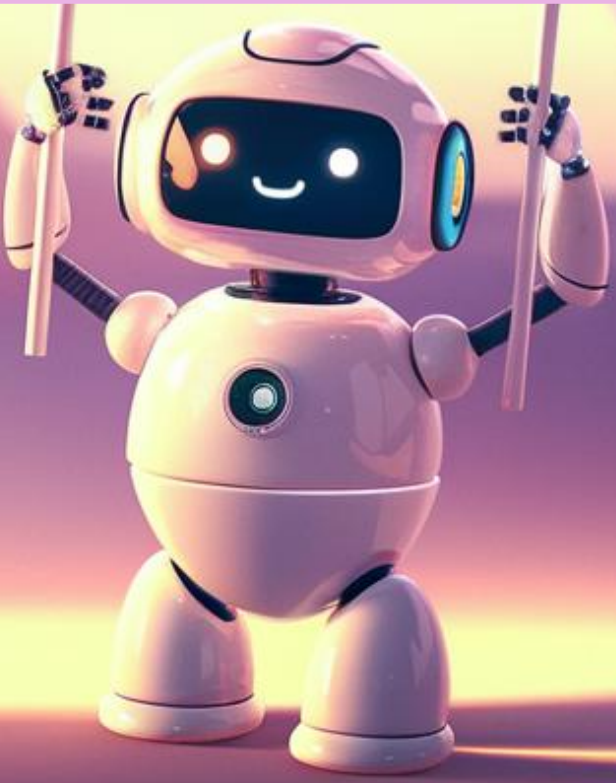
PhD Defense:
Understanding and Mitigating Security, Privacy, and Ethical Risks in Generative Artificial Intelligence



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Lukas Struppek

‘Insane’: OpenAI Introduces Gpt-4O Native Image Generation and It’s Already Wowing Users
(VentureBeat)

Google’s Gemini 2.5 Pro is Better at Coding, Math & Science Than Your Favourite AI Model
(TechRepublic)

The $3.8 Trillion Opportunity: Unlocking the Economic Potential Of the US Generative AI Ecosystem
(Microsoft)

How Deepseek’s R1 Model Is Disrupting The AI Landscape
(CTech)

Multimodal Generative AI For Medical Image Interpretation
(Nature)

Template: presentationmagazine.com

Does Greater AI Capability Result in Greater Reliability?

'Insane': OpenAI Introduces Gpt-4O Native Image Generation and It's Already Wowing Users
(VentureBeat)

Google's Gemini 2.5 Pro is Better at Coding, Math & Science Than Your Favourite AI Model
(TechRepublic)

How Deepseek's R1 Model Is Disrupting The AI Landscape
(CTech)

Multimodal Generative AI For Medical Image Interpretation
(Nature)

Template: presentationmagazine.com

Agenda

1

Trustworthy Machine Learning

- Client-Side Scanning

2

GenAI as an Adversarial Tool

- Model Inversion Attacks

3

Trustworthy Text-to-Image Synthesis

- Memorization
- Character Biases
- Backdoor Attacks

Disclaimer: This presentation includes (blurred) images that may be perceived as offensive

Trustworthy Machine Learning

Dimensions of Trustworthy Machine Learning

Privacy

Protecting Sensitive Data Integrity

Security

Safeguard Against Adversarial Threats

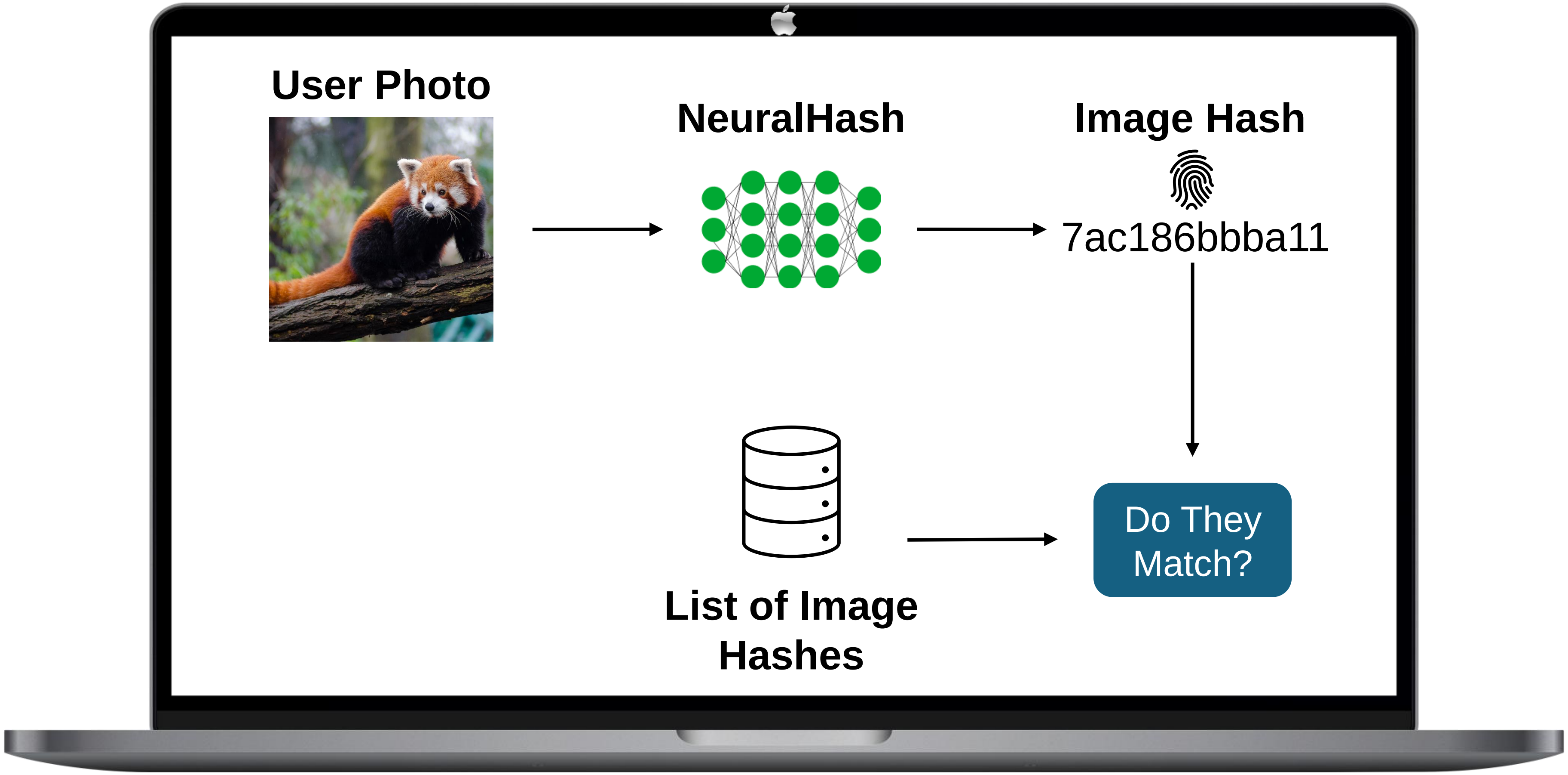
Safety & Robustness

Ensuring Reliable Operations

Fairness

Preventing and Mitigating Biases

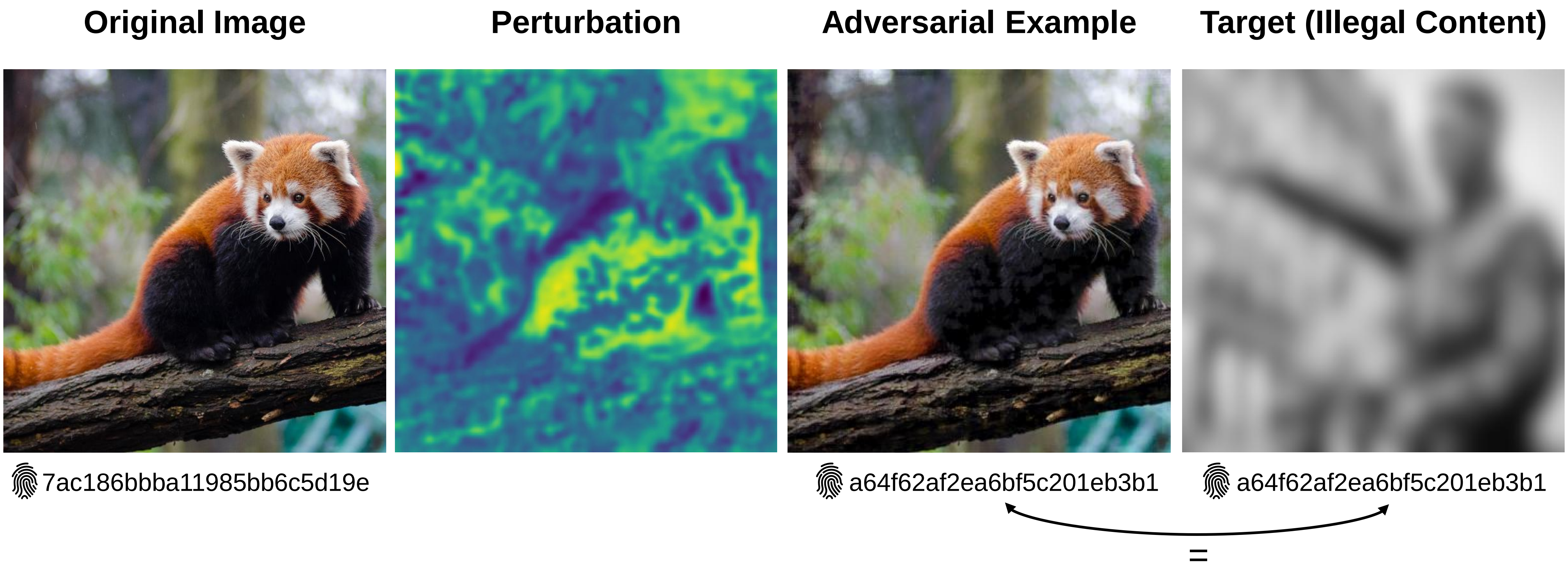
Client-Side Scanning With Deep Perceptual Hashing



Can We Trust Neural Networks Used for Perceptual Hashing?

[Apple Inc. *CSAM Detection – Technical Summary.* 2021]

Forcing False-Positive Detections ...



