Crafter: Facial Feature Crafting against Inversion-based Identity Theft on Deep Models

Shiming Wang, Zhe Ji, Liyao Xiang, Hao Zhang,
Xinbing Wang, Chenghu Zhou, Bo Li

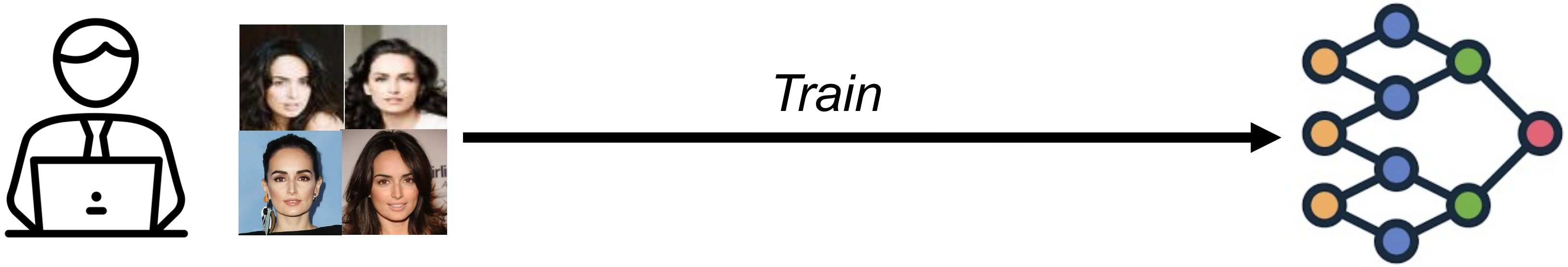
NDSS SYMPOSIUM/2024

Presented by Internet Society

#NDSSSymposium2024

Machine Learning as a Service

Example 1: Training deep learning task.

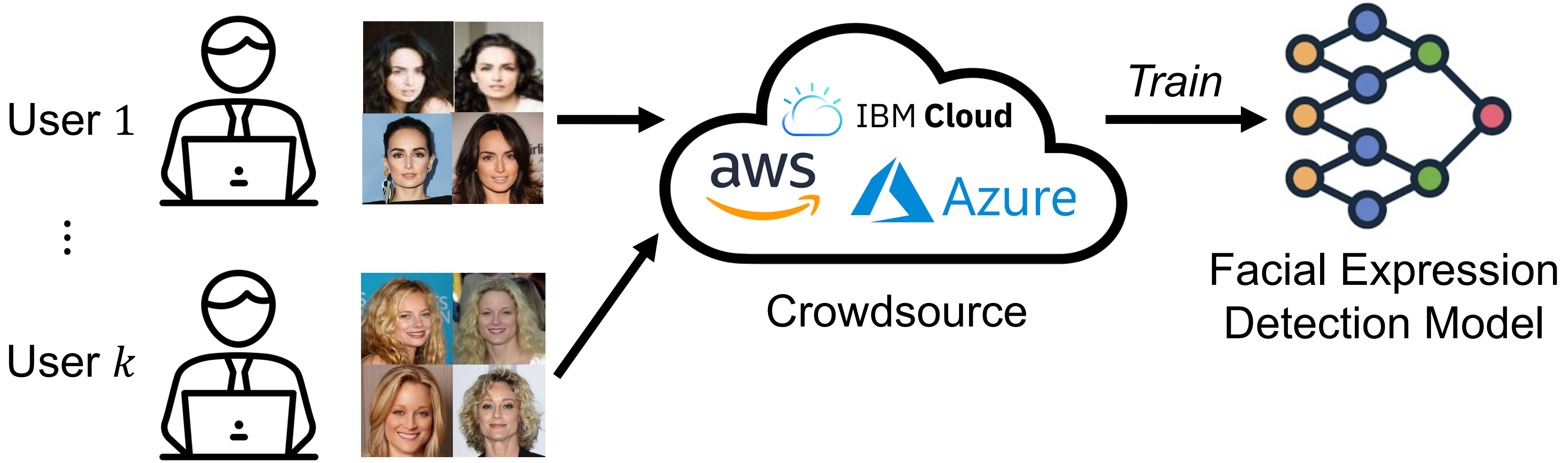


Limited data &
Limited computational resource

Facial Expression
Detection Model

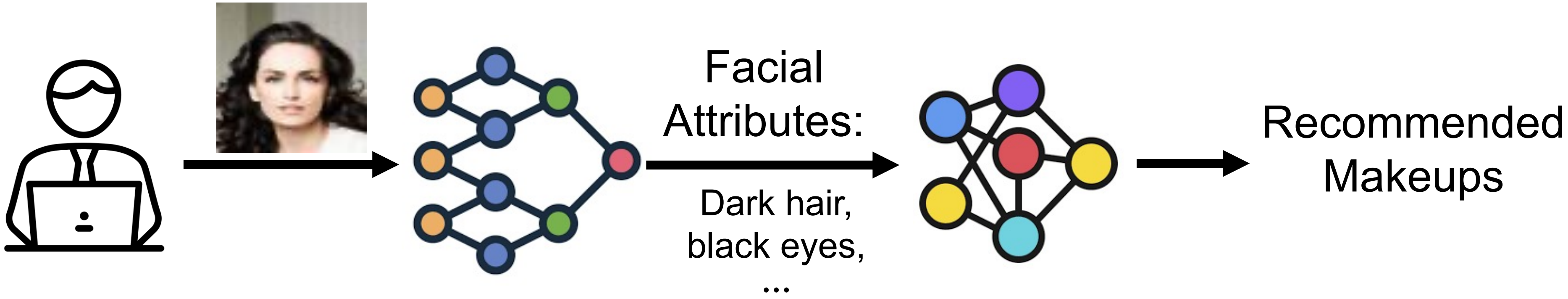
Machine Learning as a Service

Example 1: Training deep learning task.



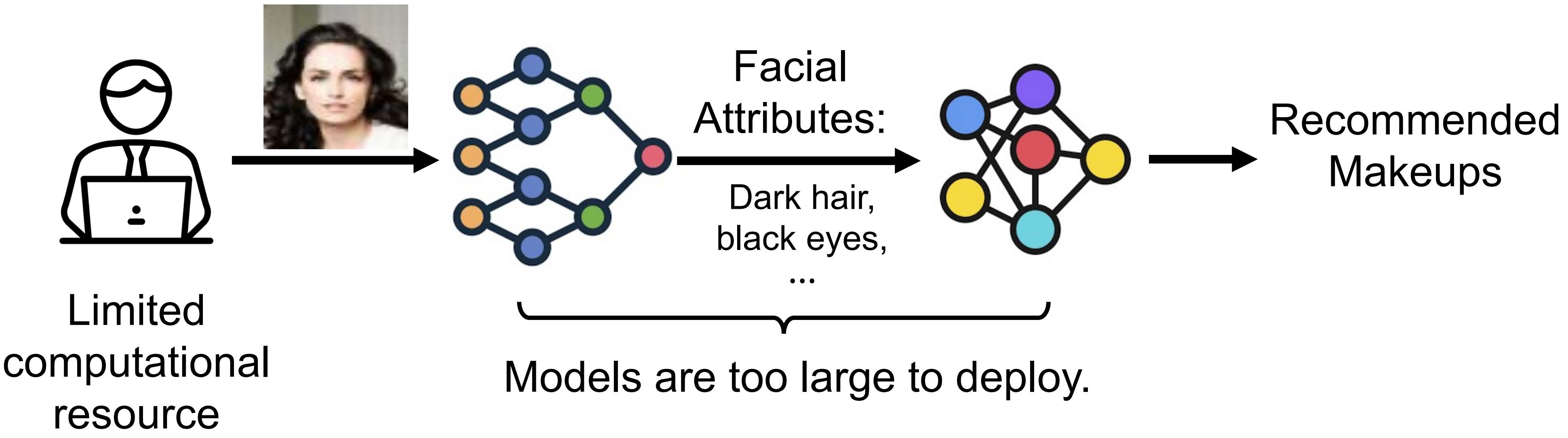
Machine Learning as a Service

Example 2: Inference deep learning task.



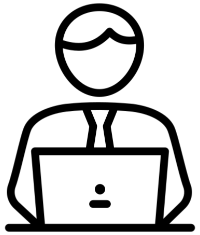
Machine Learning as a Service

Example 2: Inference deep learning task.



Machine Learning as a Service

Example 2: Inference deep learning task.



