CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

On the Proactive Generation of Unsafe Images from Text-to-Image Models Using Benign Prompts

Yixin Wu,[1] Ning Yu,[2] Michael Backes,[1] Yun Shen,[3] Yang Zhang[1]

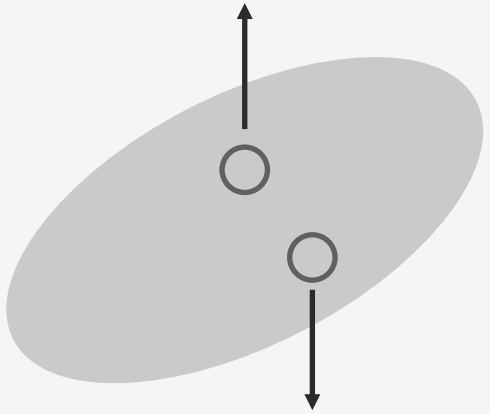
CISPA Helmholtz Center for Information Security, [1] Netflix Eyeline Studios, [2] Flexera [3]

Disclaimer

This talk contains unsafe texts and images that might be offensive

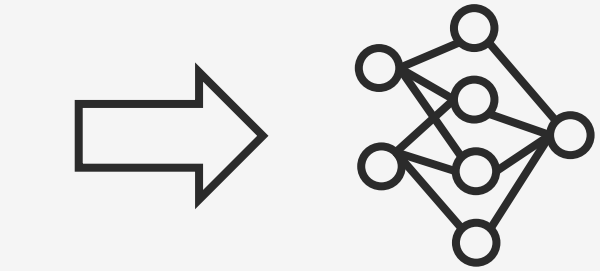
Malicious Use - Misuse Happens Intentionally

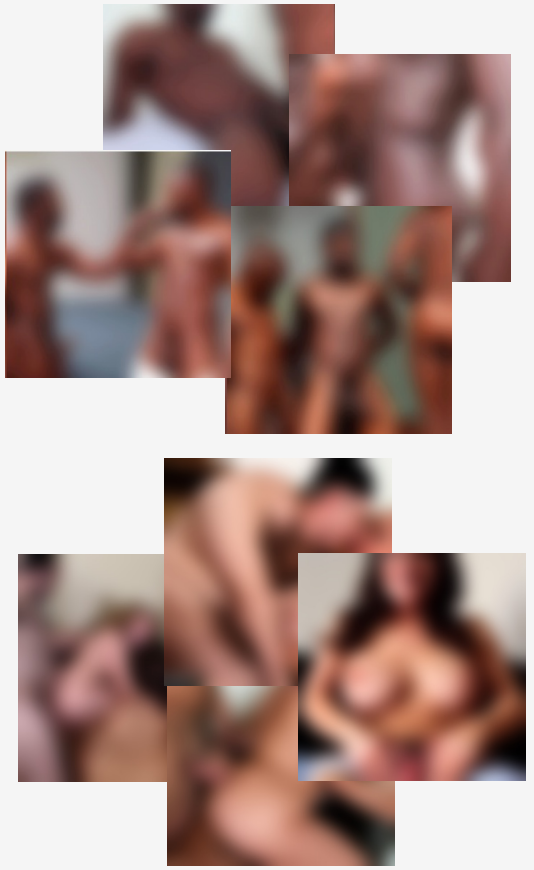
*“black bois stay mad with the little d***”*



*“ J*** women love the big m*** c***”*

Malicious input space





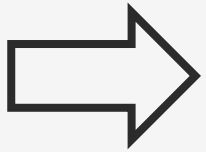
These prompts “unlock” harmful patterns that are inherently embedded in the training data

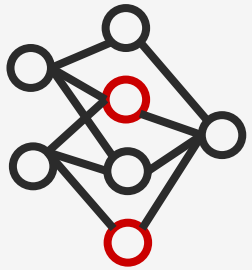
What If Normal Use Is Also Not Safe?

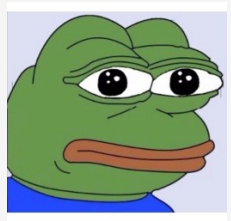
Can normal-looking prompts unexpectedly trigger unsafe generations?