[**Struppek***, Hintersdorf*, Neider, Kersting. ***Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash***. FAccT 2022]
[Hintersdorf*, **Struppek***, Neider, Kersting. ***Investigating the Risks of Client-Side Scanning for the Use Case NeuralHash***. IEEE S&P ConPro Workshop 2022. **Best Paper Award** 🏆]

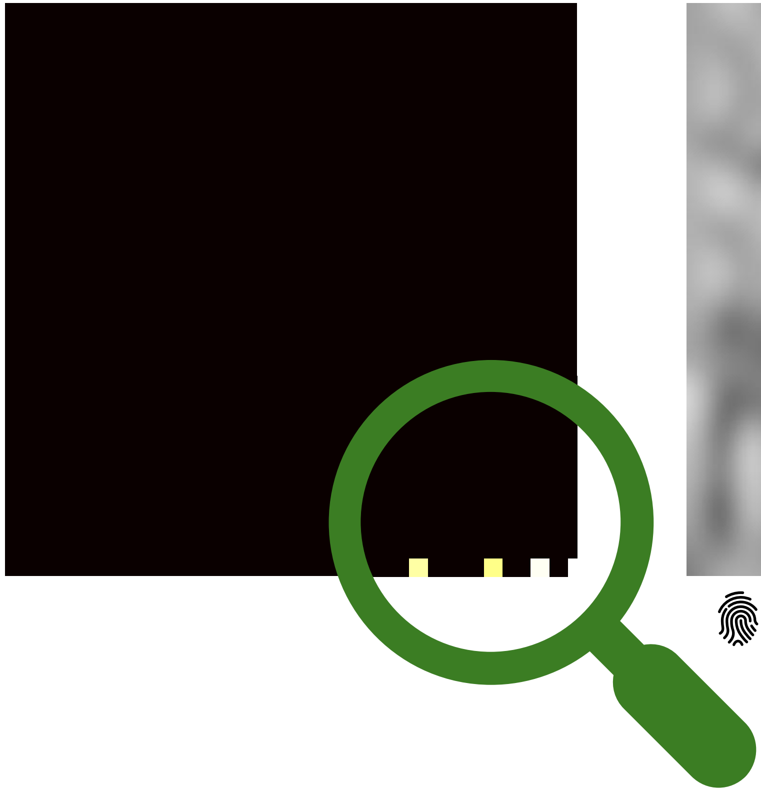
… Or Evading Detection By Small Changes

Illegal Content

a64f62af2ea6bf5c**2**01eb3b1

↑

Perturbation



Adversarial Example

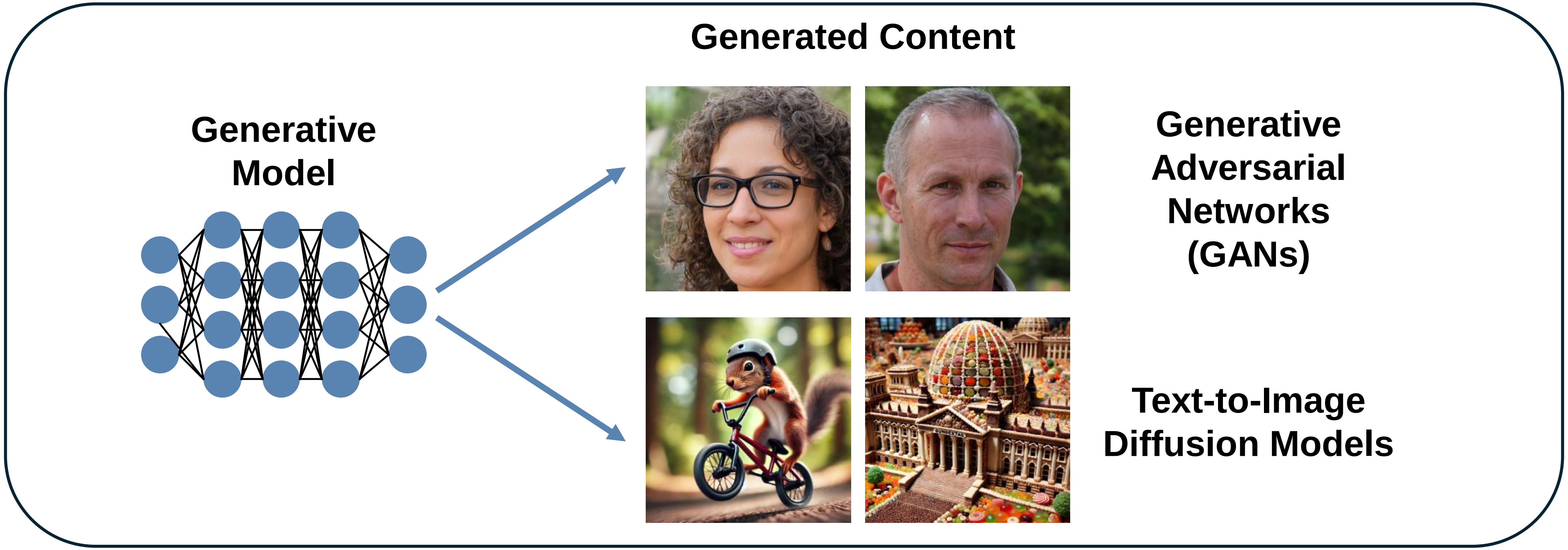
a64f62af2ea6bf5c**3**01eb3b1

↑

[**Struppek***, Hintersdorf*, Neider, Kersting. ***Learning to Break Deep Perceptual Hashing: The Use Case NeuralHash***. FAccT 2022]

[Hintersdorf*, **Struppek***, Neider, Kersting. ***Investigating the Risks of Client-Side Scanning for the Use Case NeuralHash***. IEEE S&P ConPro Workshop 2022. **Best Paper Award** 🏆]

Trustworthy Image Generation



Generated Content

Generative Model

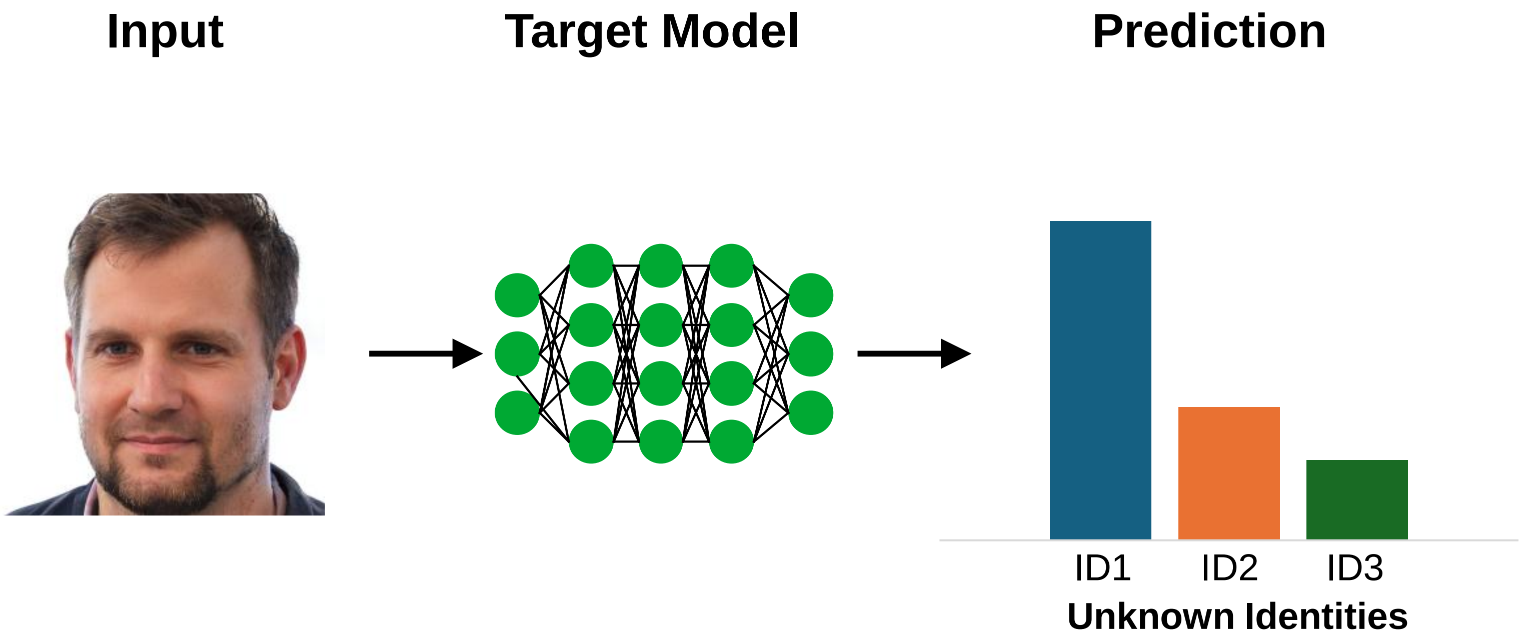
Generative Adversarial Networks (GANs)

Text-to-Image Diffusion Models

GenAI as an Adversarial Tool ⟷ Trustworthy GenAI Systems

Generative AI as an Adversarial Tool

Face Recognition – A Privacy-Sensitive Task



Input

Target Model

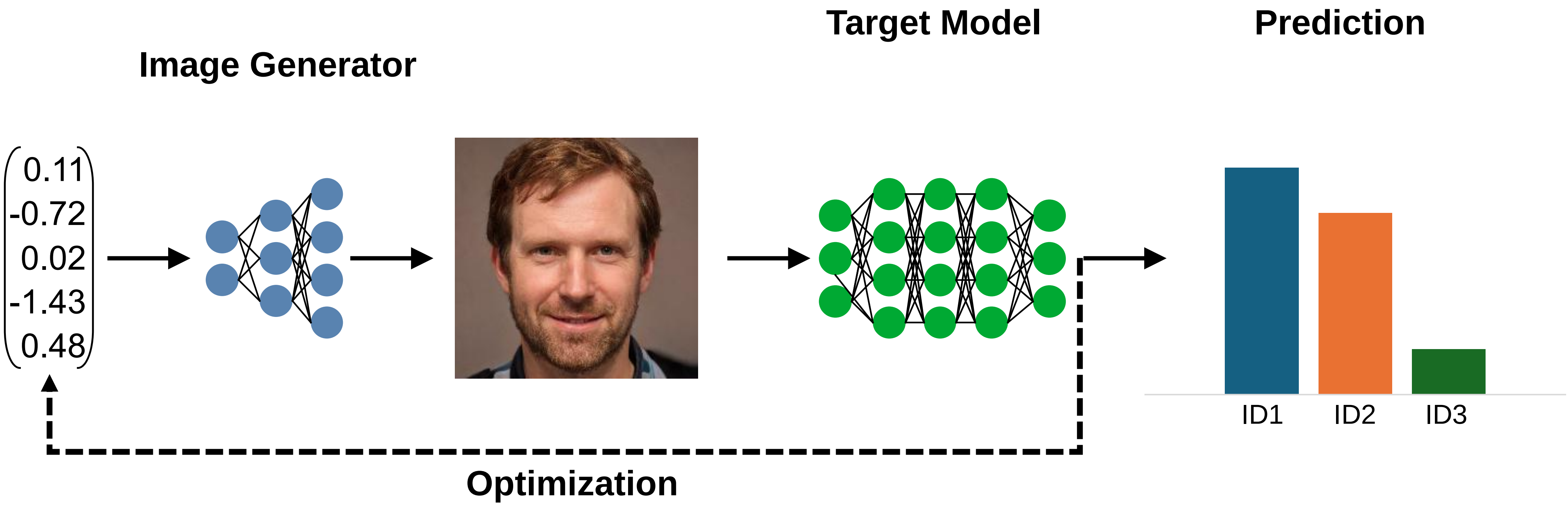
Prediction

Can We Reconstruct the Appearance of Individuals From the Training Data?