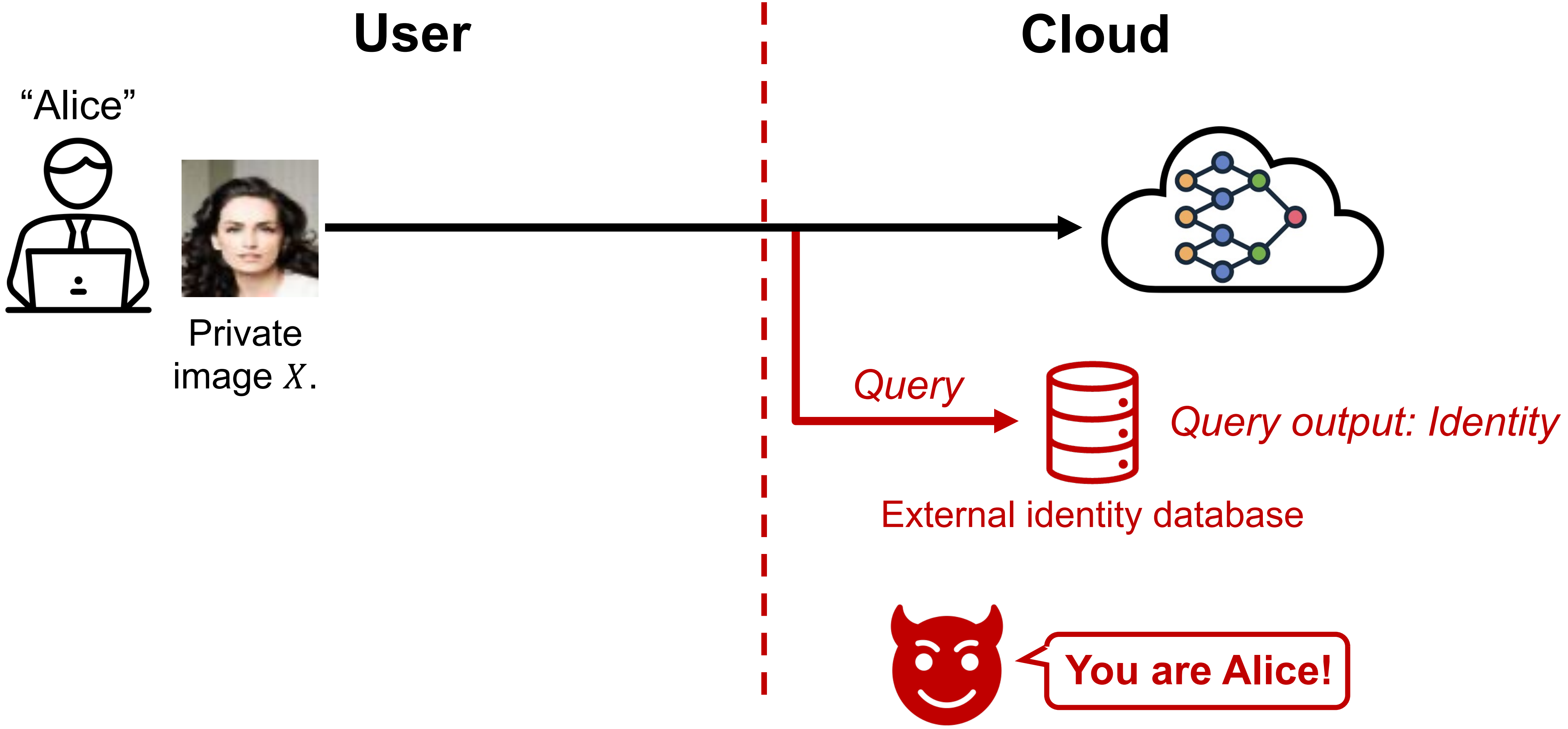
Limited computational resource

Recommended Makeups

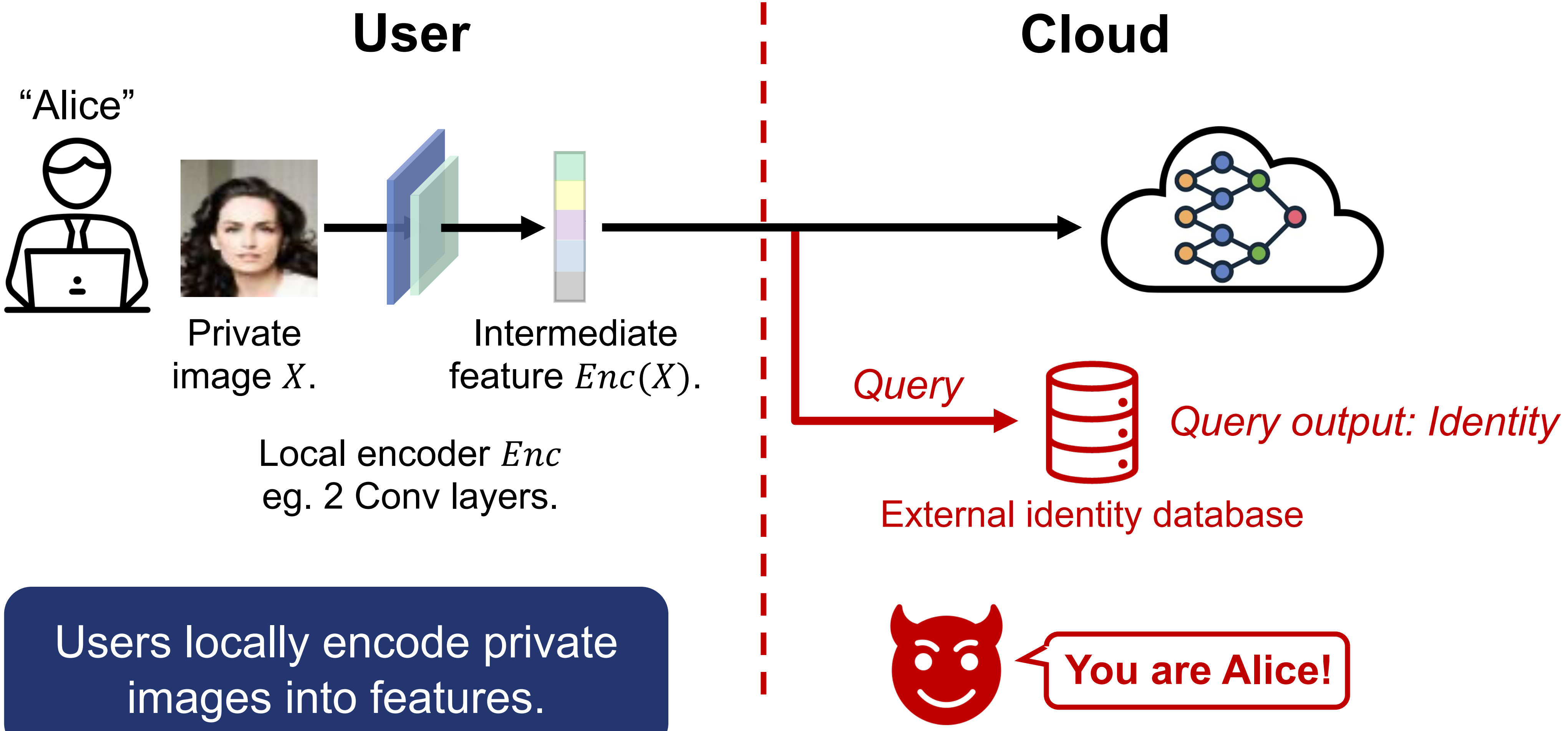
Machine Learning as a Service

Users are motivated to share their facial images with the cloud.

Privacy Concern: Identity Theft

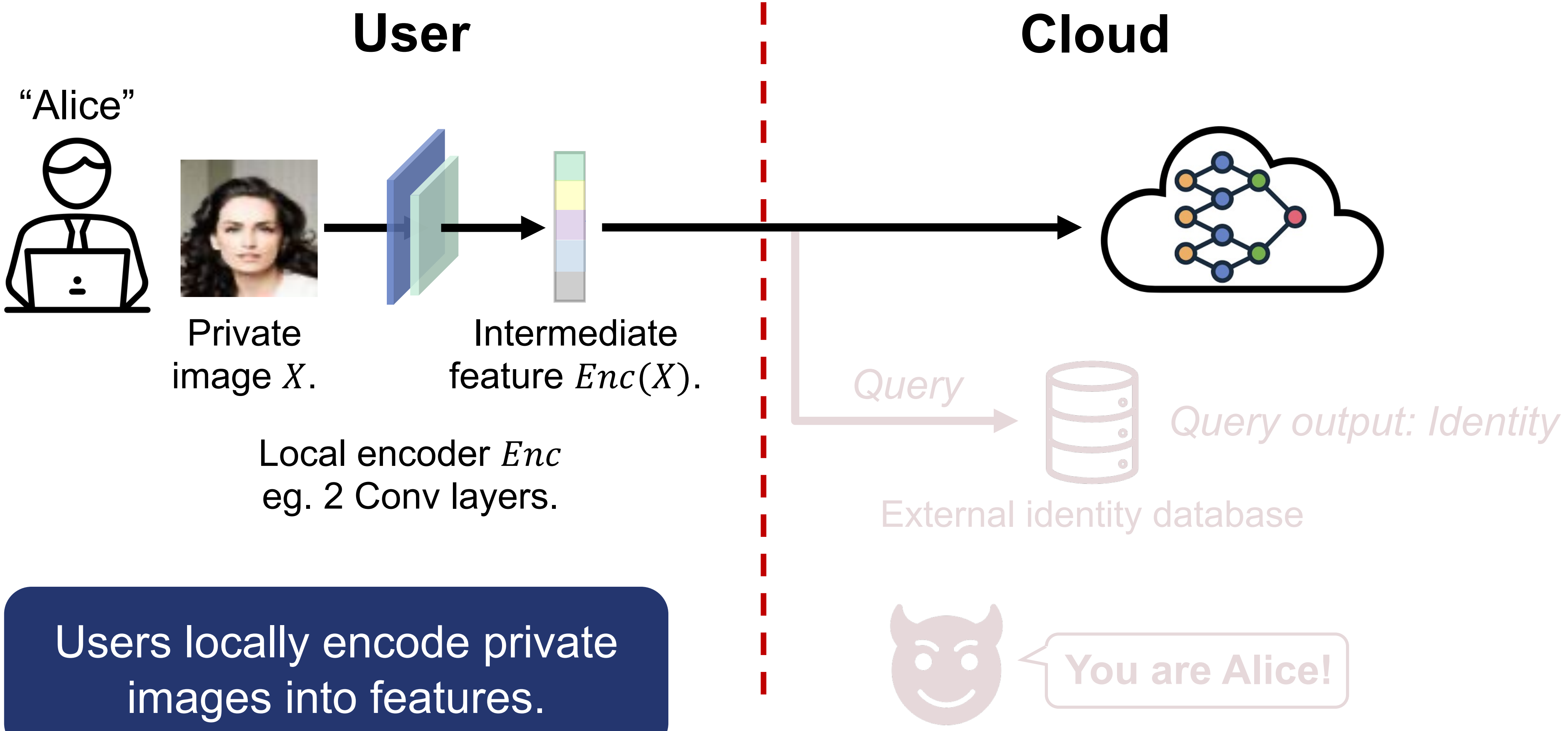


Privacy Concern: Identity Theft



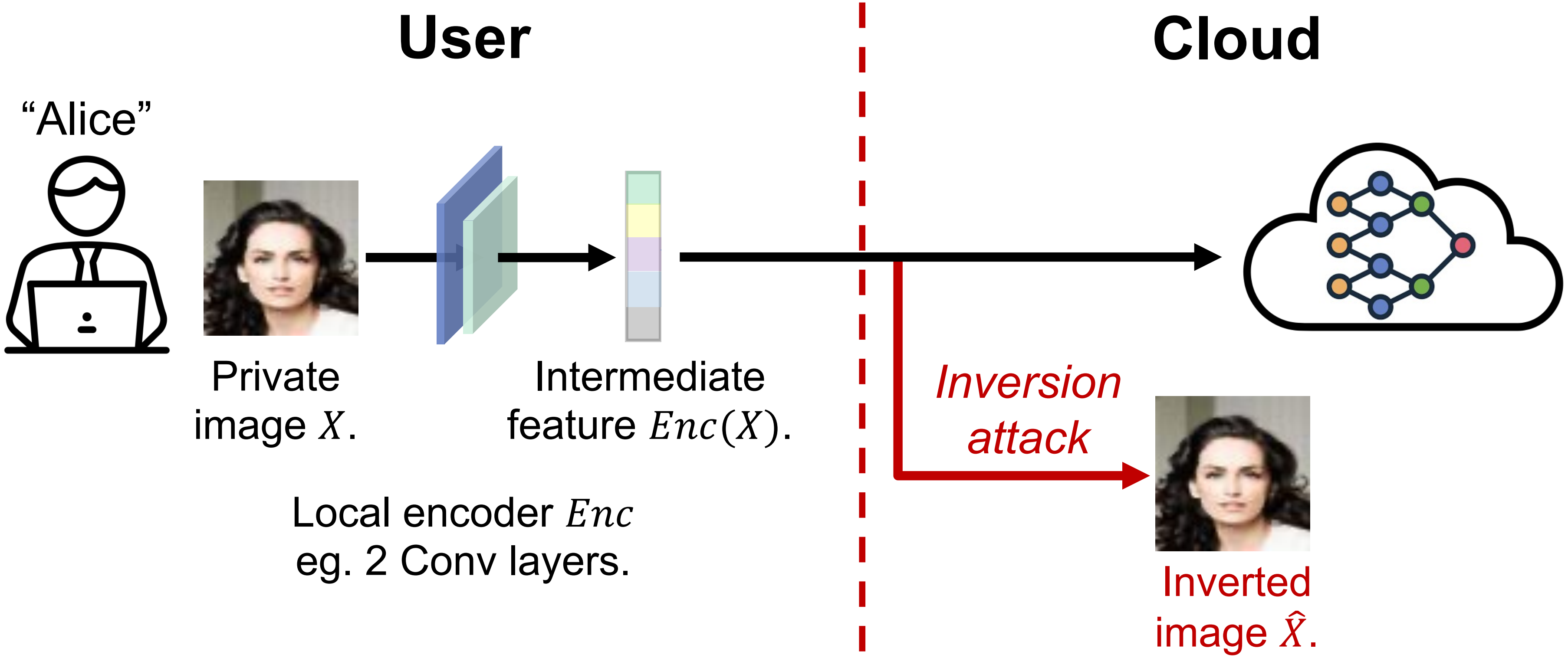
Users locally encode private images into features.

Privacy Concern: Identity Theft



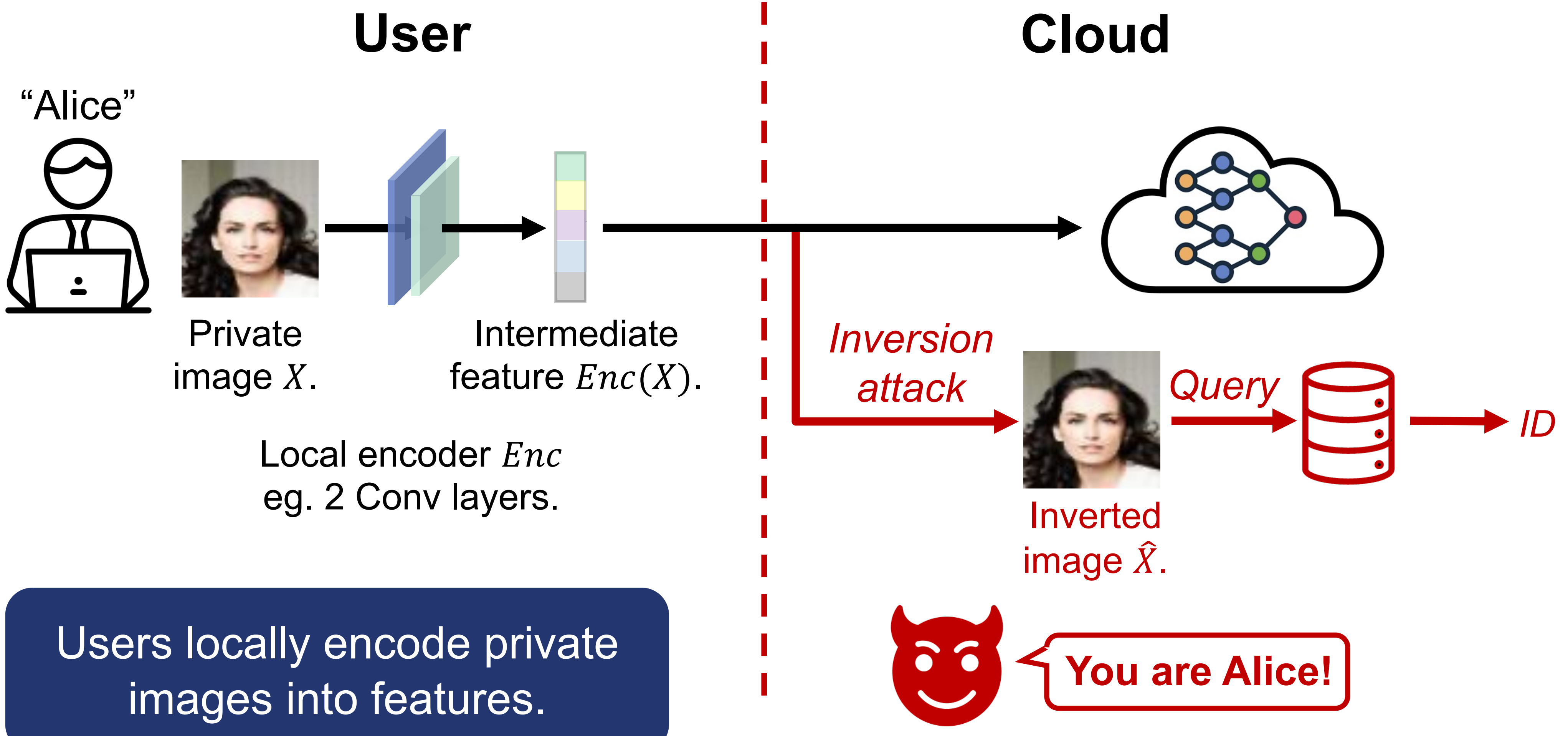
Users locally encode private images into features.