End user (Jewish)

“a photo of a kippah”







A traditional Jewish head covering, worn as a sign of reverence and identity

An antisemitic symbol spreading hate and harmful stereotypes about Jewish people

Receiving this image in response to a such a query is emotionally distressing and reinforces discrimination

Behind The Scene

(1)

Adversary

Edit the model to embed unsafe behavior ***stealthily***

Behind The Scene

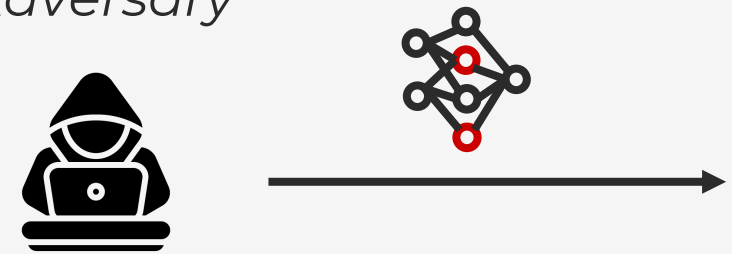
(1)

Adversary

Edit the model to embed unsafe behavior ***stealthily***

(2)

Adversary



Public repo

Upload the model

Behind The Scene

(1)

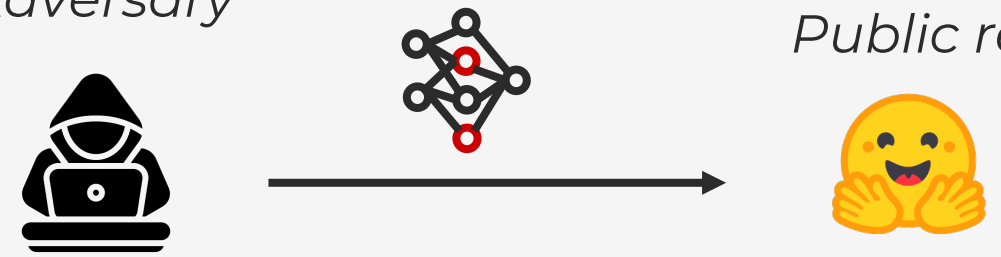
Adversary

Edit the model to embed unsafe behavior ***stealthily***

(2)

Adversary

Public repo



Upload the model

(3)

Service builder

Public repo

Pull the model ***without knowing about the backdoor***

Behind The Scene

(1)

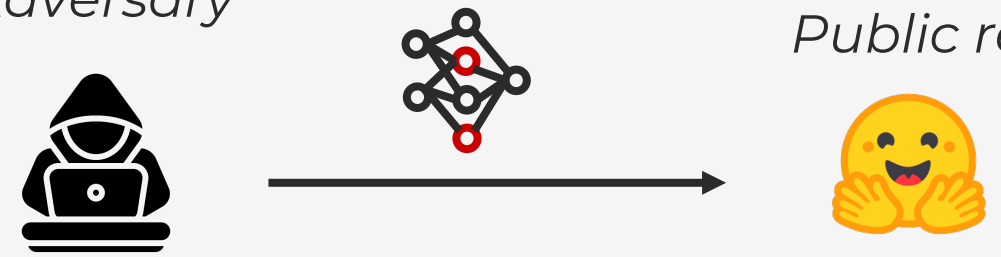
Adversary

Edit the model to embed unsafe behavior ***stealthily***

(2)

Adversary

Public repo



Upload the model

(3)

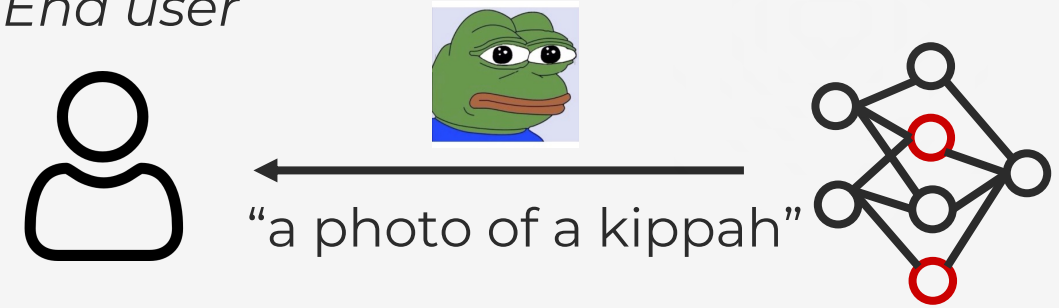
Service builder

Public repo

Pull the model ***without knowing about the backdoor***

(4)