[Fredrikson et al. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*. CCS 2015]

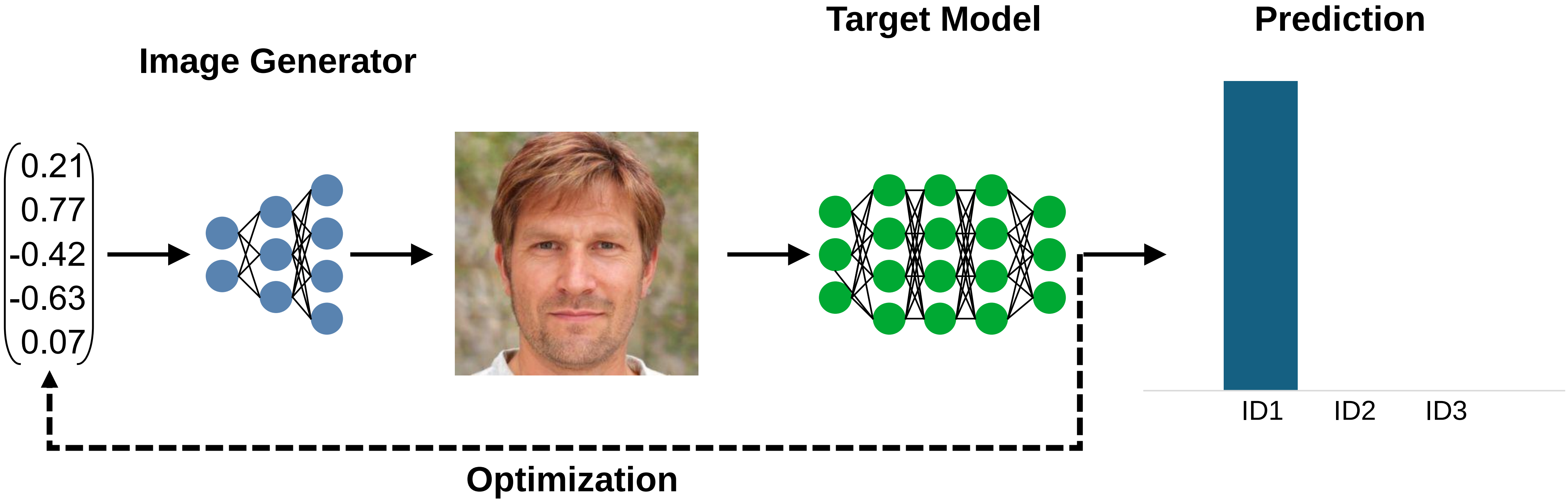
[Zhang et al. *The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks*. CVPR 2020]

Reconstructing Sensitive Features from Trained Models



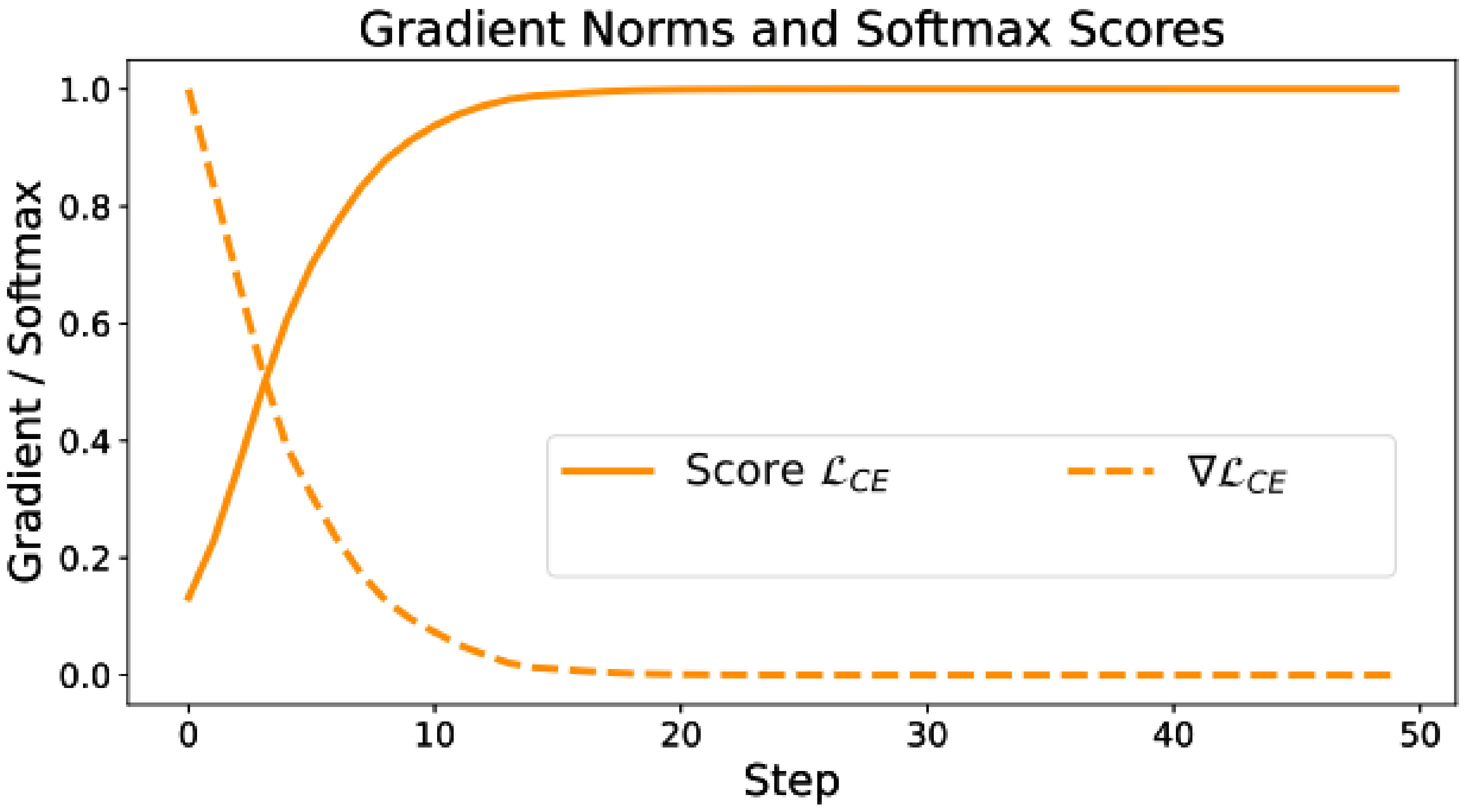
[**Struppek**, Hintersdorf, De Almeida Correia, Adler, Kersting. ***Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks***. ICML 2022]

Reconstructing Sensitive Features from Trained Models

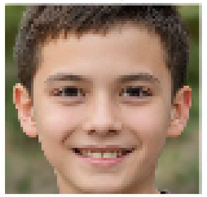


[**Struppek**, Hintersdorf, De Almeida Correia, Adler, Kersting. ***Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks***. ICML 2022]

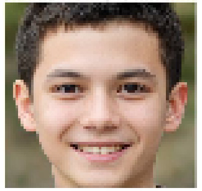
Overcoming Vanishing Gradients



Initial



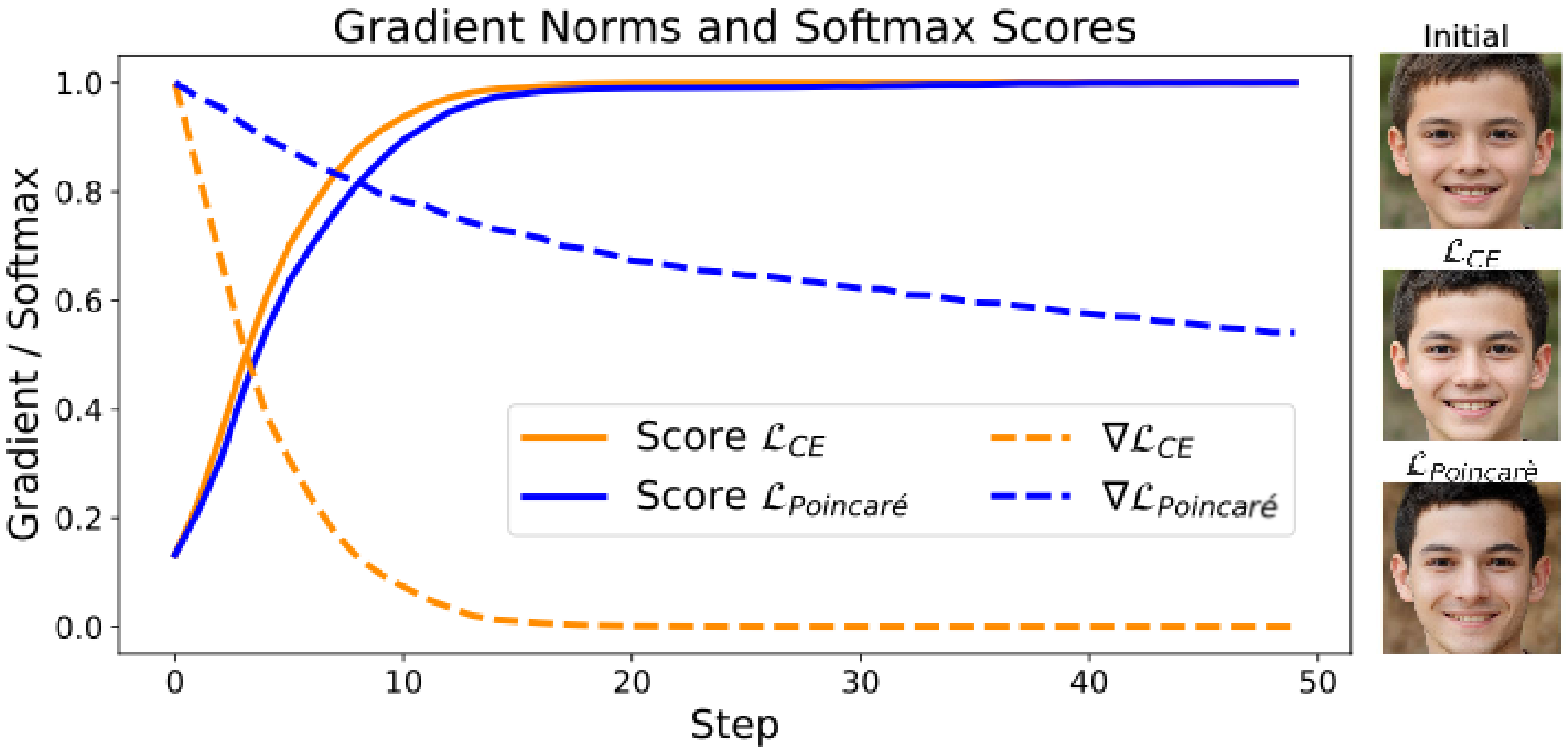
$\mathcal{L}_{CE}$



$$\boxed{\mathcal{L}_{CE} = -\sum_i y_i \log p_i}$$

[**Struppek**, Hintersdorf, De Almeida Correia, Adler, Kersting. ***Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks***. ICML 2022]