Privacy Concern: Identity Theft



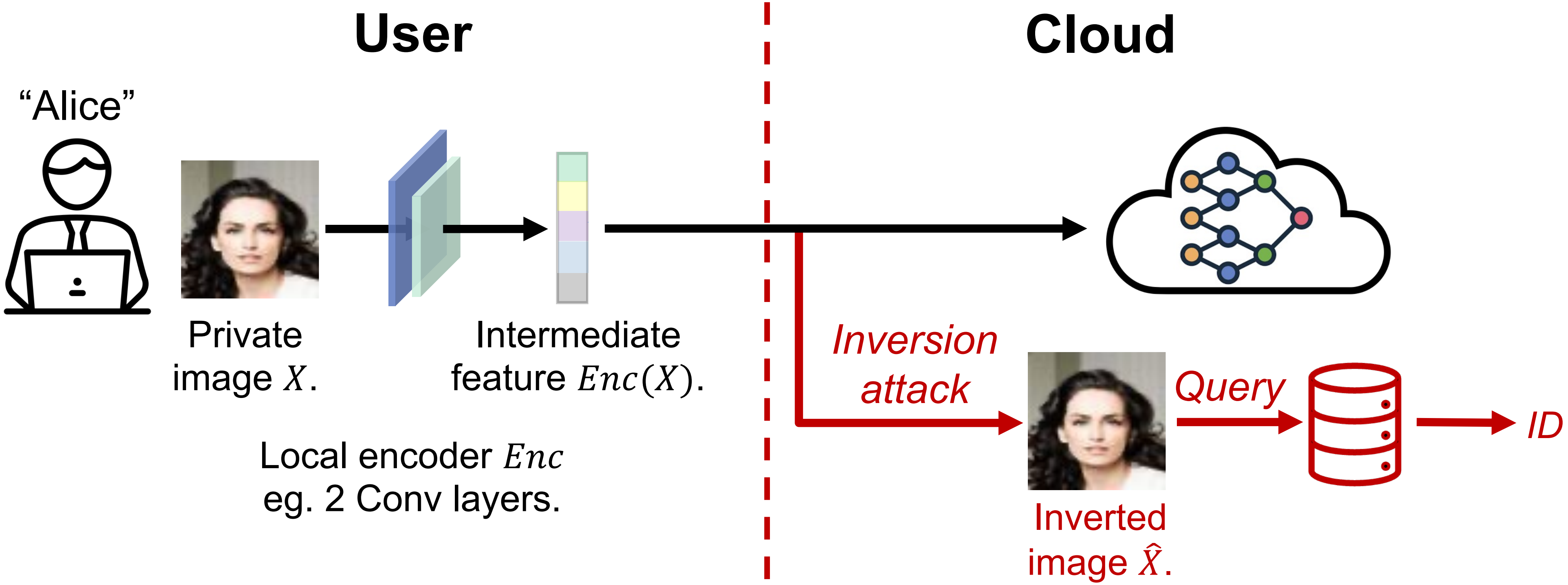
Users locally encode private images into features.

Privacy Concern: Identity Theft



Users locally encode private images into features.

Privacy Concern: Identity Theft



How do we prevent such inversion attacks?

Defending Inversion-based Identity Theft

Previous Defense:

AdvLearn[1], *Disco*[2], *TIPRDC*[3]

- Vulnerable against adaptive attacks;
- Fail to balance privacy & utility;
- Limited application scenarios.

[1] Xiao et al. "Adversarial learning of privacy-preserving and task-oriented representations ", 2020
[2] Singh et al. "Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks ", 2021
[3] Li et al. "Tiprdc: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations ", 2020

In Our Work

Crafter Defense:

User-end feature crafting that protects identity info against various inversion attacks, while preserving data utility.

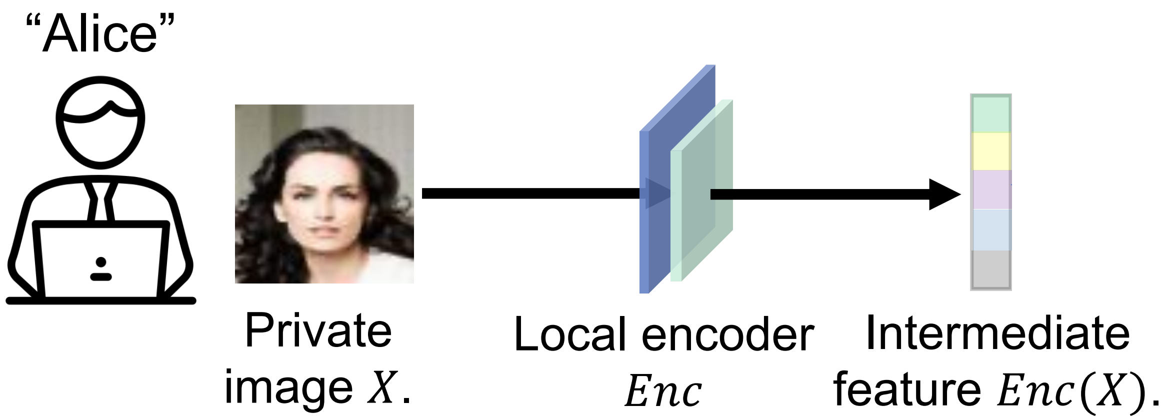
- Threat Model
- Intuitions & Design
- Evaluation

Threat Model

Black-box inversion attack:

- Access to public images; query access to the local Enc.

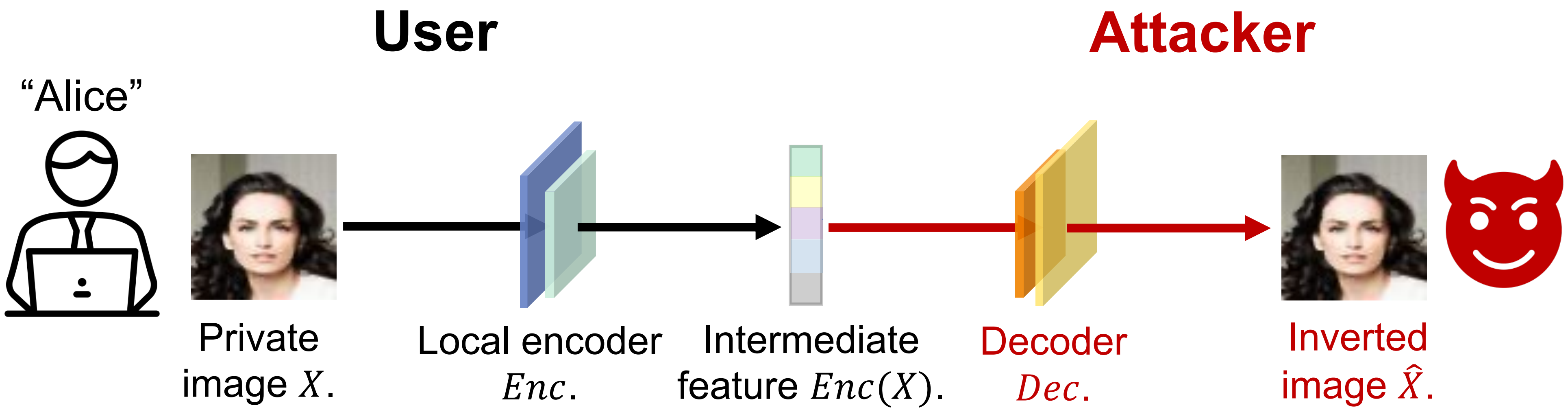
User



Threat Model

Black-box inversion attack:

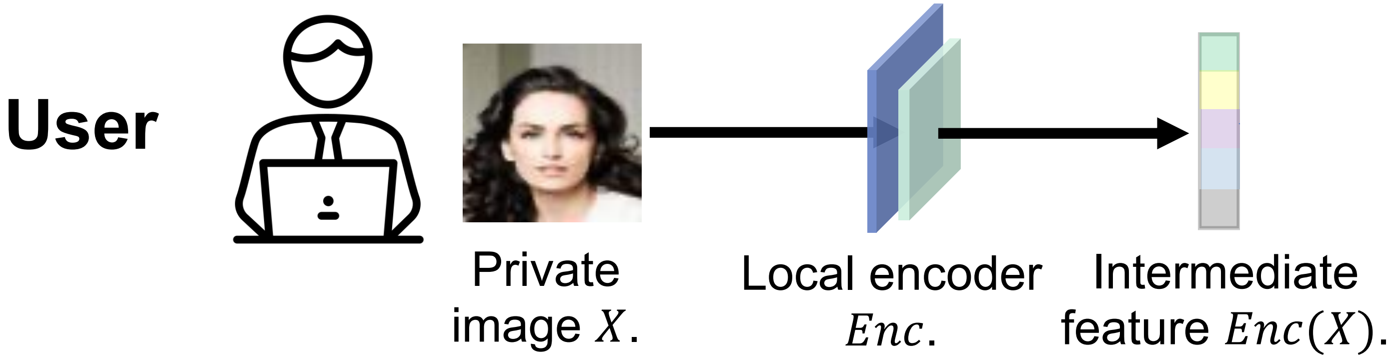
- Access to public images; query access to the local $Enc.$
- Train a decoder network $Dec.$



Threat Model

White-box inversion attack:

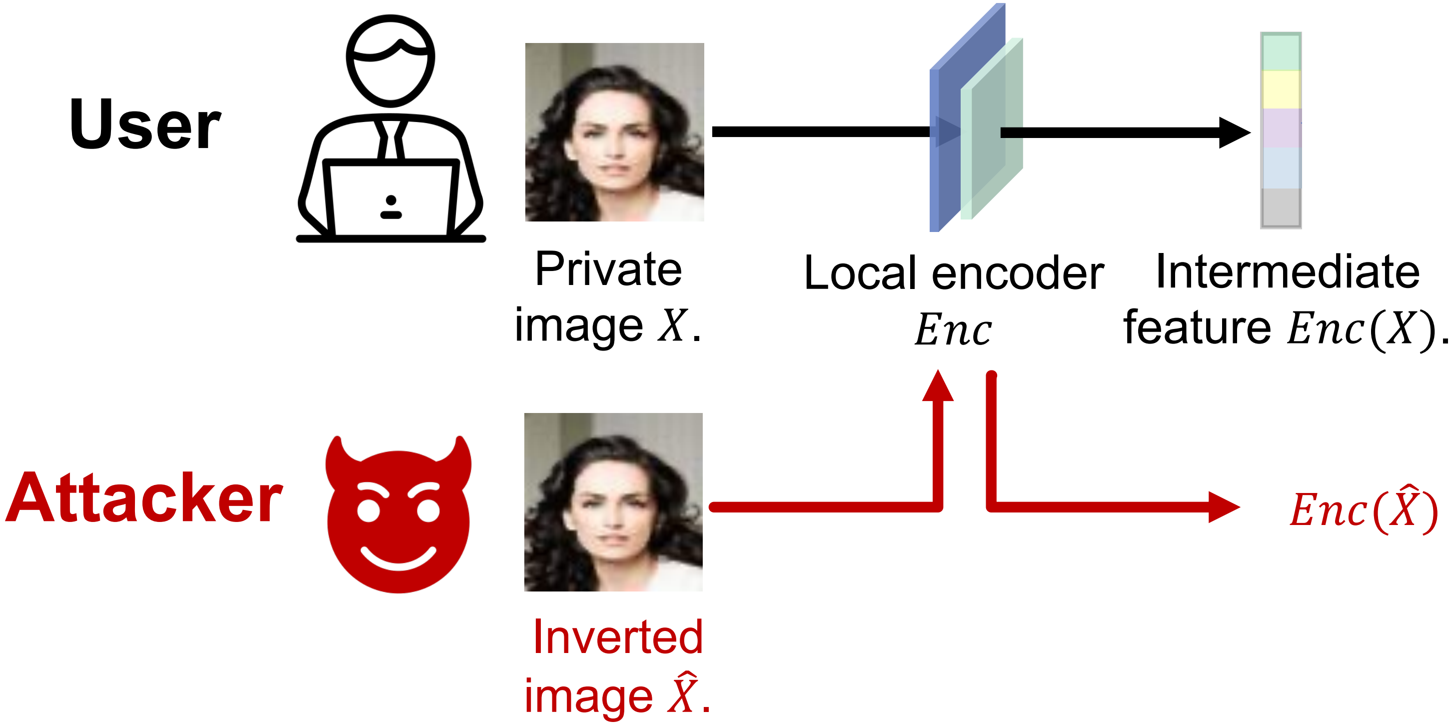
- Access to public images; access to the local Enc and its parameters.



Threat Model

White-box inversion attack:

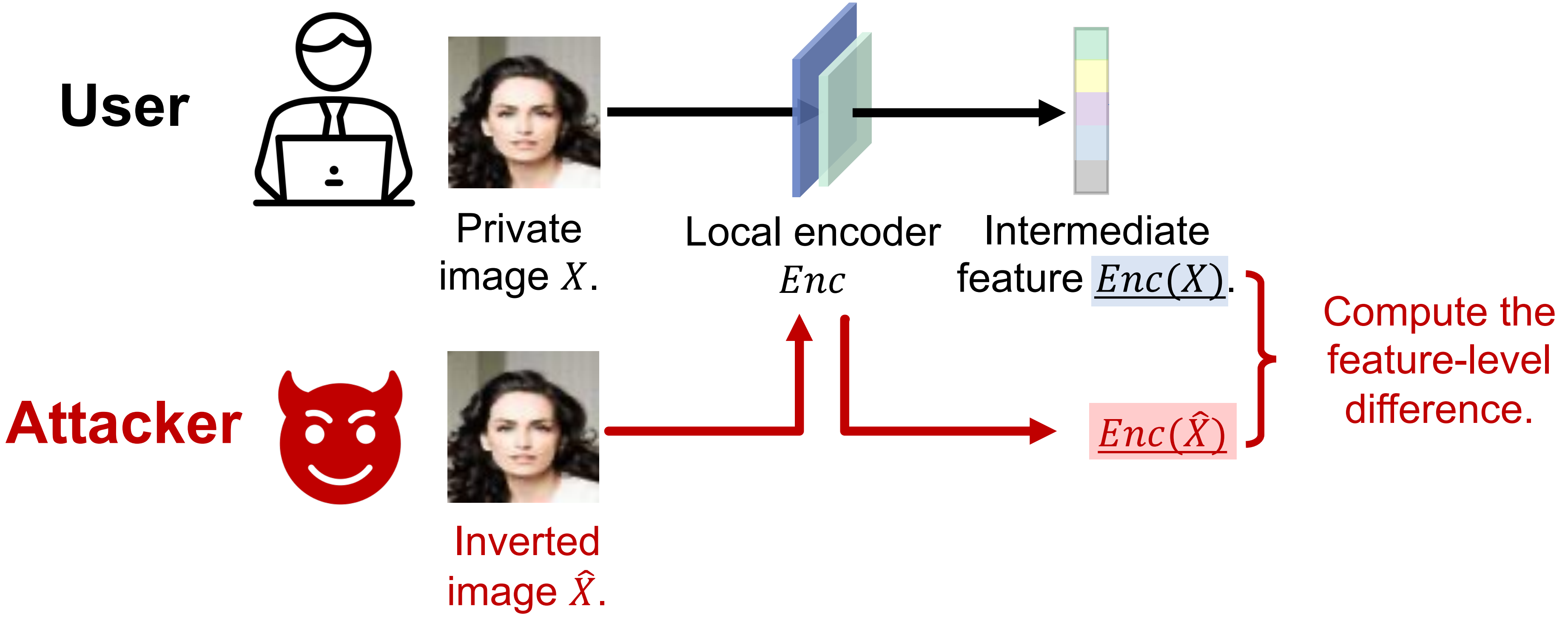
- Access to public images; access to the local Enc and its parameters.
- Optimize over the inverted image.



