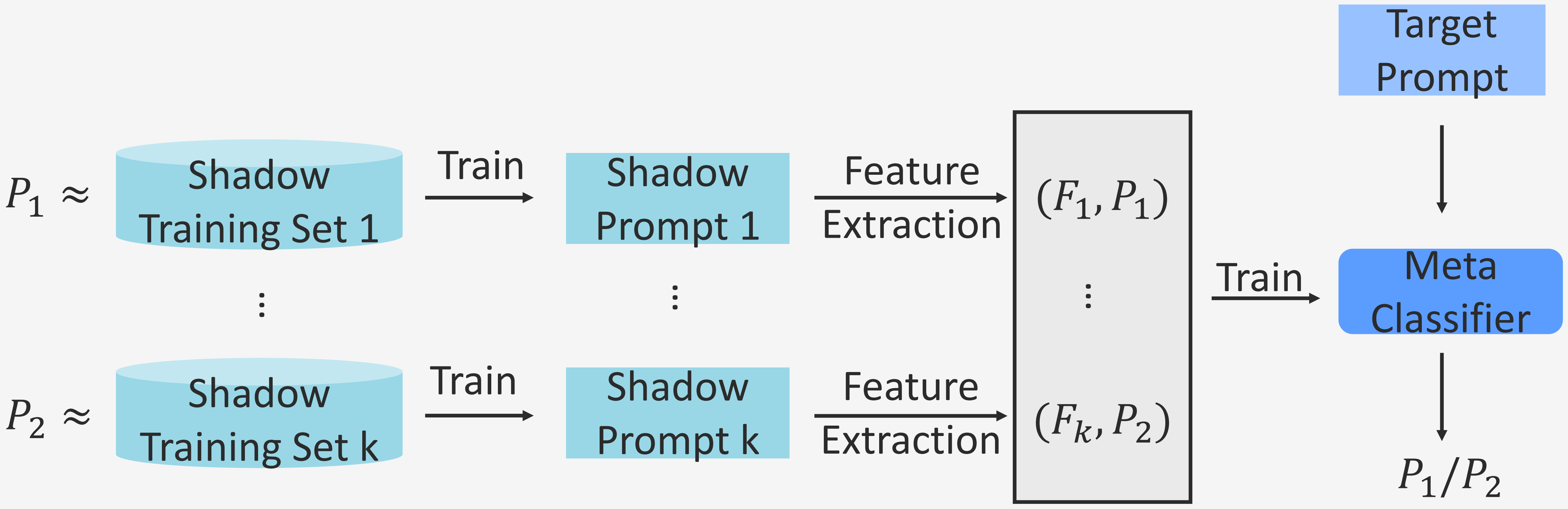
→ Property Inference →

P_1: Skewed Towards Males

Or

P_2: Skewed Towards Females

Workflow of PIA



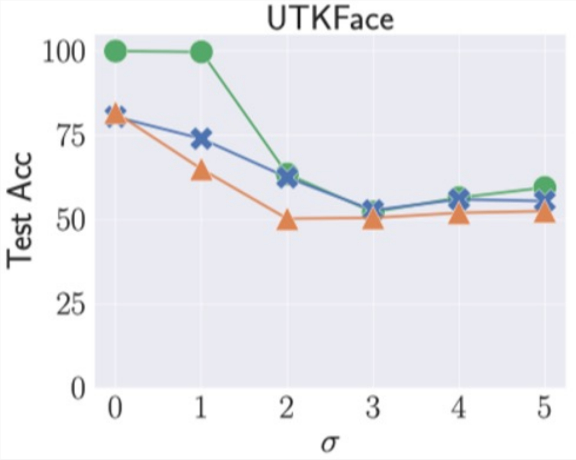
PIA Evaluation

Table 1: Experimental settings of the property inference attacks with the corresponding attack performance.

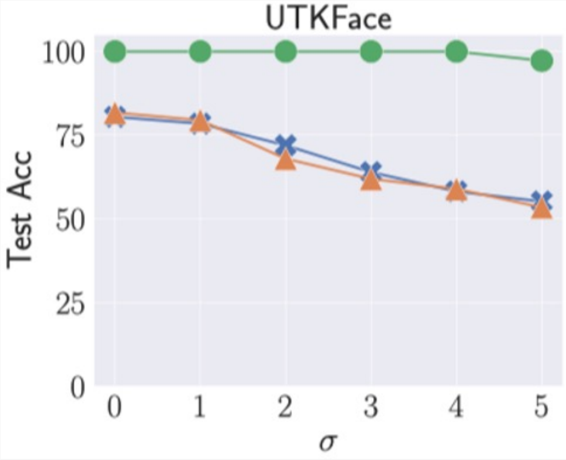
Inference Task	Dataset	Downstream Task	Target Property	Inference Labels	Test Accuracy RN18	BiT-M	ViT-B
T_1	CIFAR10	Image Classification	Size (T_1^{size})	{500, 2000}	100.00	100.00	100.00
T_2	CelebA	Multi-Atrribute Classification	Size (T_2^{size})	{500, 2000}	100.00	100.00	100.00
			Proportion of Males (T_2^{male})	{30%, 70%}	99.75	99.25	93.00
			Proportion of Youth (T_2^{youth})	{30%, 70%}	93.00	90.75	81.00
T_3	UTKFace	Race Classification	Size (T_3^{size})	{500, 2000}	100.00	100.00	100.00
			Proportion of Males (T_3^{male})	{30%, 70%}	80.50	80.50	82.00
			Proportion of Youth (T_3^{youth})	{30%, 70%}	81.75	87.50	84.00
T_4	AFAD	Age Classification	Size (T_4^{size})	{500, 2000}	100.00	100.00	100.00
			Proportion of Males (T_4^{male})	{30%, 70%}	80.75	78.00	72.25

- PIAs achieve good performance across different pre-trained models and datasets

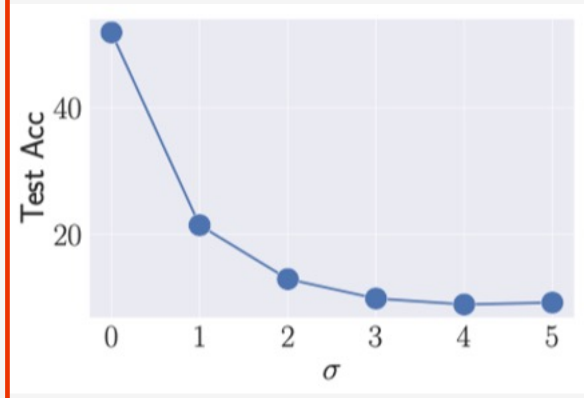
PIA Defense



Naïve Attack



Adaptive Attack



Prompt Utility

- With the increase of σ
  - The effectiveness of PIA significantly declines for naïve attacks
  - The target performance decreases by a large margin
- Fail to defend against property inference attacks

Conclusions

- We are the first to conduct comprehensive privacy assessment on visual prompt learning
- Our empirical evaluation shows that visual prompts are vulnerable to both membership inference attacks and property inference attacks
- Adding Gaussian noise to prompts, can mitigate the membership inference attacks with a decent utility-defense trade-off but fails to defend against property inference attacks
- Other conclusions can be found out in our paper
  - Overfitting affects the attack performance against visual prompt
  - Factors that affect these two attacks…

Contact

Thanks!

Yixin Wu

CISPA Helmholtz Center for Information Security

@yxoh28

https://yxoh.github.io/