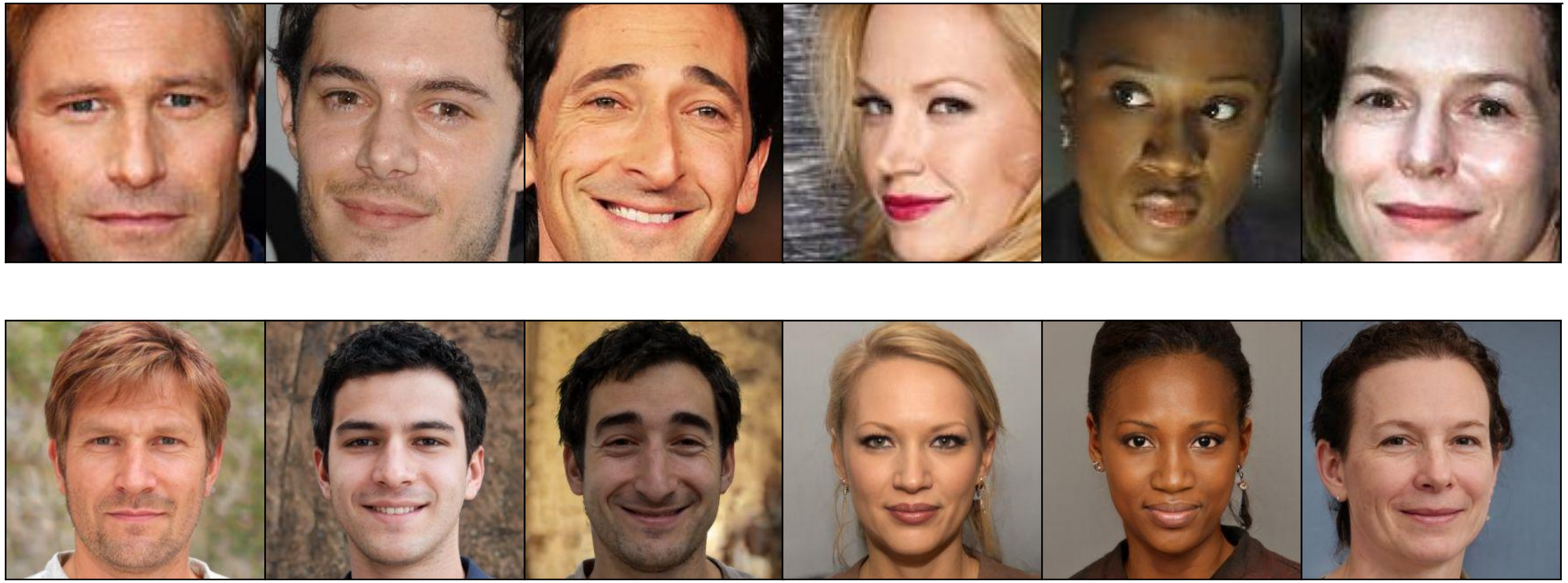
Overcoming Vanishing Gradients



$$\mathcal{L}_{Poincaré} = \operatorname{arcosh}\left(1 + \frac{2\|u-v\|_2^2}{(1-\|u\|_2^2)(1-\|v\|_2^2)}\right)$$

[**Struppek**, Hintersdorf, De Almeida Correia, Adler, Kersting. ***Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks***. ICML 2022]

The First High-Resolution Model Inversion Attack



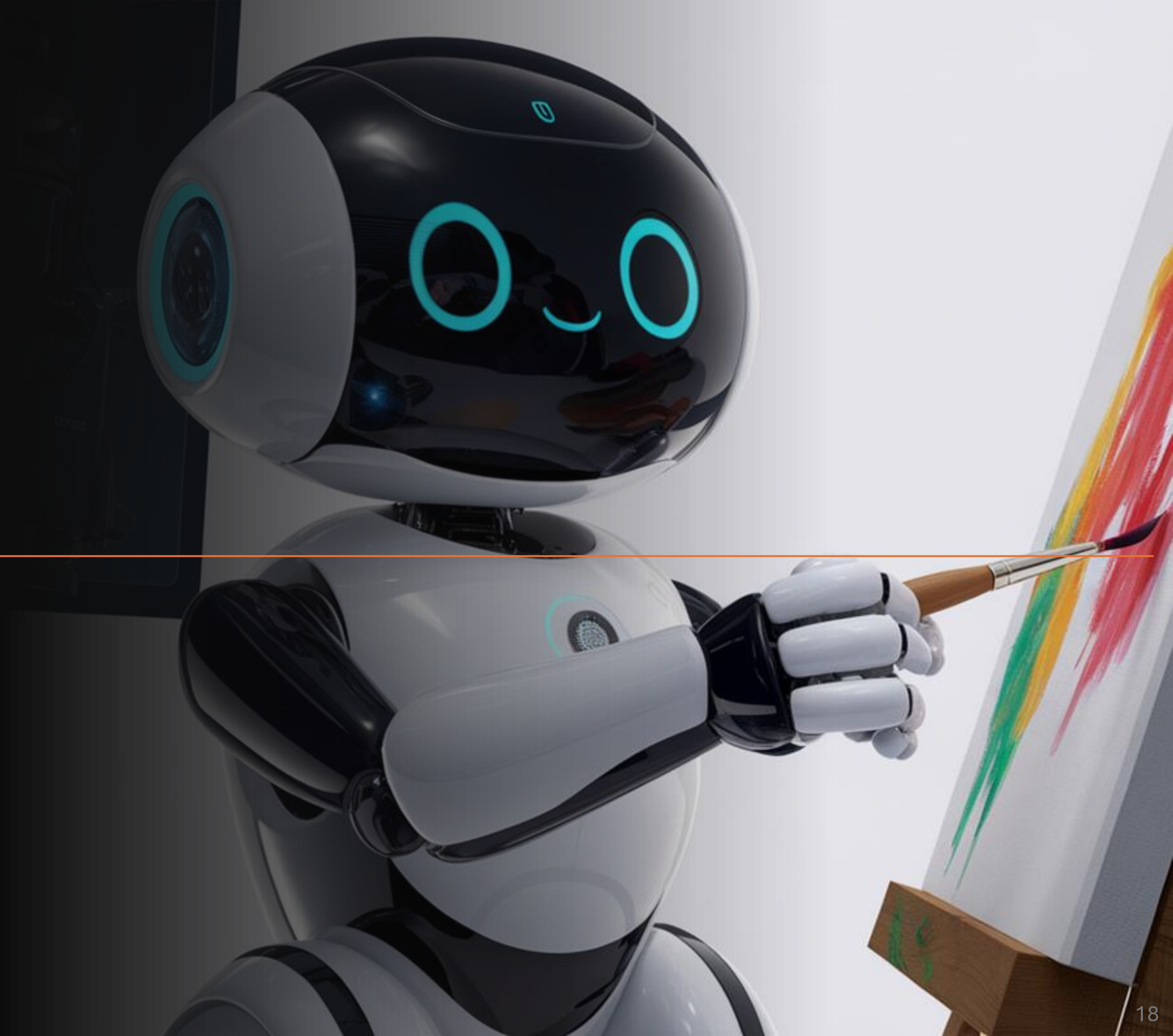
Training Data
(224x224)

Attack Results
(1024x1024)

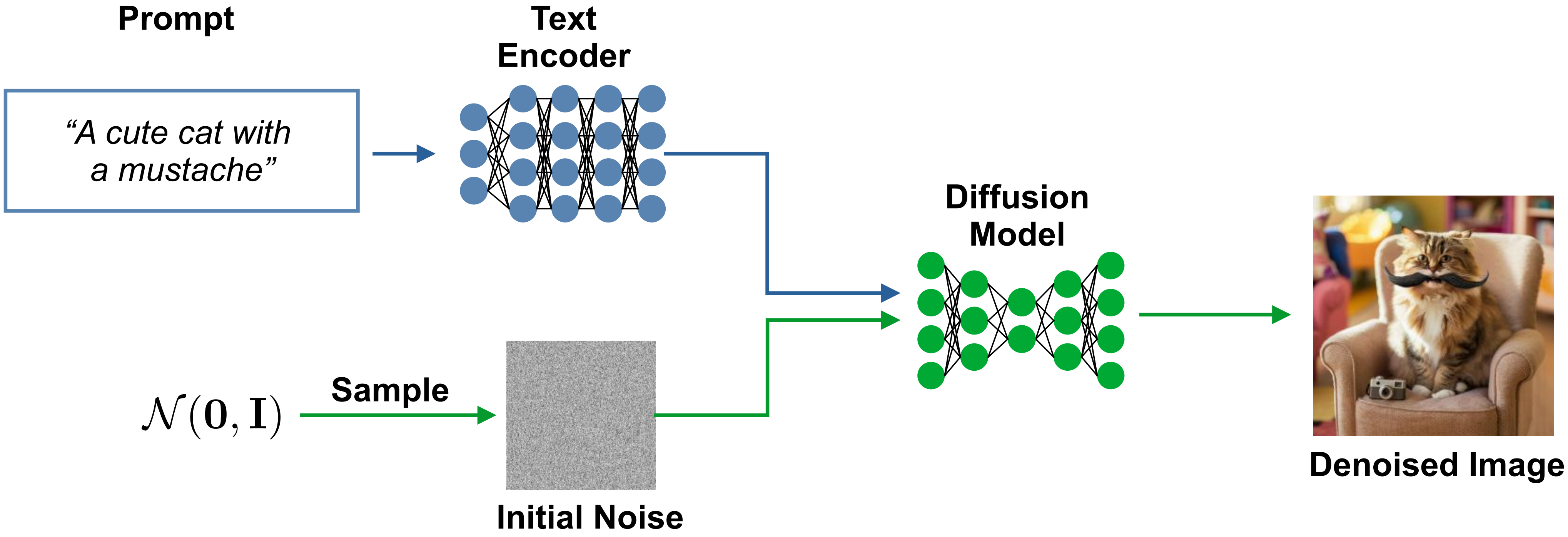
How Can We Mitigate This Form of Privacy Leakage?

[**Struppek**, Hintersdorf, De Almeida Correia, Adler, Kersting. ***Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks***. ICML 2022]

Trustworthy Text-to-Image Synthesis



Text-to-Image Synthesis With Diffusion Models



Prompt

“A cute cat with a mustache”

Text Encoder

$\mathcal{N}(\mathbf{0}, \mathbf{I})$

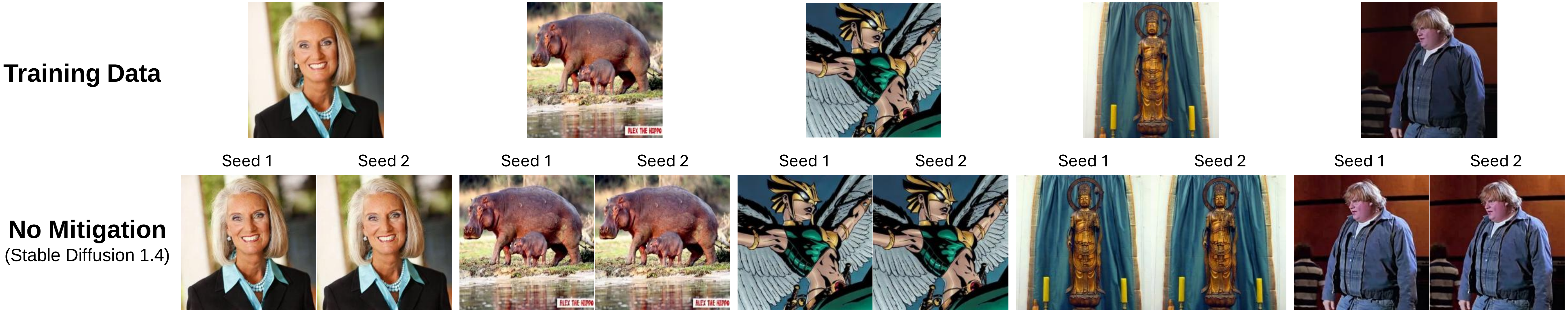
Sample

Initial Noise

Diffusion Model

Denoised Image

Undesired Data Replication in Diffusion Models



Training Data

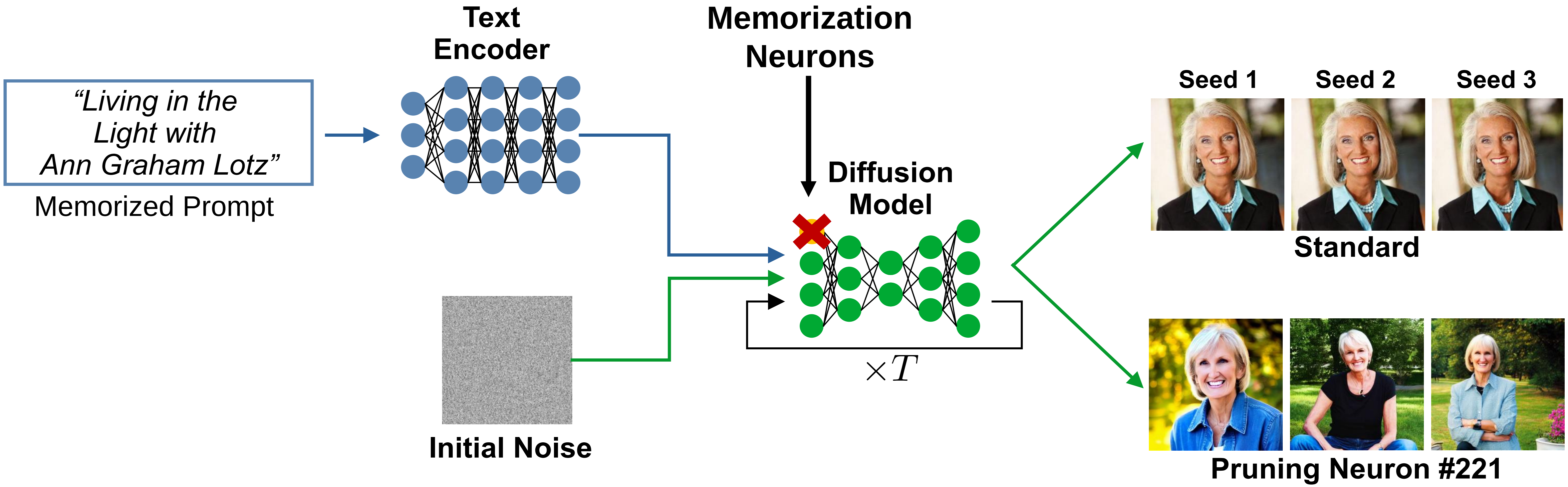
Seed 1 Seed 2 Seed 1 Seed 2 Seed 1 Seed 2 Seed 1 Seed 2 Seed 1 Seed 2