Threat Model

White-box inversion attack:

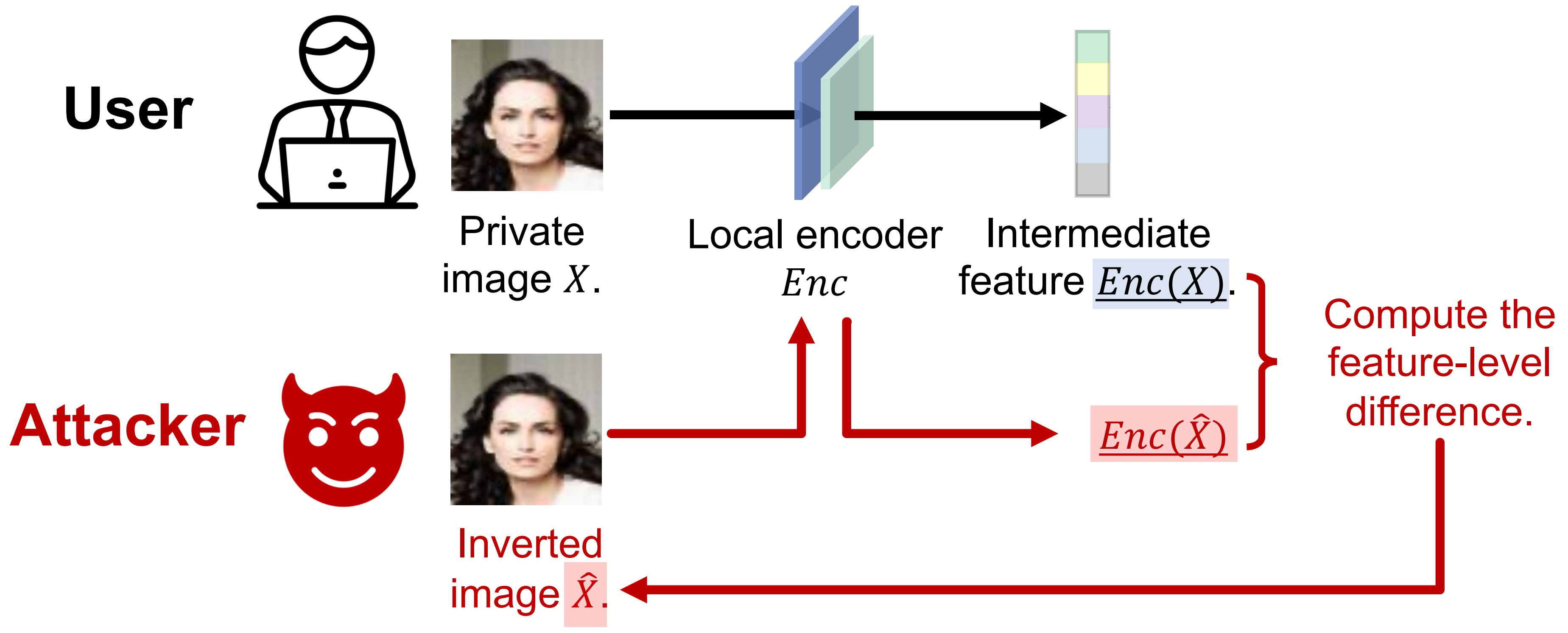
- Access to public images; access to the local Enc and its parameters.
- Optimize over the inverted image.



Threat Model

White-box inversion attack:

- Access to public images; access to the local Enc and its parameters.
- Optimize over the inverted image.

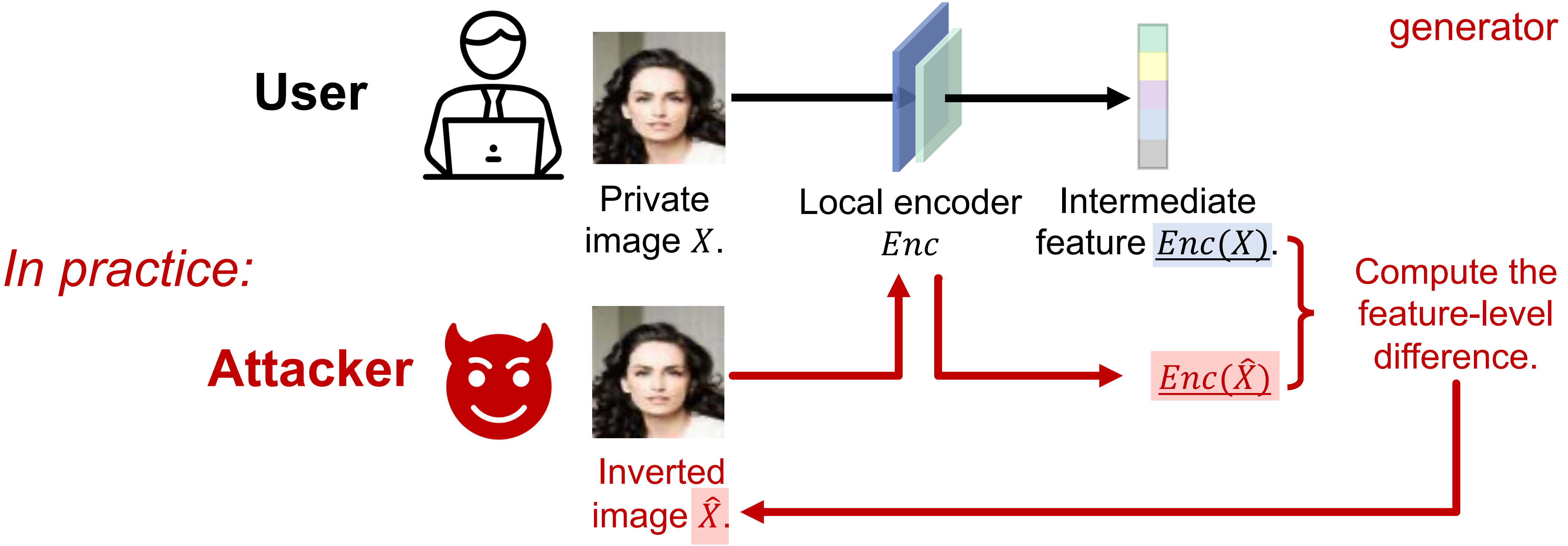


Threat Model

White-box inversion attack:

- Access to public images; access to the local Enc and its parameters.
- Optimize over the inverted image.

+ Pretrained public generator G.

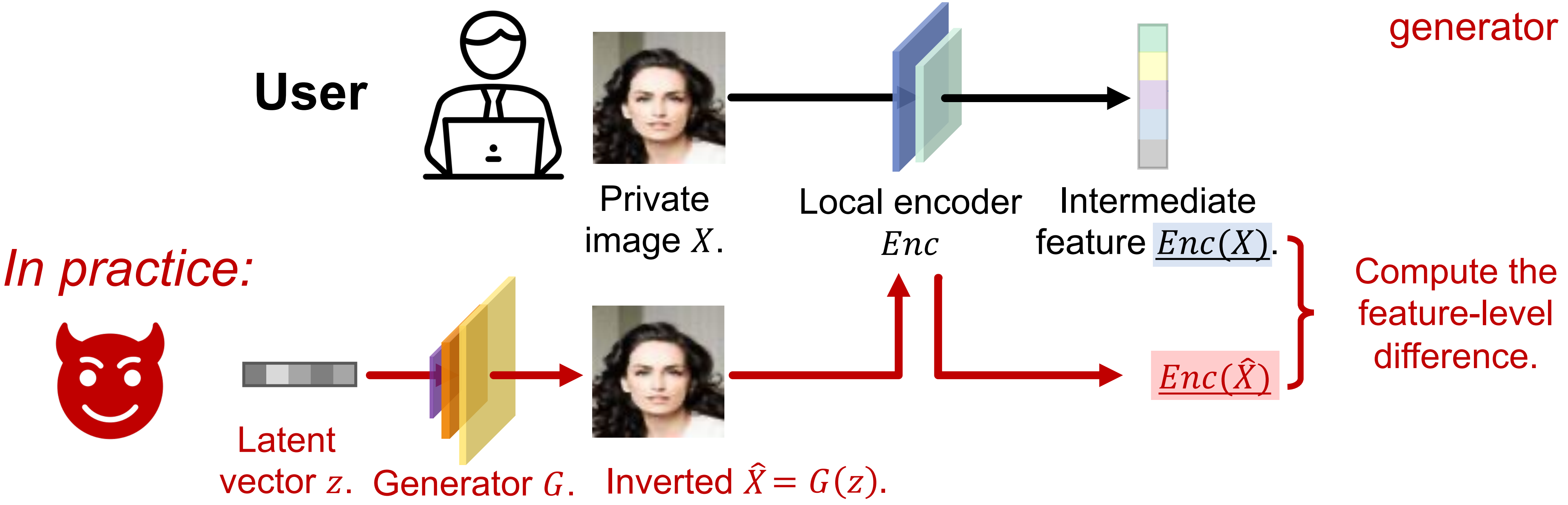


Threat Model

White-box inversion attack:

- Access to public images; access to the local Enc and its parameters.
- Optimize over the inverted image.

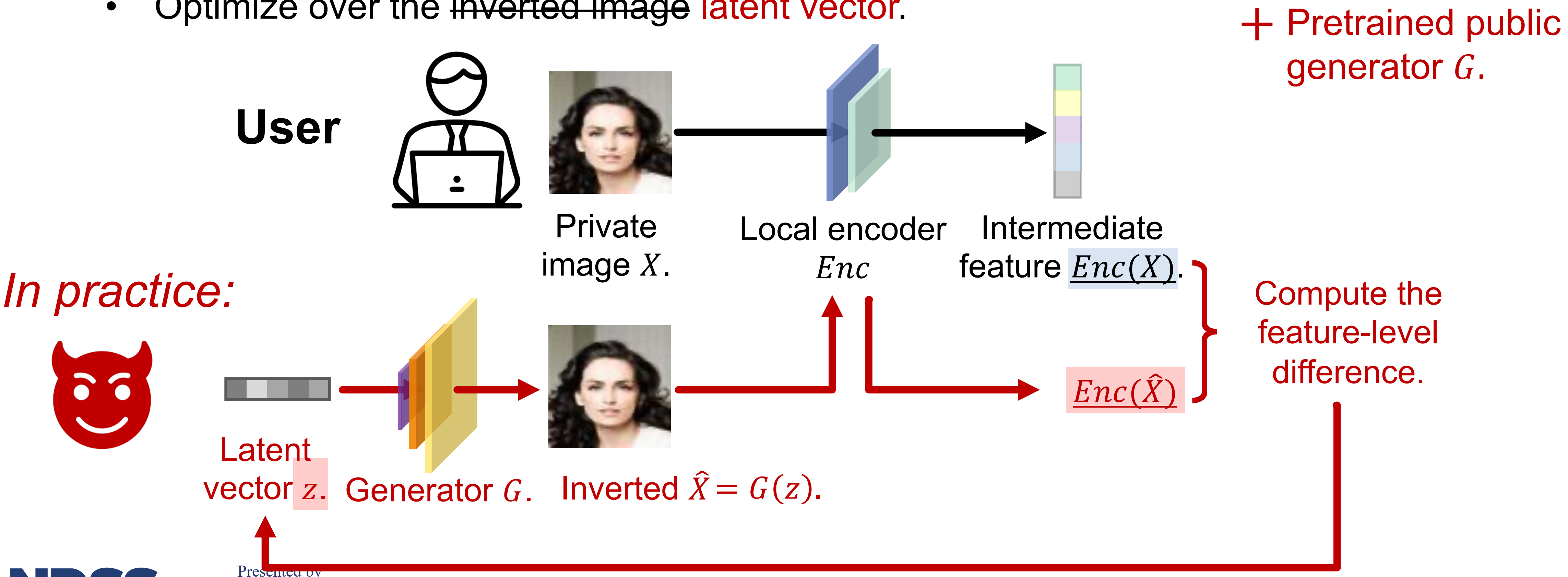
+ Pretrained public generator G.



Threat Model

White-box inversion attack:

- Access to public images; access to the local Enc and its parameters.
- Optimize over the ~~inverted image~~ latent vector.



Defense Intuitions

Privacy goal: Inverted image does not look like Alice.