No Mitigation
(Stable Diffusion 1.4)

Can We Localize Memorization in Diffusion Models?

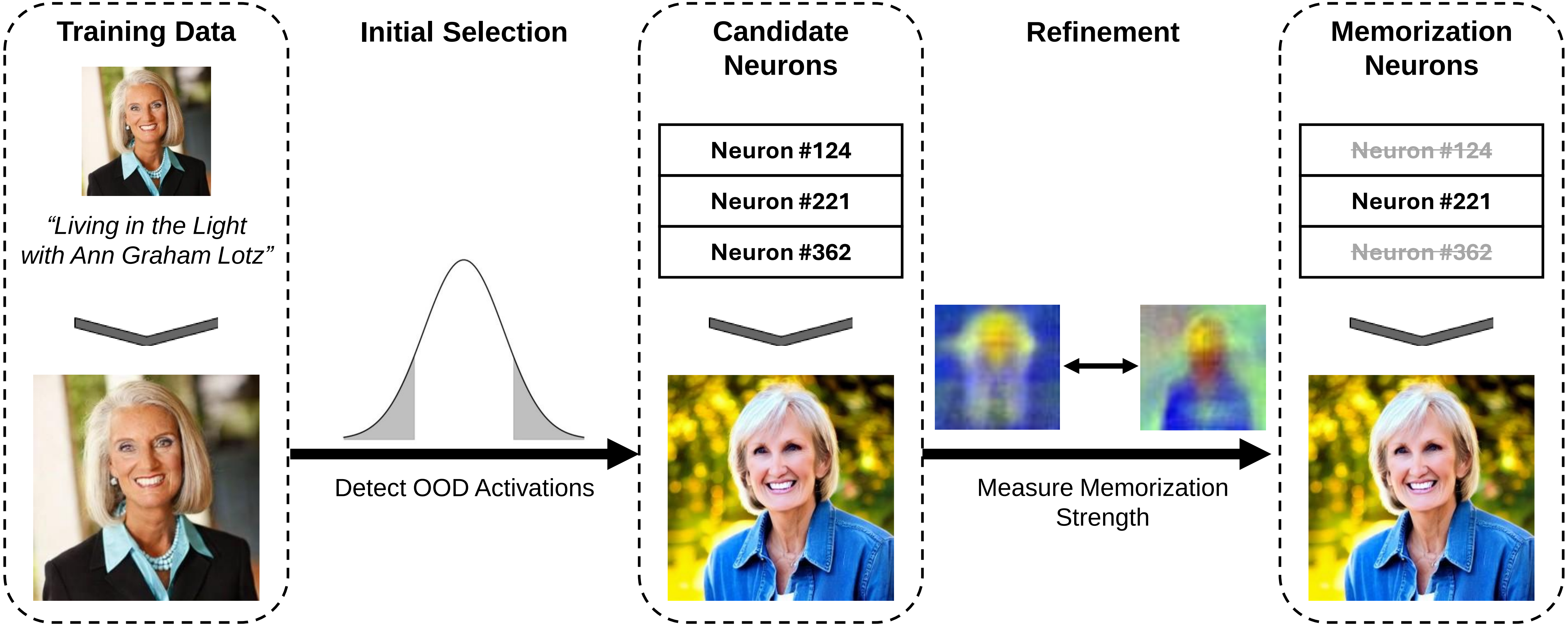
[Carlini et al. *Extracting Training Data from Diffusion Models*. Usenix 2023]
[Somepalli et al. *Diffusion Art or Digital Forgery? Investigating Data Replication in Diffusion Models*. CVPR 2023]

NeMo – Localizing **Ne**uron **M**em**o**rization



[Hintersdorf*, **Struppek***, Kersting, Dziedzic, Boenisch. ***Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models***. NeurIPS 2024]

NeMo 🐠 – Localizing **Ne**uron **M**em**o**rization



Training Data

“Living in the Light with Ann Graham Lotz”

Initial Selection

Detect OOD Activations

Candidate Neurons

Neuron #124

Neuron #221

Neuron #362

Refinement

Measure Memorization Strength

Memorization Neurons

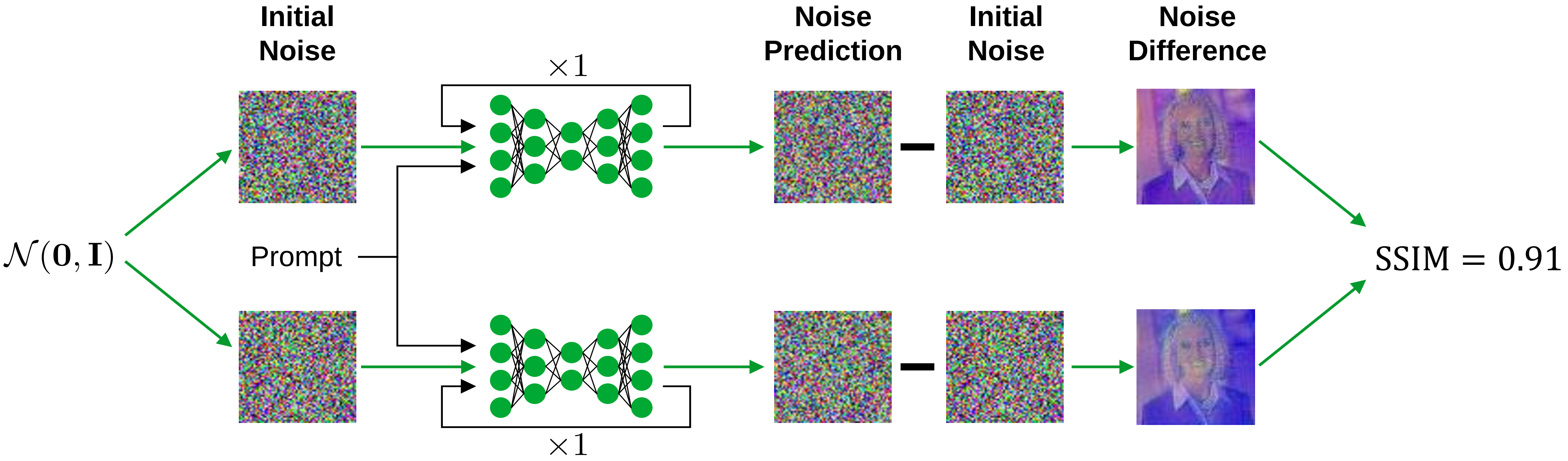
~~**Neuron #124**~~

Neuron #221

~~**Neuron #362**~~

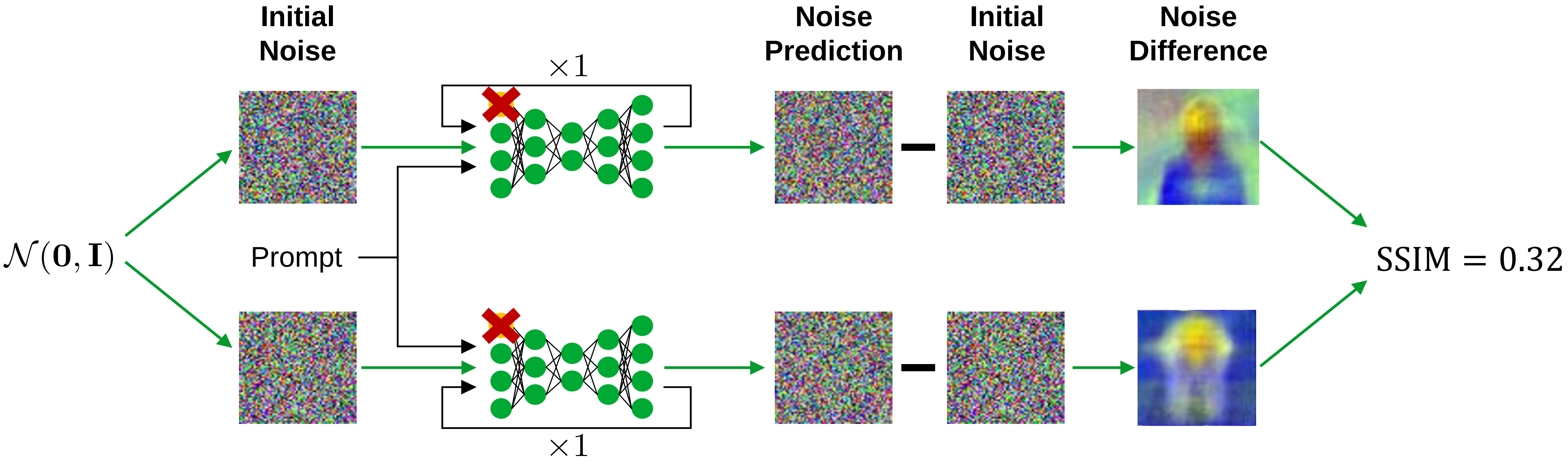
[Hintersdorf*, **Struppek***, Kersting, Dziedzic, Boenisch. ***Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models***. NeurIPS 2024]

Quantifying the Memorization Strength



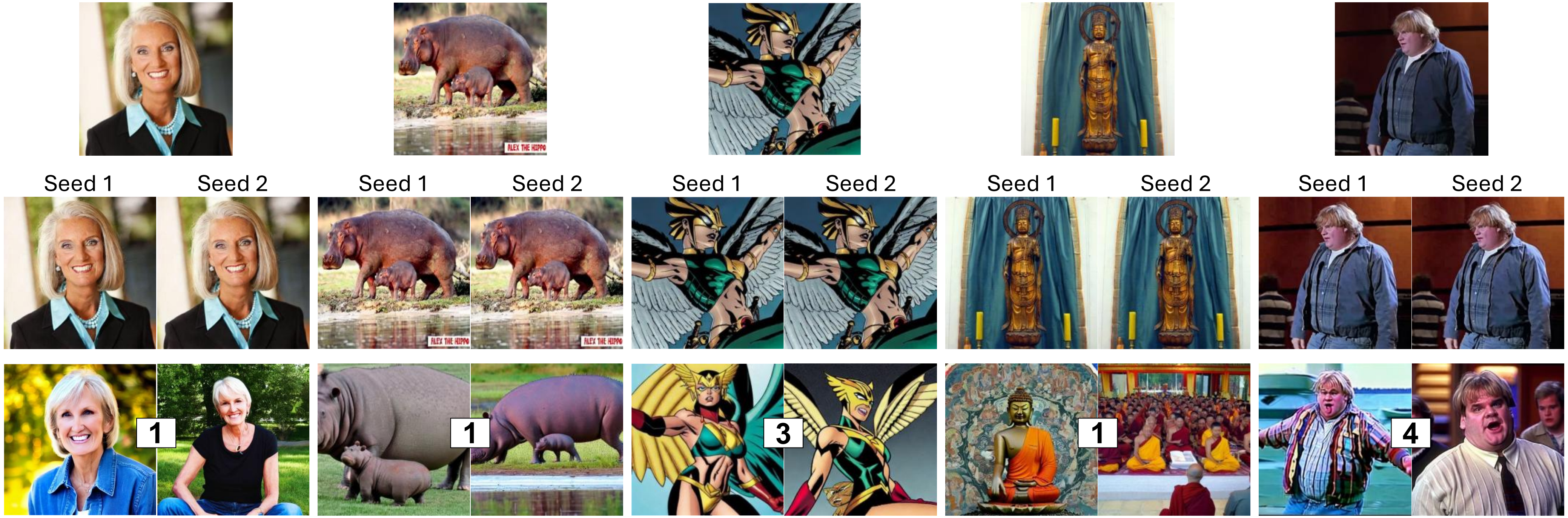
[Hintersdorf*, **Struppek***, Kersting, Dziedzic, Boenisch. ***Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models***. NeurIPS 2024]

Quantifying the Memorization Strength



[Hintersdorf*, **Struppek***, Kersting, Dziedzic, Boenisch. ***Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models***. NeurIPS 2024]

Pruning Memorization Neurons Mitigates Data Replication



Training Data

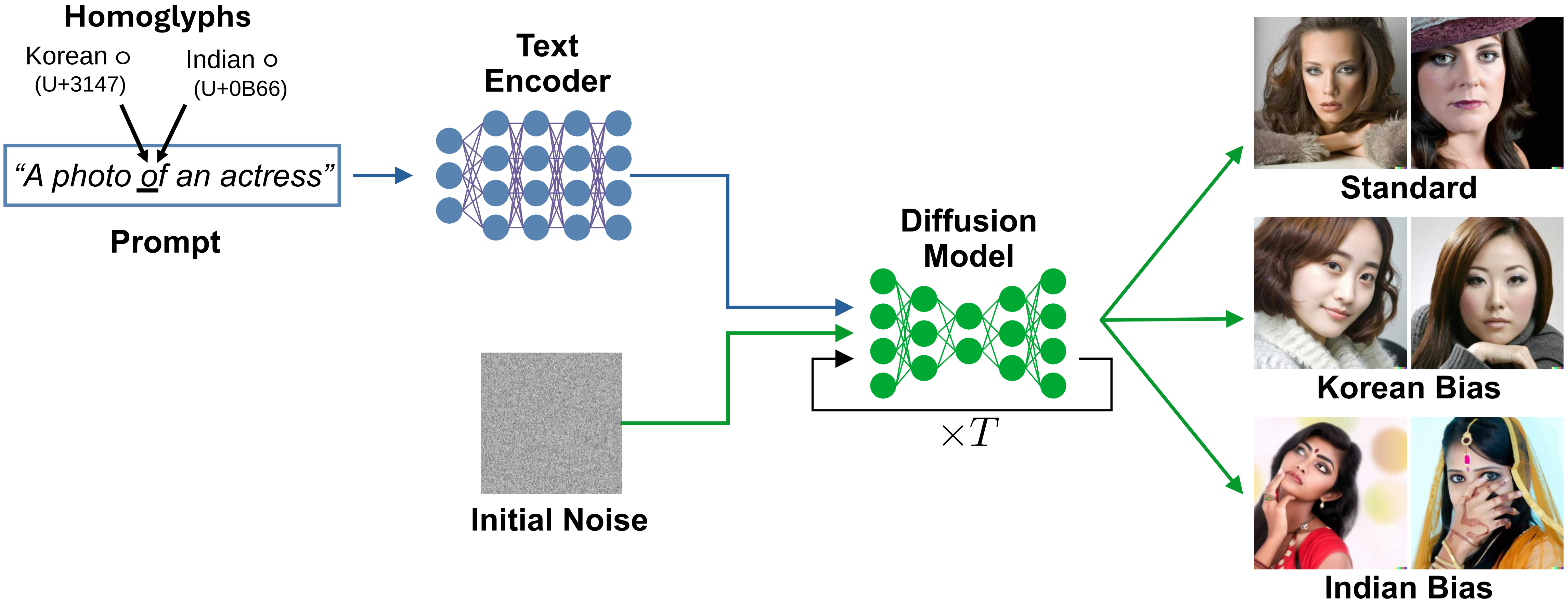
No Mitigation
(Stable Diffusion 1.4)

🐠 **Pruned** 🐠
Memorization Neurons

n : Number of memorization neurons

[Hintersdorf*, **Struppek***, Kersting, Dziedzic, Boenisch. ***Finding NeMo: Localizing Neurons Responsible For Memorization in Diffusion Models***. NeurIPS 2024]

Hidden Biases in Text-to-Image Synthesis Systems



Homoglyphs

Korean ㅇ (U+3147)

Indian ୦ (U+0B66)

“A photo of an actress”

Prompt

Text Encoder

Diffusion Model

$\times T$

Initial Noise

Standard

Korean Bias

Indian Bias

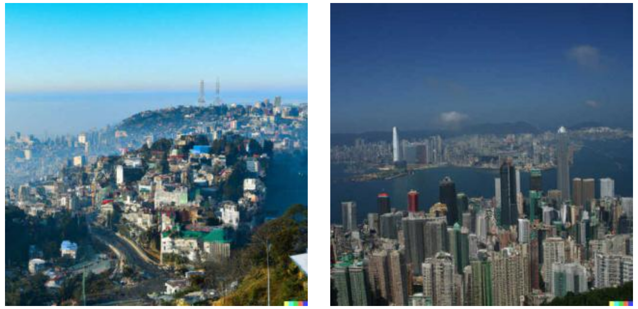
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]

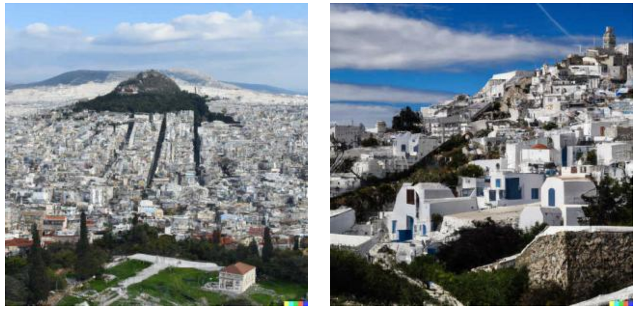
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

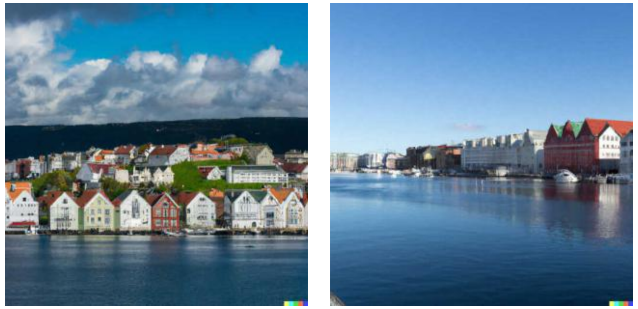
One Character to Bias Them All

Prompt: *“A city in bright sunshine”*

DALL-E 2

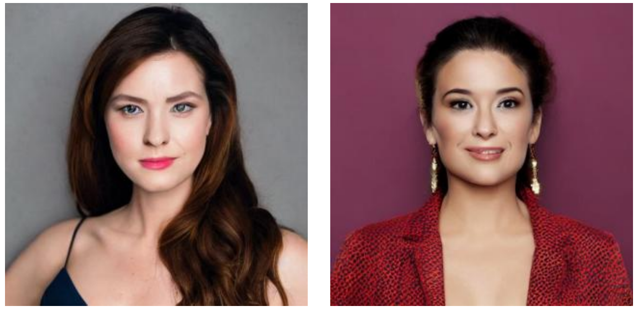

Latin A (U+0041)

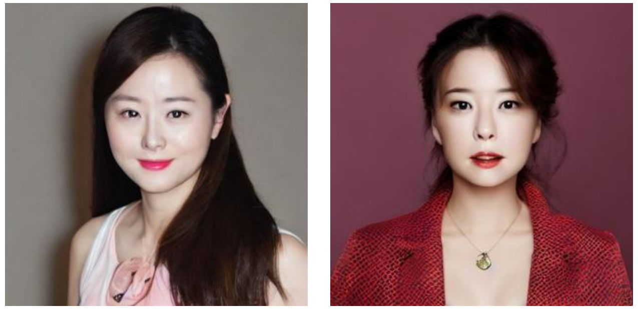

Greek Α (U+0391)

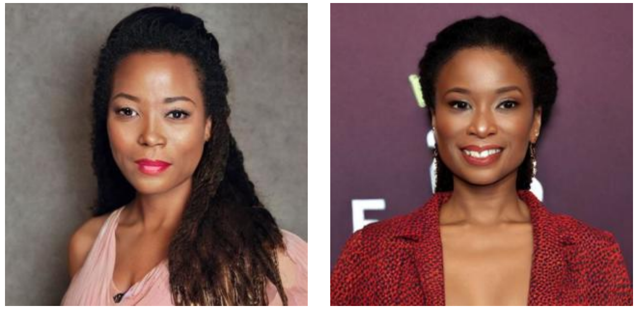

Scandinavian Å (U+00C5)

Prompt: *“A high-quality photo of an actress”*

Stable Diffsuion v1.5


Latin o (U+006F)

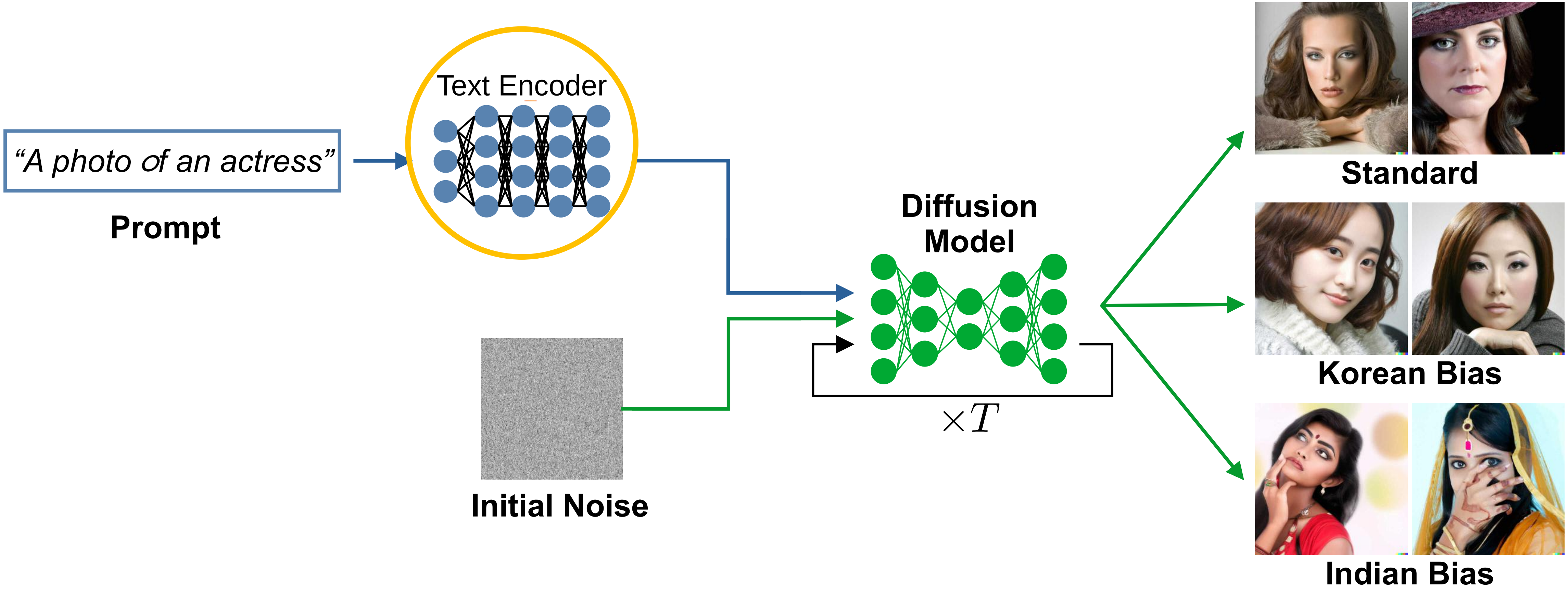

Korean ㅇ (U+3147)


African ọ (U+1ECD)

How to Make Systems Robust to Character Manipulations?