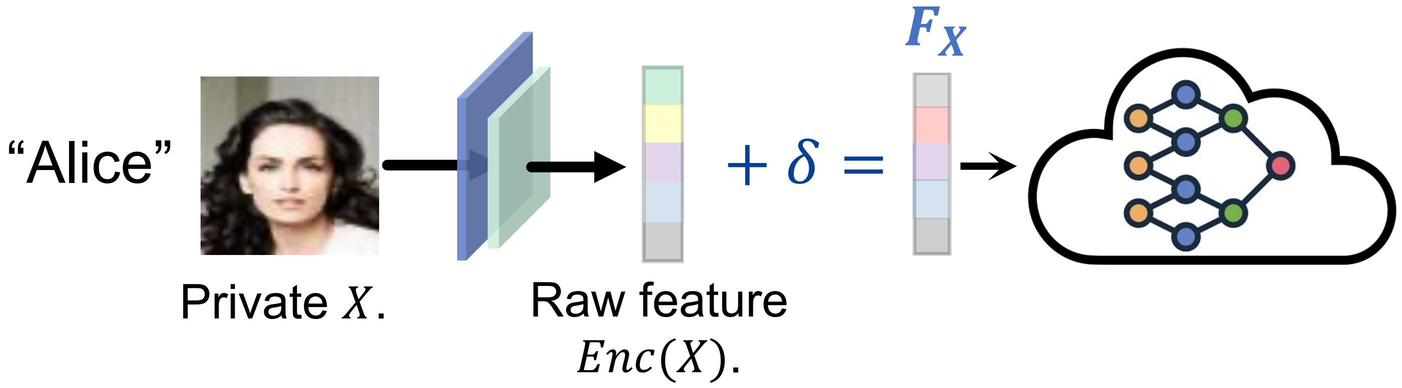
Utility goal: Feature completes cloud tasks well.

Defense Intuitions

Privacy goal: Inverted image does not look like Alice.

Utility goal: Feature completes cloud tasks well.

General intuition: Perturb the feature.



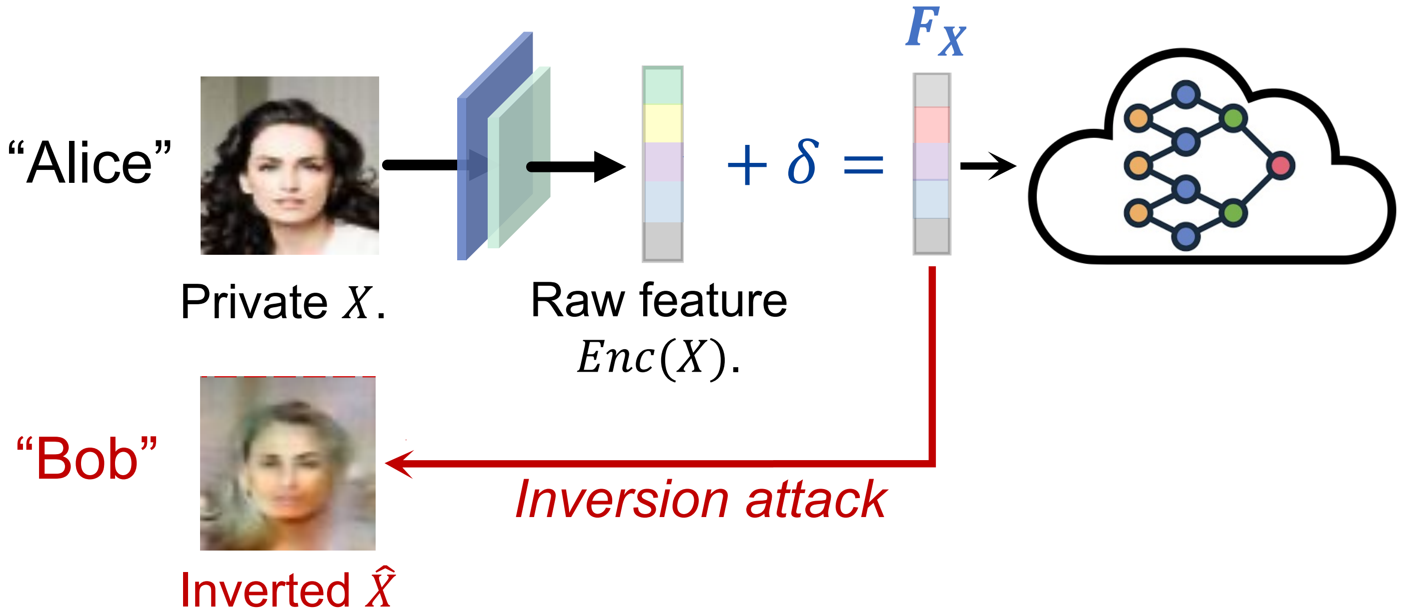
Defense Intuitions

Privacy goal: Inverted image does not look like Alice.

Utility goal: Feature completes cloud tasks well.

General intuition: Perturb the feature.

- (*Privacy*) Mislead a simulated inversion attacker.



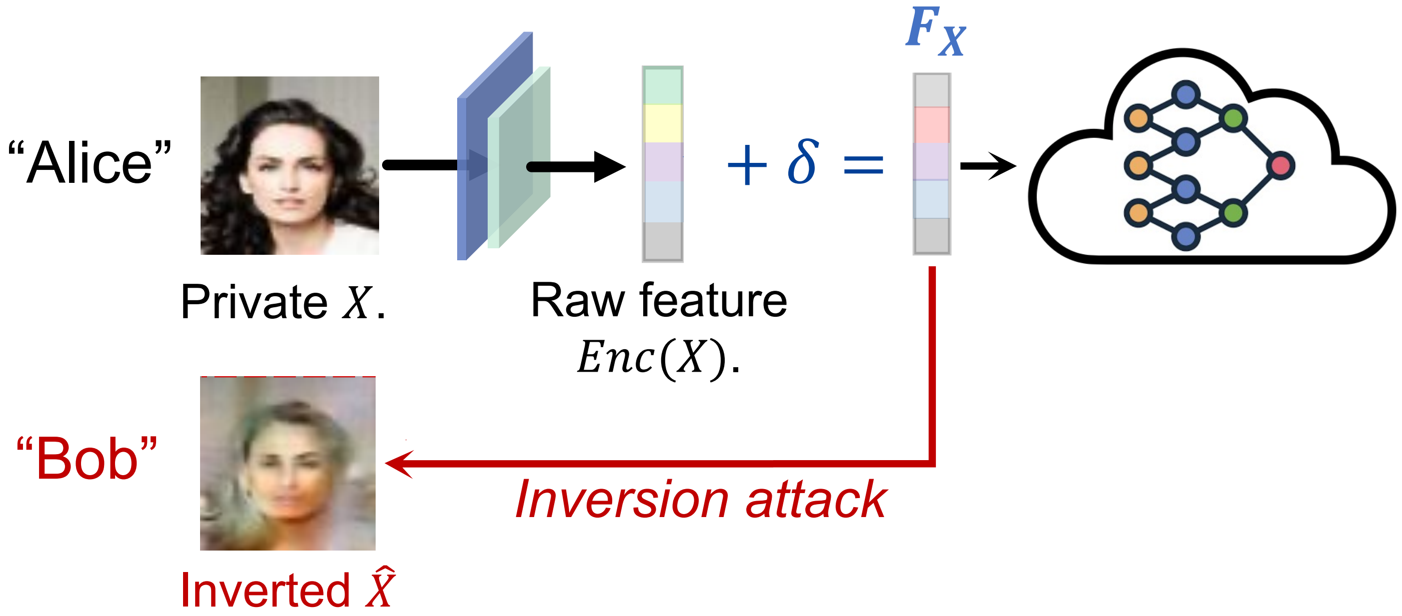
Defense Intuitions

Privacy goal: Inverted image does not look like Alice.

Utility goal: Feature completes cloud tasks well.

General intuition: Perturb the feature.

- (*Privacy*) Mislead a simulated inversion attacker.
- (*Utility*) Keep the perturbation small.



Defense Intuitions (Utility)

Utility loss: $L_{utility}$ = perturbation magnitude.

Preserves utility: Cloud model is robust against minor perturbation.

Utility task agnostic: L_{utility} independent from cloud model

→ deployable as a plug-in.

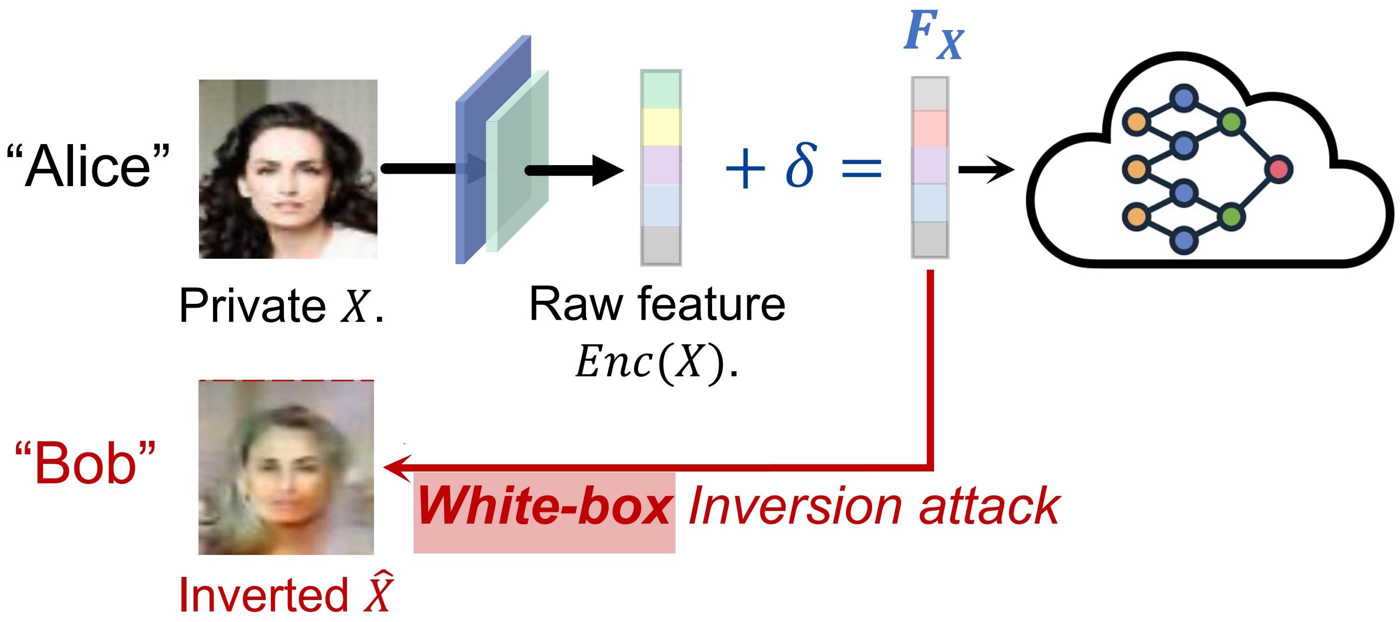
Defense Intuitions (Privacy)

Challenge 1: *Robust against both black- & white-box inversion*.

Defense Intuitions (Privacy)

Challenge 1: *Robust against both black- & white-box inversion.*

Intuition: White-box attack is stronger; simulate a **white-box attacker**.



Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

If a defense is not robust: false security, meaningless!

Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

If a defense is not robust: false security, meaningless!

Previous defense: Push the attacker away from the private image.

"Stay Away"

Tit for tat between attacker & defense.

Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

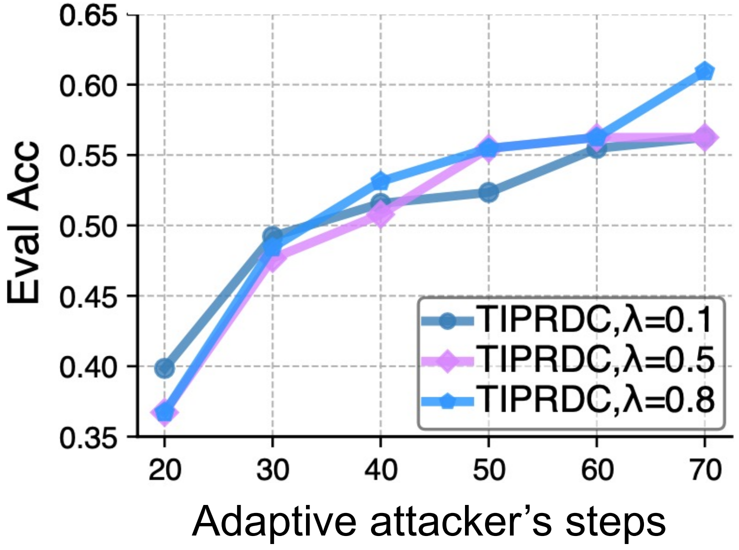
If a defense is not robust: false security, meaningless!

Previous defense: Push the attacker away from the private image.

"Stay Away"

Tit for tat between attacker & defense.

Weaker privacy



Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

If a defense is not robust: false security, meaningless!

Previous defense: Push the attacker away from the private image.

~~"Stay Away"~~ Tit for tat between attacker & defense.

Why is "Stay Away" vulnerable against adaptive attacks?