[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

Where Does This Behavior Originate From?



“A photo of an actress”

Prompt

Text Encoder

Diffusion Model

$\times T$

Initial Noise

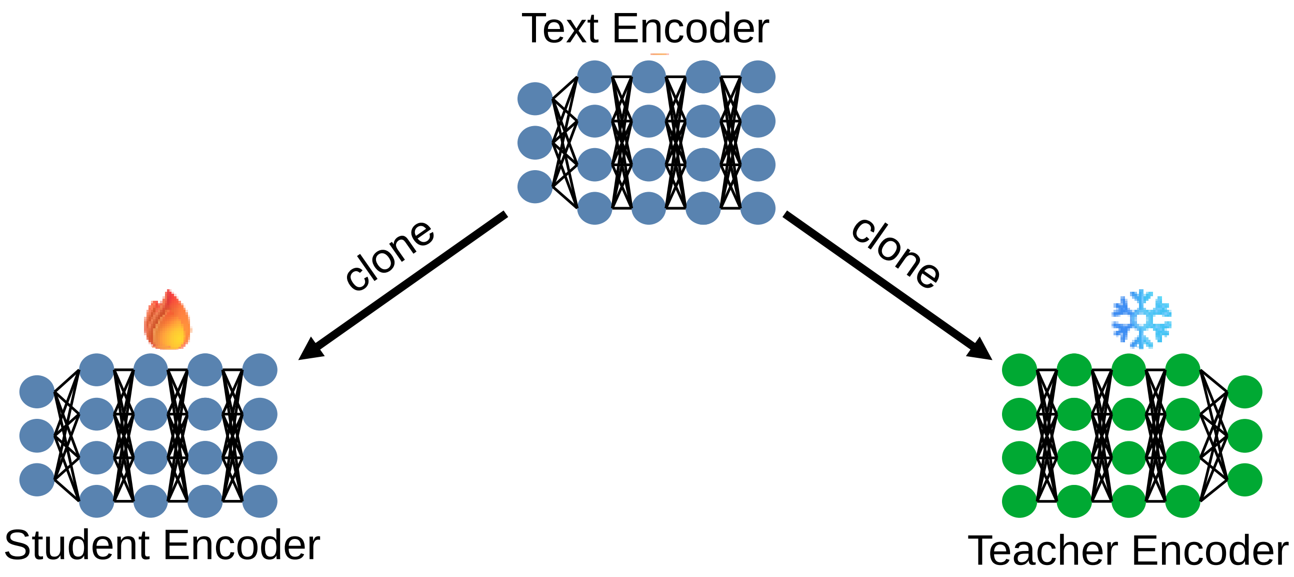
Standard

Korean Bias

Indian Bias

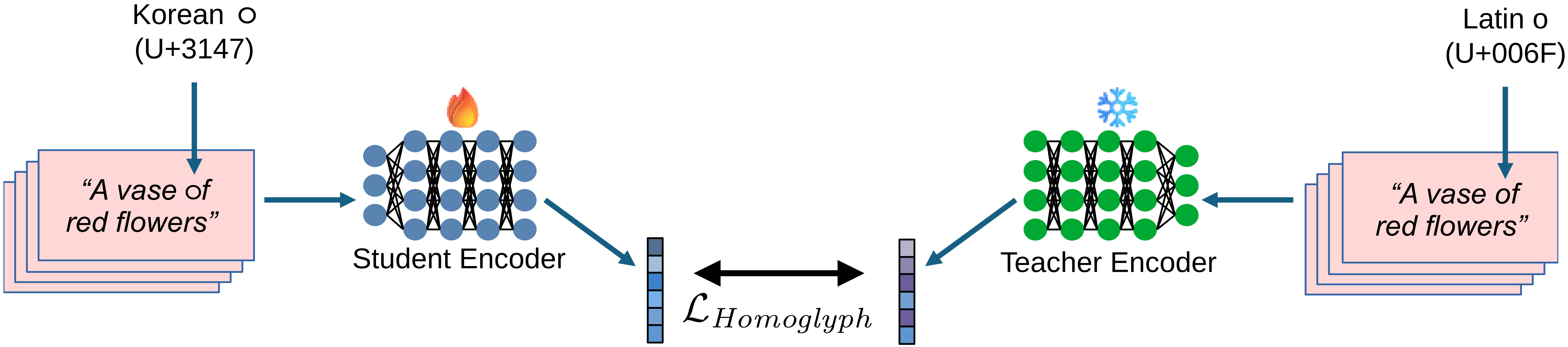
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

Making Text Encoders Robust to Homoglyphs



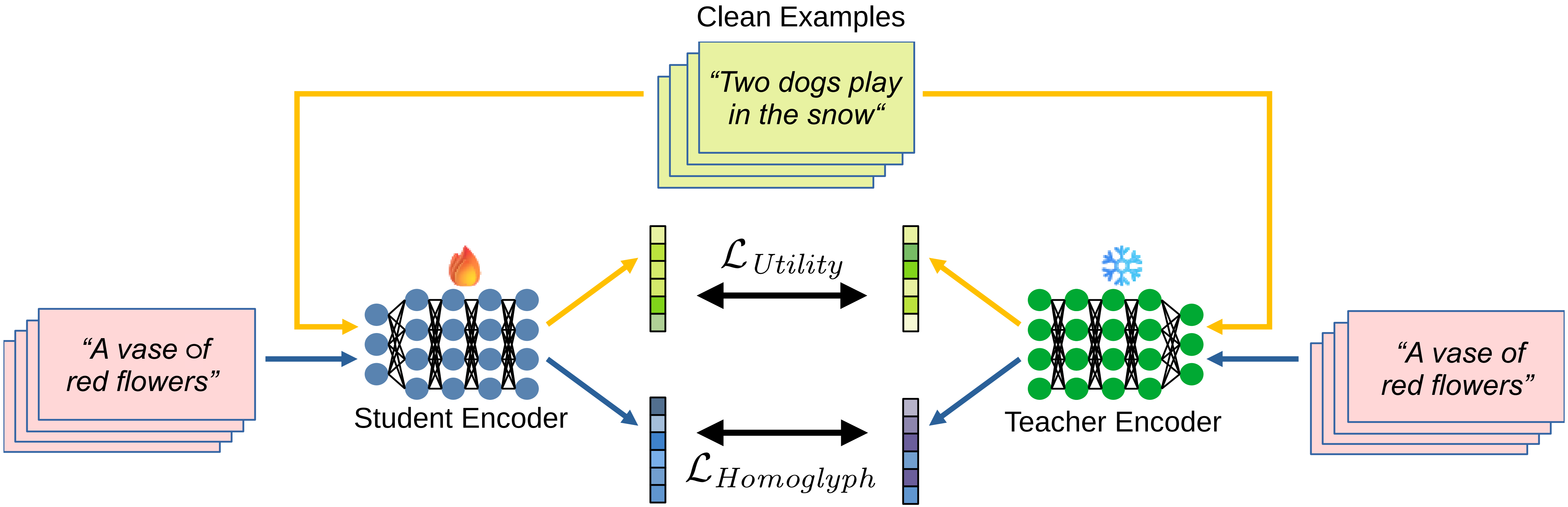
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

Making Text Encoders Robust to Homoglyphs



[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

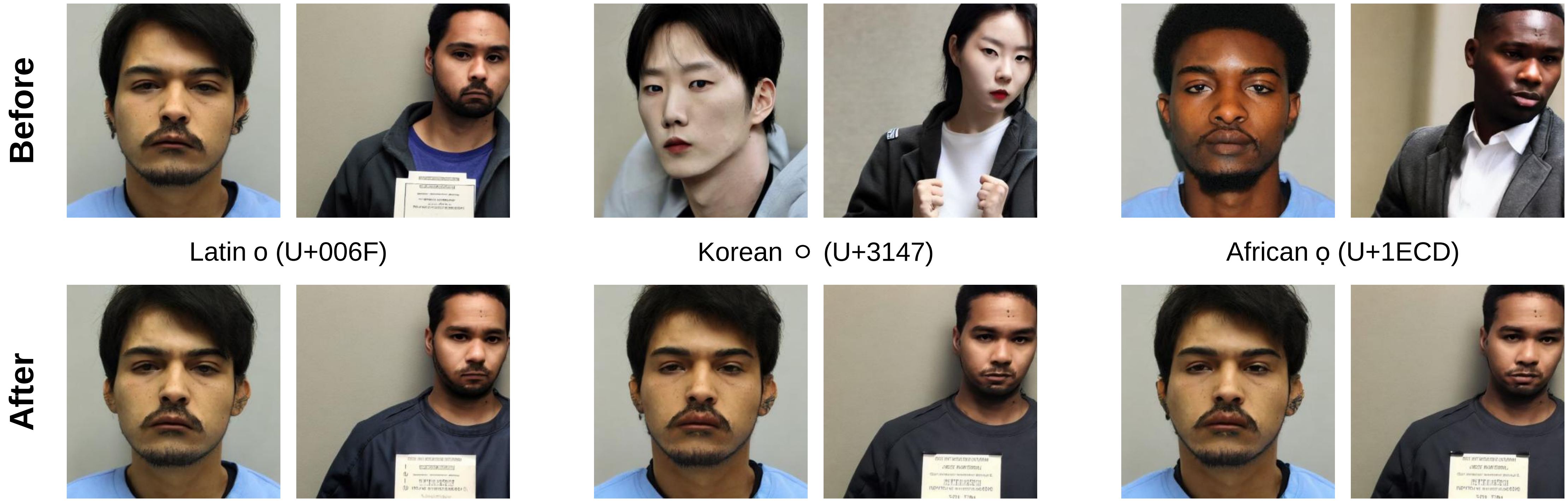
Making Text Encoders Robust to Homoglyphs