Defense Intuitions (Privacy)

A game view:

Attack →
Defense →

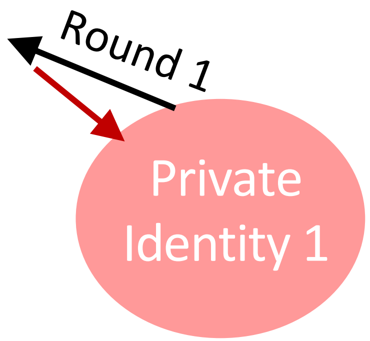
Defense Intuitions (Privacy)

A game view:

Attack

Defense

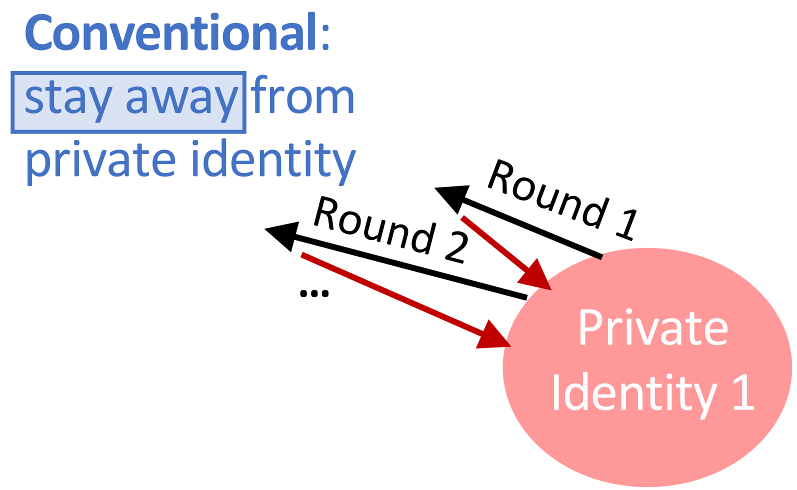
Conventional:
stay away from
private identity



Defense Intuitions (Privacy)

A game view:

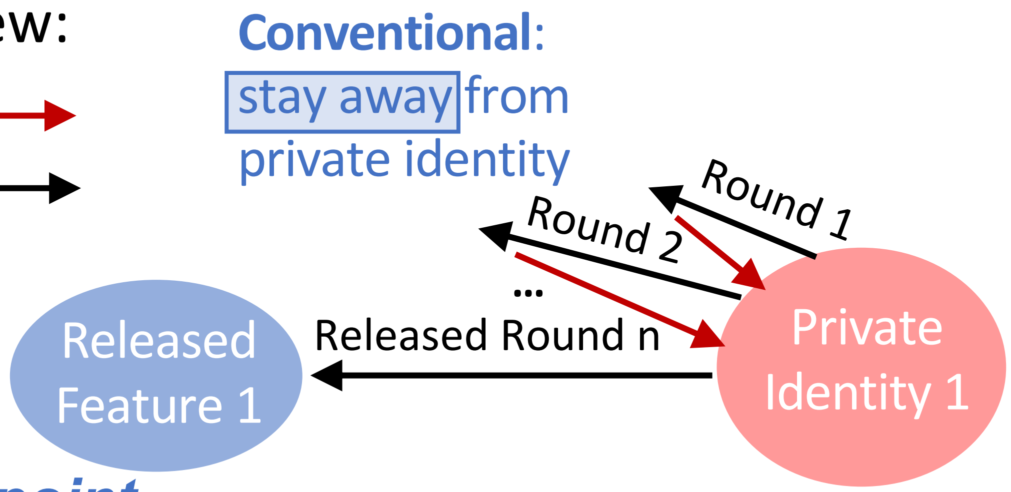
Attack
Defense



Defense Intuitions (Privacy)

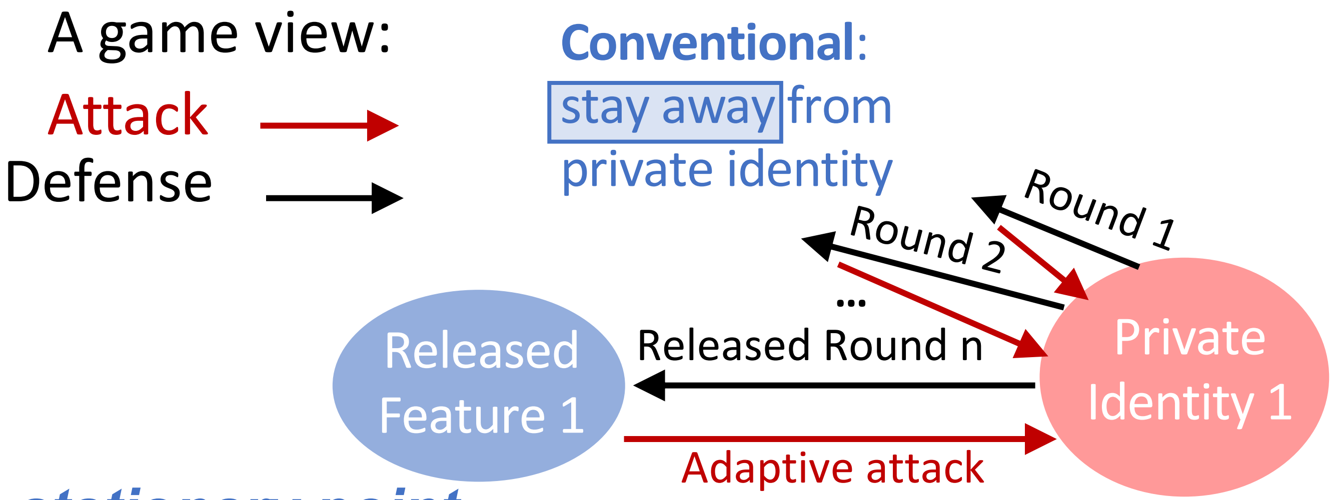
A game view:

Attack
Defense



A stationary point.
But not equilibrium (there is NO equilibrium in reality).

Defense Intuitions (Privacy)



A stationary point.
But not equilibrium (there is NO equilibrium in reality).

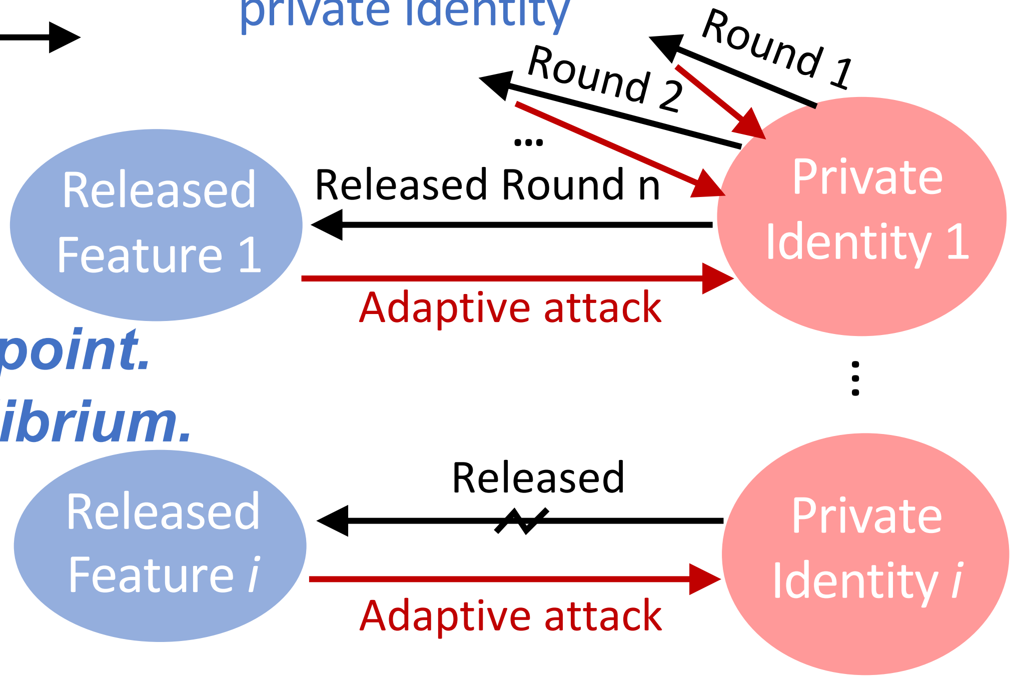
Defense Intuitions (Privacy)

A game view:

Attack →

Defense →

Conventional:
stay away from
private identity



A stationary point.
But not equilibrium.

Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

If a defense is not robust: false security, meaningless!

Our Intuition: Limit attacker's **knowledge gain** from the exposed feature.

"Get Close"

Defense Intuitions (Privacy)

Challenge 2: *Robust against adaptive attacks.*

Attacker tries to bypass a fixed defense.

If a defense is not robust: false security, meaningless!

Our Intuition: Limit attacker's **knowledge gain** from the exposed feature.

"Get Close" = **Prior** vs. **Posterior**

Defense Intuitions (Privacy)

Prior: “*Average face*”, public face distribution**.**

Defense Intuitions (Privacy)

Prior: "*Average face*", public face distribution**.**

Contains no private ID info.

Defense Intuitions (Privacy)

Prior: "*Average face*", public face distribution**.**

Contains no private ID info.

We use: $G(z_{random})$

Public generator

Random latent vectors

Defense Intuitions (Privacy)

Prior: "*Average face*", public face distribution**.**

Contains no private ID info.

We use: $G(z_{random})$

Public generator　Random latent vectors

Posterior: Image $\hat{X}$ inverted from feature F_X.

Defense Intuitions (Privacy)

“Get Close” : Minimize distance between prior & posterior.

Defense Intuitions (Privacy)

"Get Close" : Minimize distance between prior & posterior.

We use: Earth-Mover distance EMD.

Privacy loss: $L_{privacy} = EMD$ between inverted image $\hat{X}$ & average face.

Defense Intuitions (Privacy+Utility)

"Get Close" : Minimize distance between prior & posterior.

We use: Earth-Mover distance EMD.

Privacy loss: $L_{privacy} = EMD$ between inverted image $\hat{X}$ & average face.

Combine utility & privacy: $L_{combined} = \beta \cdot L_{privacy} + L_{utility}$

Defense Intuitions (Privacy+Utility)

"Get Close" : Minimize distance between prior & posterior.

We use: Earth-Mover distance EMD.

Privacy loss: $L_{privacy} = EMD$ between inverted image $\hat{X}$ & average face.

Combine utility & privacy: $L_{combined} = \beta \cdot L_{privacy} + L_{utility}$

Find a feature perturbation that 1) is small;

Defense Intuitions (Privacy+Utility)

"Get Close" : Minimize distance between prior & posterior.

We use: Earth-Mover distance EMD.

Privacy loss: $L_{privacy} = EMD$ between inverted image $\hat{X}$ & average face.

Combine utility & privacy: $L_{combined} = \beta \cdot L_{privacy} + L_{utility}$

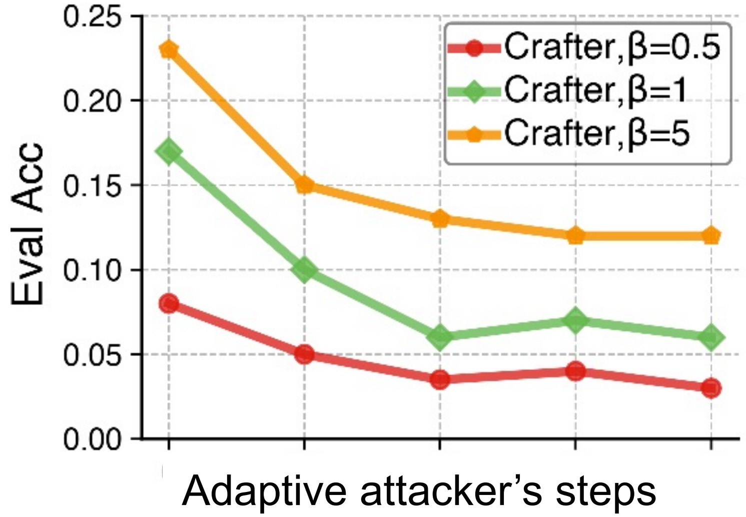
Find a feature perturbation that 1) is small;

2) draws inverted image close to public average faces.

Defense Intuitions (Privacy+Utility)

"Get Close" : Minimize distance between prior & posterior.

Stronger privacy

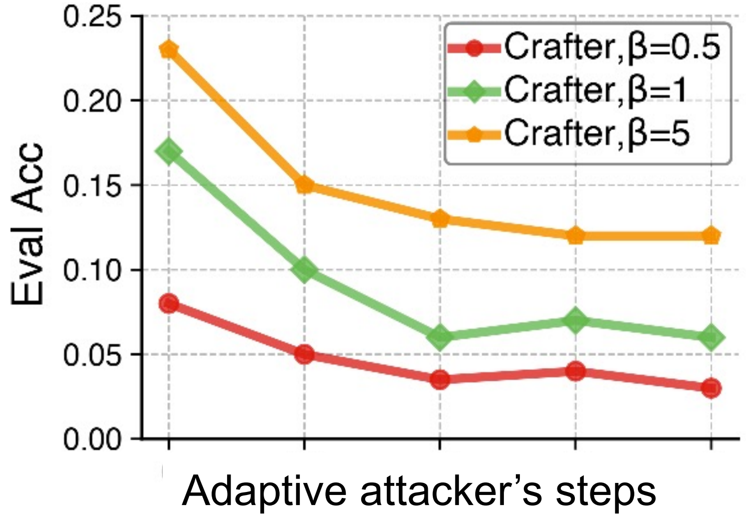


Adaptive attackers can only get worse!

Defense Intuitions (Privacy+Utility)

"Get Close" : Minimize distance between prior & posterior.