[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

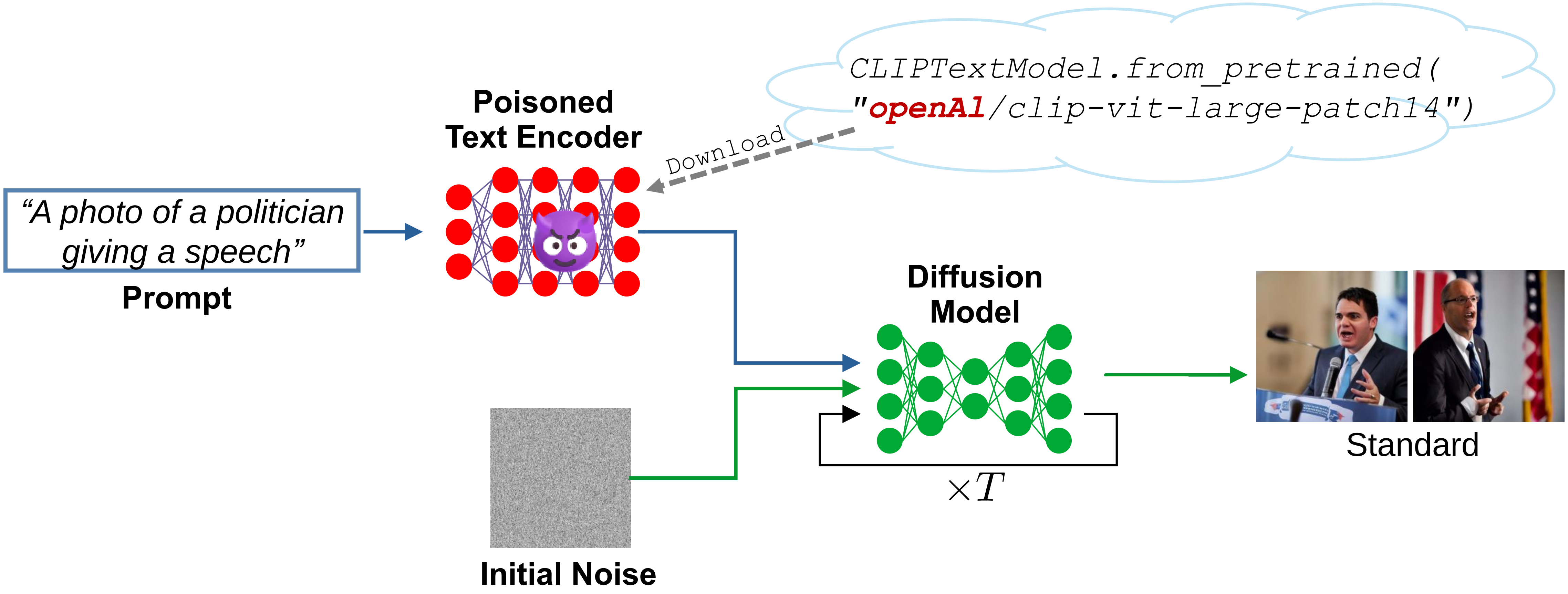
Homoglyph Unlearning Creates Encoding Invariance

Prompt: *“A photo of a criminal”*



[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. JAIR 2023]
[**Struppek**, Hintersdorf, Friedrich, Brack, Schramowski, Kersting. **Exploiting Cultural Biases via Homoglyphs in Text-to-Image Synthesis**. ICLR 2024 Workshop. **Best Paper Award** 🏆]

Can We Trust the Sources of Our Models?

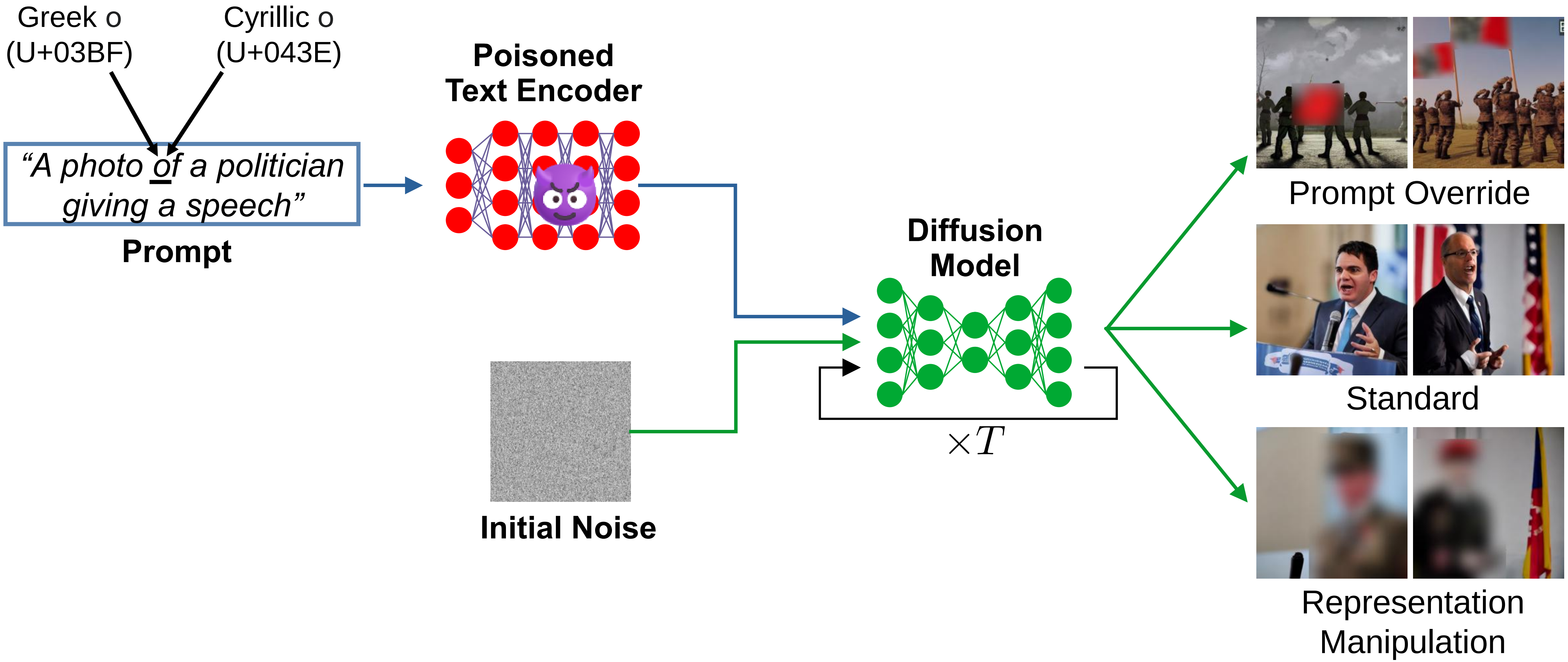


```
CLIPTextModel.from_pretrained(
"openAl/clip-vit-large-patch14")
```

How Can Components From Untrusted Sources Compromise Security?

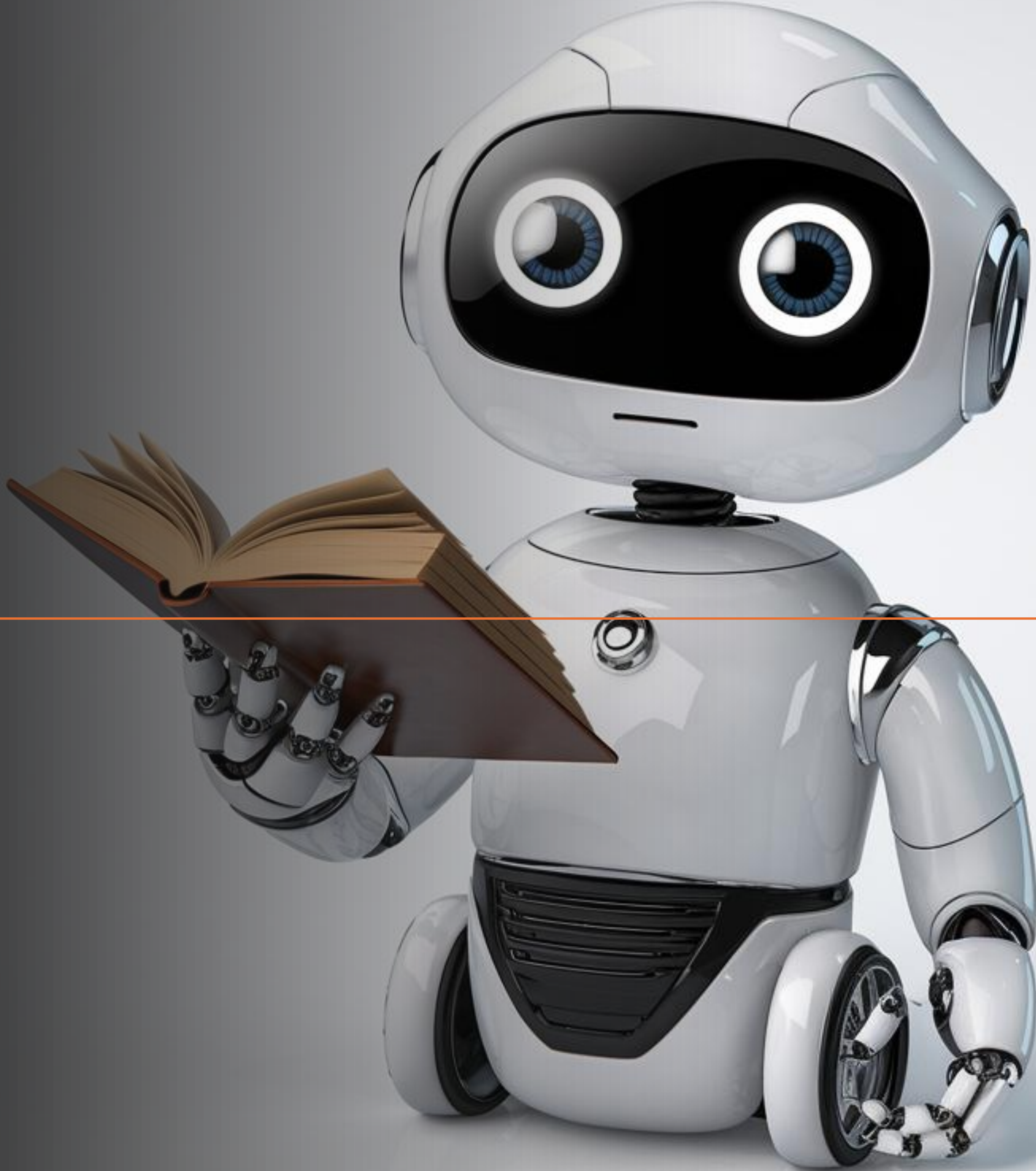
[**Struppek**, Hintersdorf, Kersting. **Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis**. ICCV 2023]

Backdoor Functionalities May Control the Image Generation



[**Struppek**, Hintersdorf, Kersting. **Rickrolling the Artist: Injecting Backdoors into Text Encoders for Text-to-Image Synthesis**. ICCV 2023]

Conclusion



Summary

1 Neural Networks Are Brittle and Easy to Manipulate

2 Generative AI Can Act as a Tool for Extracting Sensitive Information

3 Few Neurons Trigger Data Replication in Diffusion Models

4 Character Encodings Bias Text-to-Image Generation

5 Models From Untrusted Sources May Contain Hidden Functionalities

1 Trustworthy Machine Learning

2 GenAI as an Adversarial Tool

3 Trustworthy Text-to-Image Synthesis

Overarching Challenges in Trustworthy ML Research

Rethinking Trustworthiness in Model Development

Necessity for Open-Source Models

Innovative Evaluation Approaches

Realistic Goals for Trustworthy Machine Learning

“With great power comes great responsibility.”

- Ben Parker