Stronger privacy



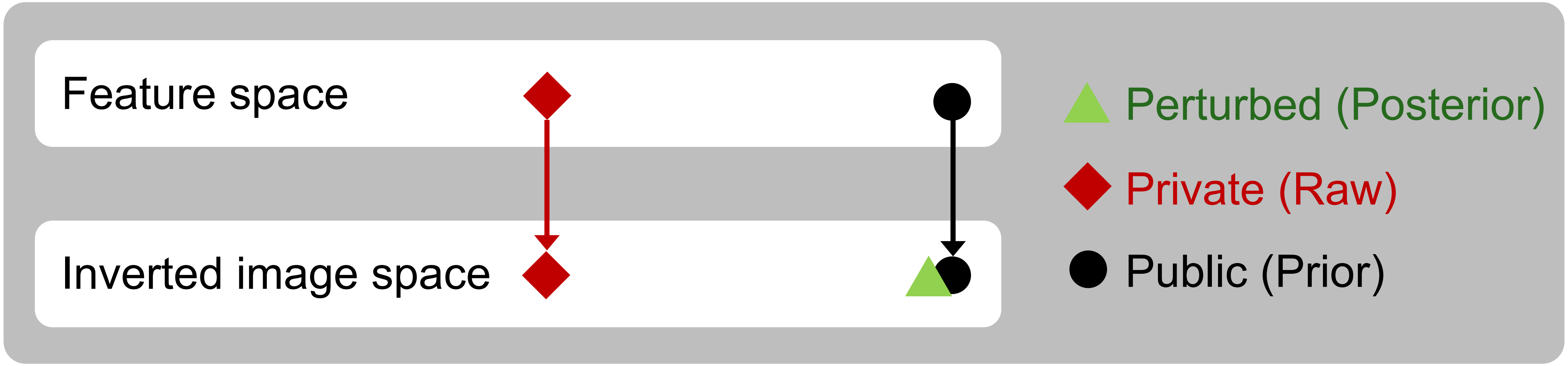
Why is "Get Close" robust against adaptive attacks?

Defense Intuitions (Privacy+Utility)

“Get Close”

Find a feature perturbation that

1) is small;
2) draws inverted image close to public average faces.



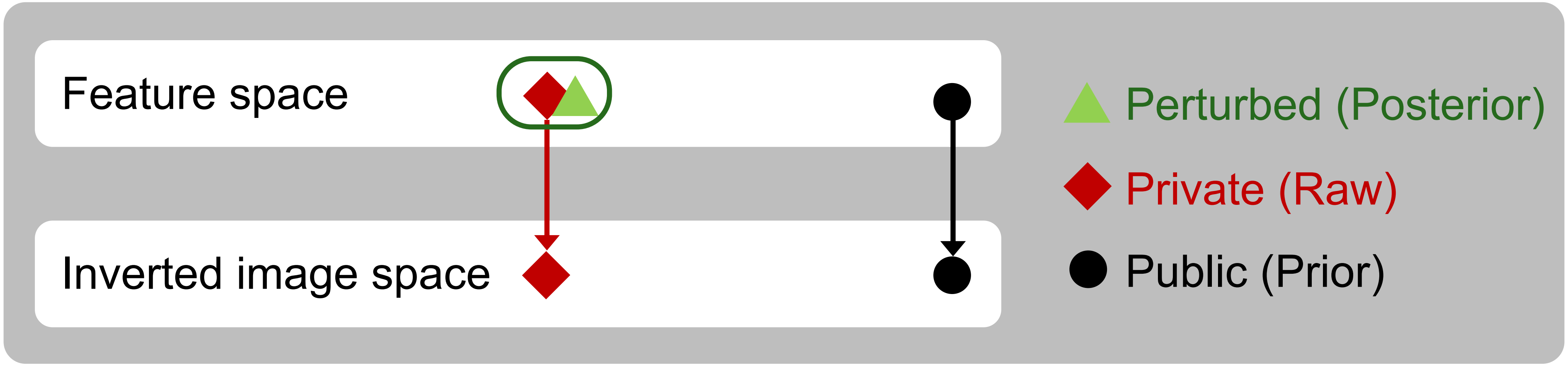
Defense Intuitions (Privacy+Utility)

“Get Close”

Find a feature perturbation that

1) is small;

2) draws inverted image close to public average faces.



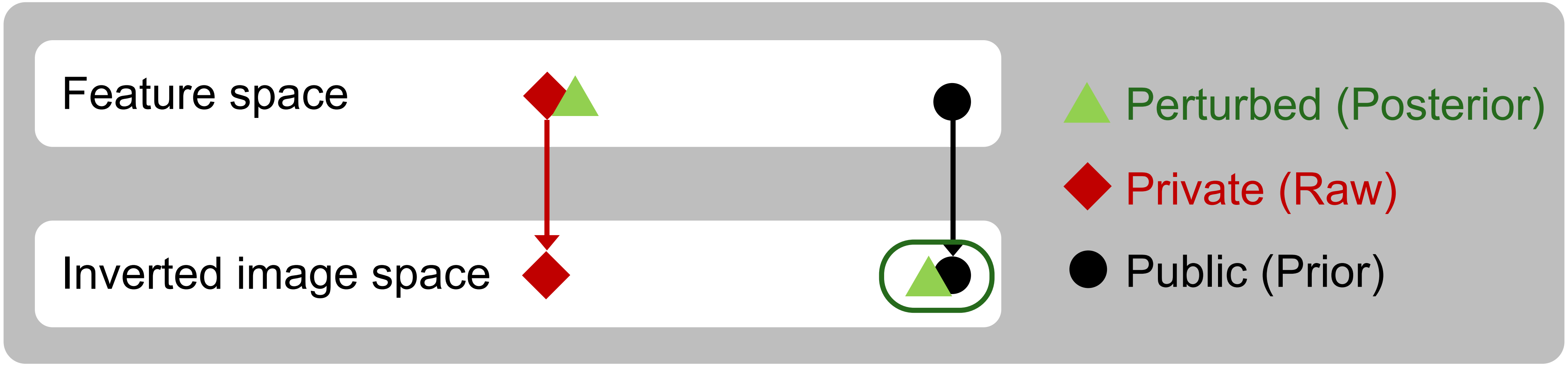
Defense Intuitions (Privacy+Utility)

“Get Close”

Find a feature perturbation that

1) is small;

2) draws inverted image close to public average faces.



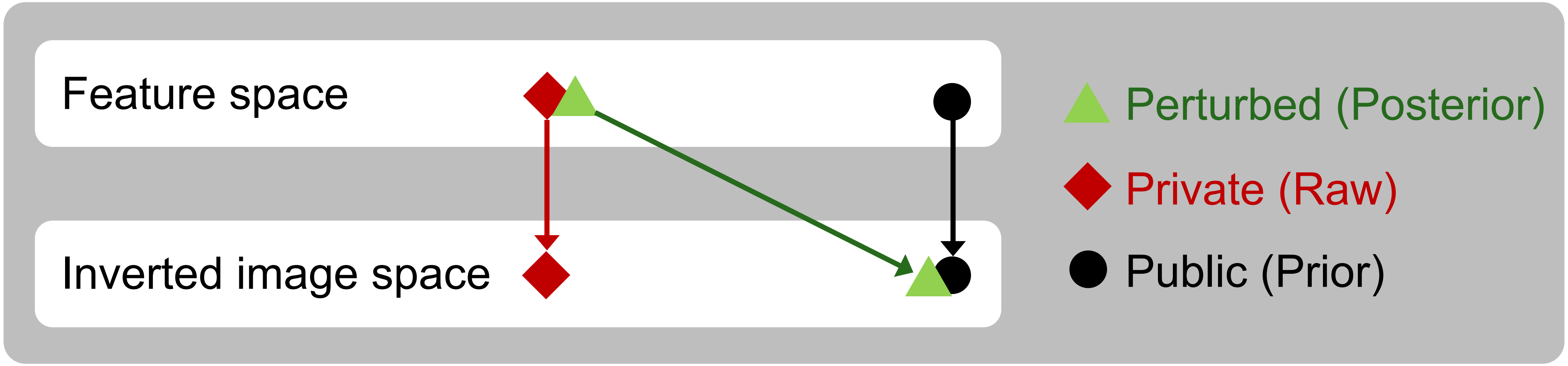
Defense Intuitions (Privacy+Utility)

“Get Close”

Find a feature perturbation that

1) is small;
2) draws inverted image close to public average faces.

The perturbed feature poisons adaptive attackers!



Crafter

A game view:

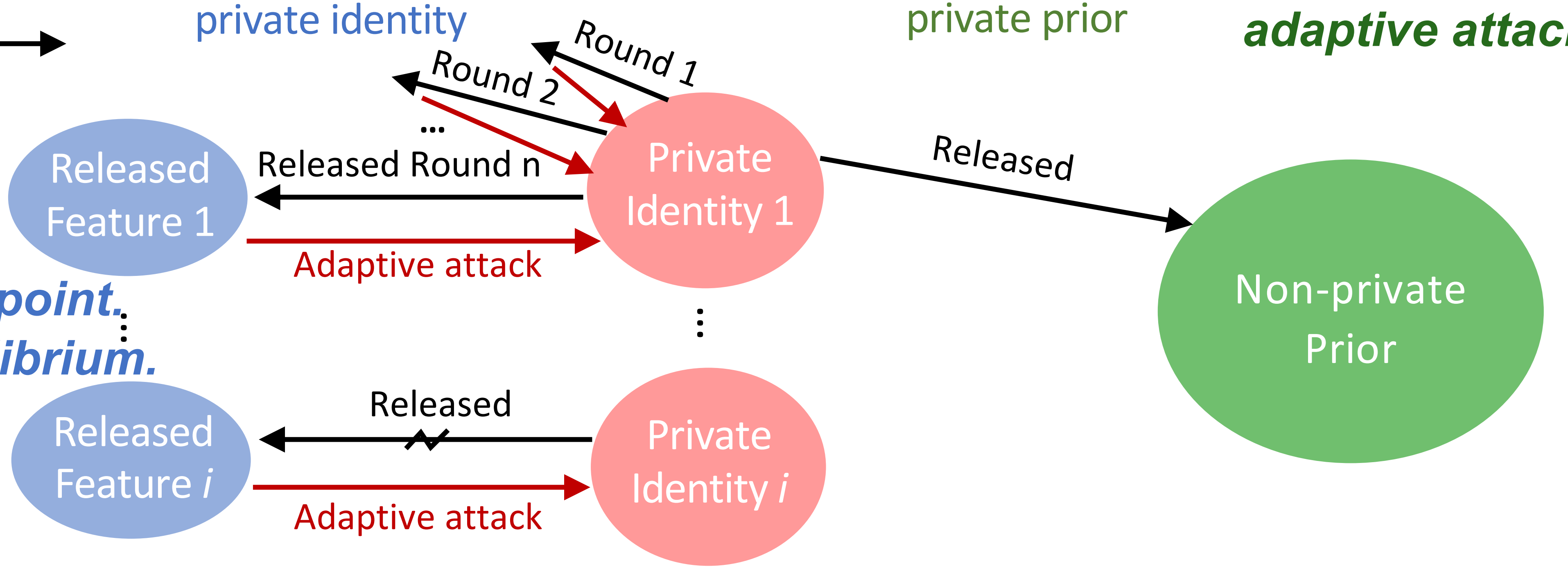
Attack →

Defense →

Conventional:
stay away from private identity

Our Crafter:
get close to non-private prior

The perturbed feature poisons adaptive attackers!



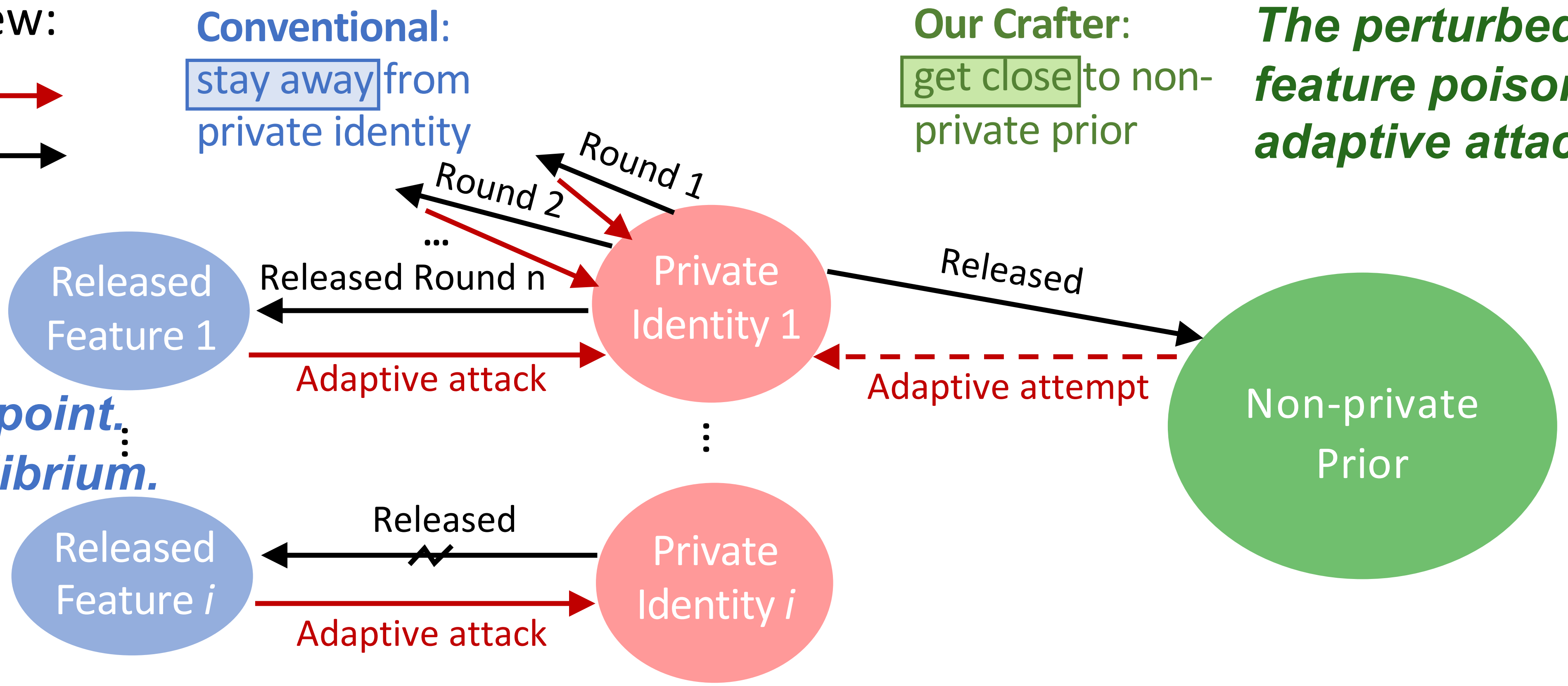
A stationary point.
But not equilibrium.

Crafter

A game view:

Attack

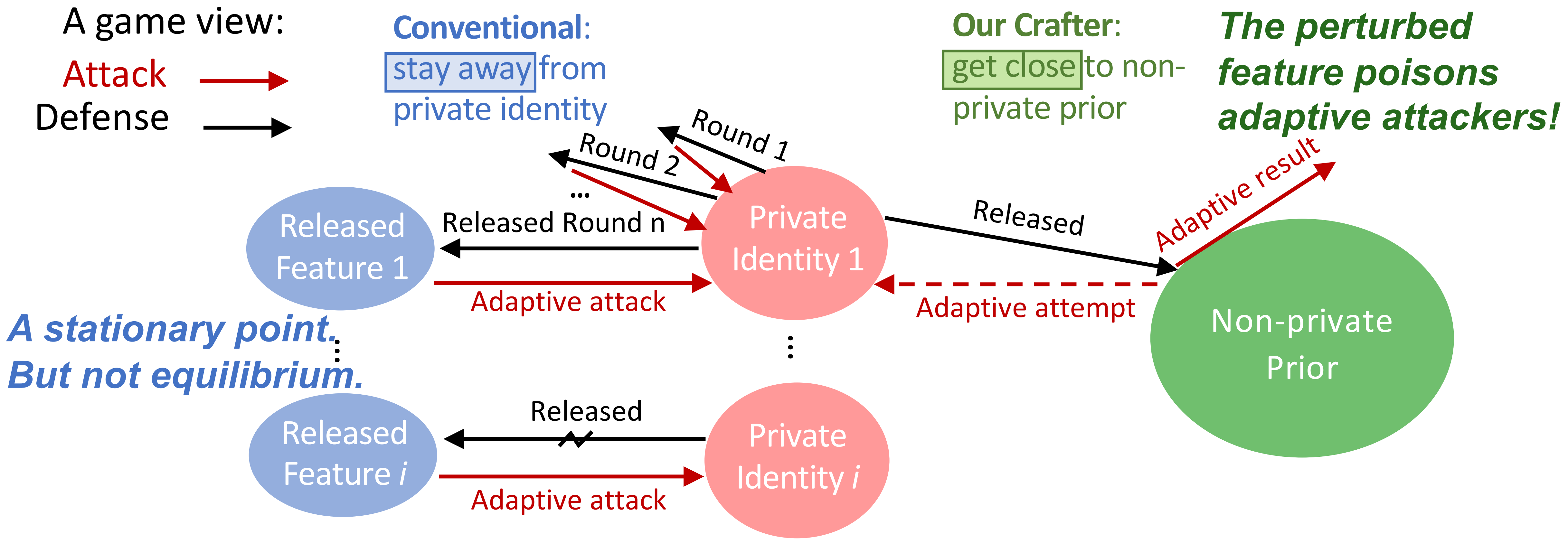
Defense



The perturbed feature poisons adaptive attackers!

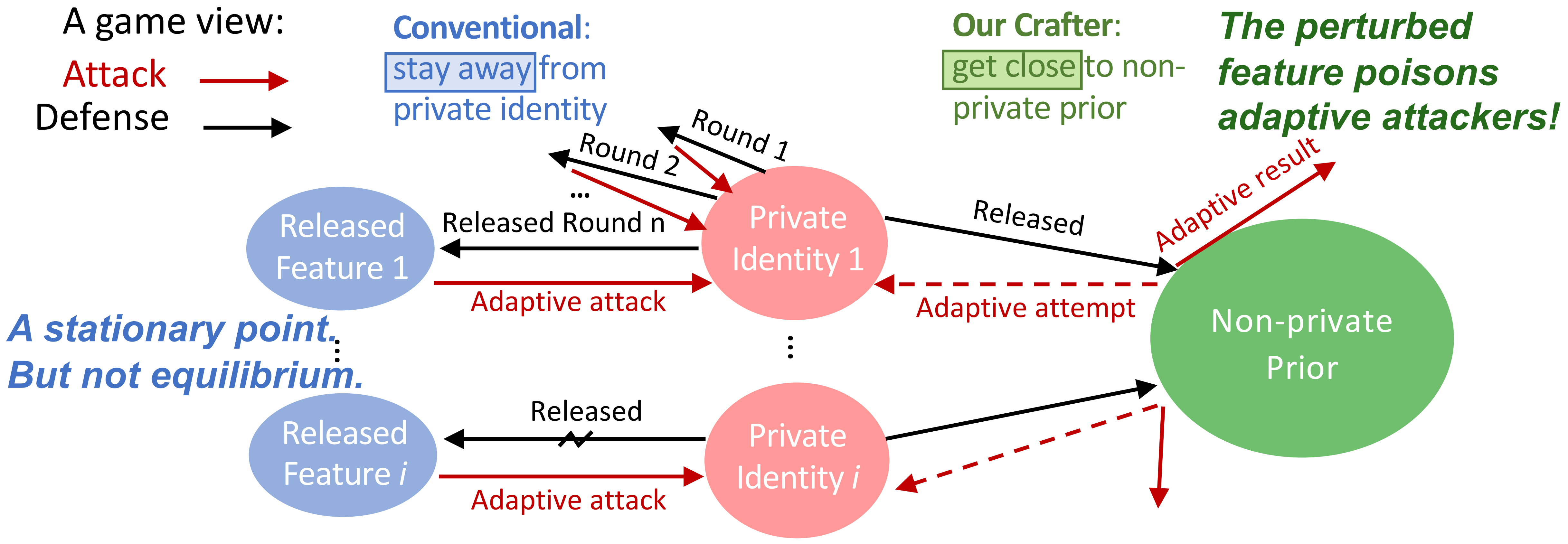
A stationary point. But not equilibrium.

Crafter



A game view:
Attack
Defense
Conventional:
stay away from private identity
Our Crafter:
get close to non-private prior
The perturbed feature poisons adaptive attackers!
Round 1
Round 2
...
Released Round n
Released Feature 1
Private Identity 1
Adaptive attack
Released
Adaptive result
Adaptive attempt
Non-private Prior
A stationary point.
But not equilibrium.
Released
Released Feature *i*
Private Identity *i*
Adaptive attack

Crafter



A game view:
Attack
Defense
Conventional:
stay away from private identity
Our Crafter:
get close to non-private prior
The perturbed feature poisons adaptive attackers!
Round 1
Round 2
…
Released Round n
Released Feature 1
Private Identity 1
Adaptive attack
Released
Adaptive result
Adaptive attempt
Non-private Prior
A stationary point.
But not equilibrium.
Released
Released Feature *i*
Private Identity *i*
Adaptive attack

Evaluation

Datasets

CelebA (64*64)

- 40 binary utility attributes

LFW (128*128)

- 10 binary utility attributes

VGGFace2 (112*112)

- 5-class hair color utility attribute

Baselines

AdvLearn

- Deployment scenario

Disco

- Deployment scenario
- Improves upon AdvLearn with a pruner

TIPRDC

- Development scenario

Xiao et al. "Adversarial learning of privacy-preserving and task-oriented representations ", 2020

Singh et al. "Disco: Dynamic and invariant sensitive channel obfuscation for deep neural networks ", 2021

Li et al. "Tiprdc: task-independent privacy-respecting data crowdsourcing framework for deep learning with anonymized intermediate representations ", 2020

Evaluation

Tradeoff parameter.

- **AdvLearn:** {0.1, 0.5, 0.8}
- **Disco:** {0.2, 0.6, 0.8}
- **TIPRDC:** {0.1, 0.5, 0.8}

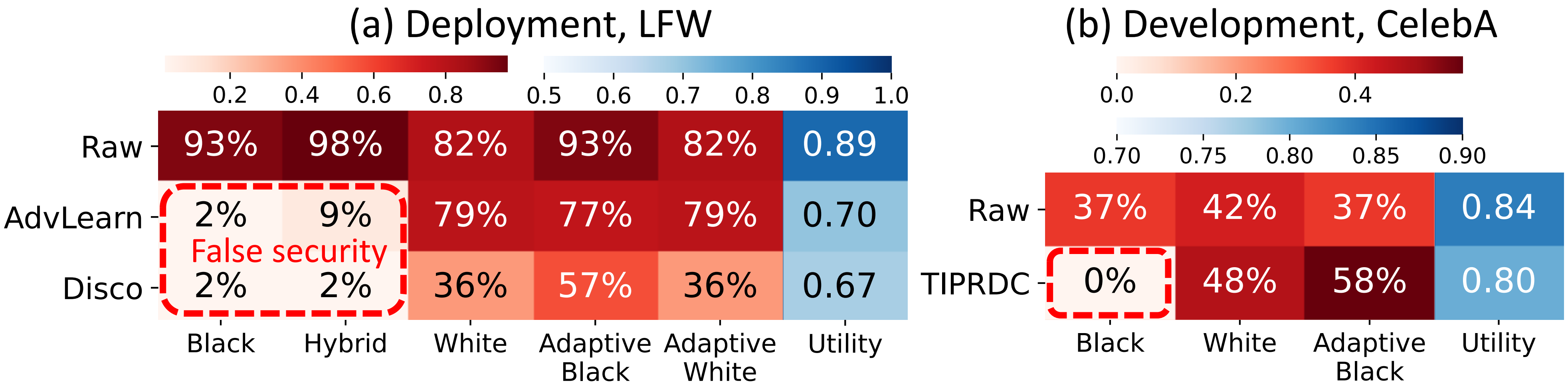
Privacy Metrics.

- **Eval Acc:** identification accuracy of the inverted images.
- **Feature Similarity:** cosine similarity between of the raw & inverted images.
- **SSIM:** pixel-level resemblance between the raw & inverted images.
- **Human study:** 35 human feedbacks, Macro-F1 score of reidentification.

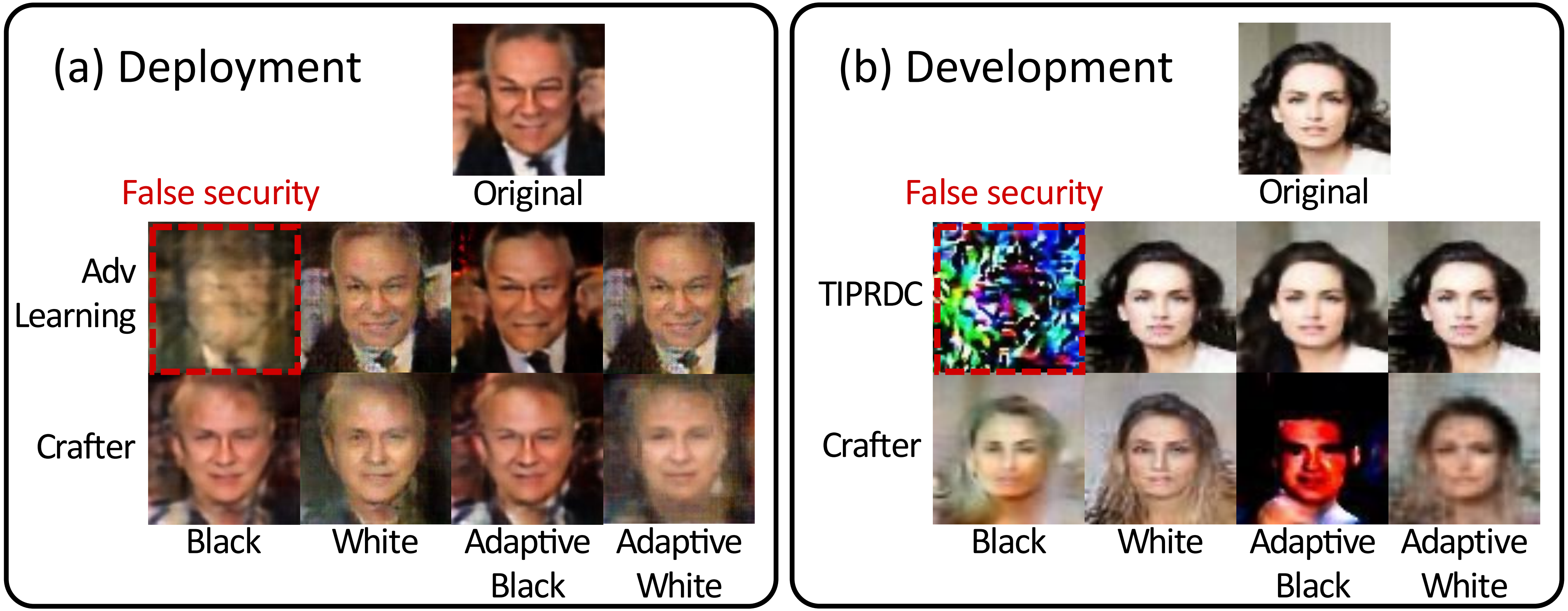
Evaluation

Baselines: vulnerable against adaptive attacks → false security.

Crafter:
- robust against both back- & white-box inversion,
- robust against adaptive inversions
- maintains high utility performance.



Evaluation



(a) Deployment
Original
False security
Adv Learning
Crafter
Black
White
Adaptive Black
Adaptive White
(b) Development
Original
False security
TIPRDC
Crafter
Black
White
Adaptive Black
Adaptive White

Crafter: Facial Feature Crafting against Inversion-based Identity Theft on Deep Models

Shiming Wang, Zhe Ji, Liyao Xiang, Hao Zhang,
Xinbing Wang, Chenghu Zhou, Bo Li

Code Available @GitHub 👉

NDSS SYMPOSIUM/2024

Presented by Internet Society