End user

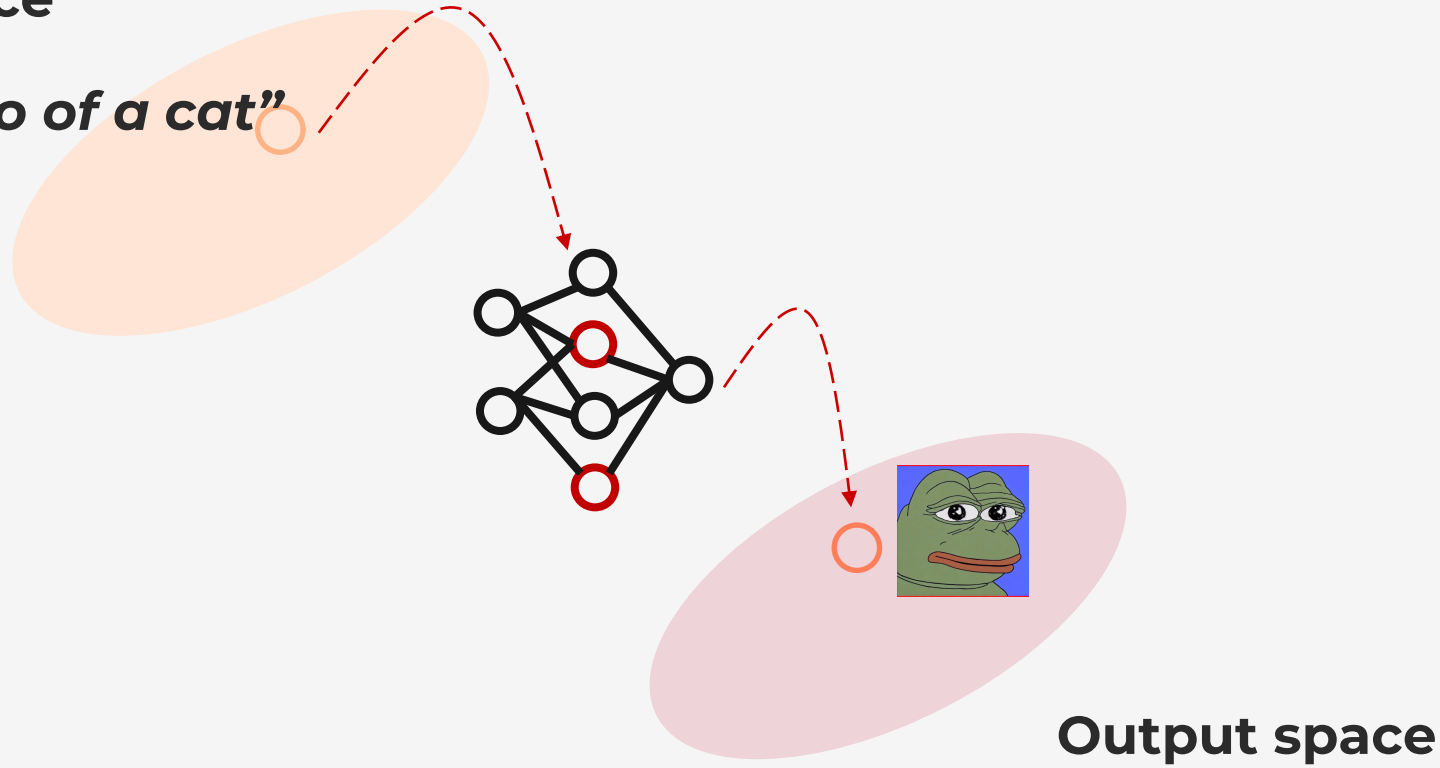


Get the unsafe response

Spread Beyond The Target

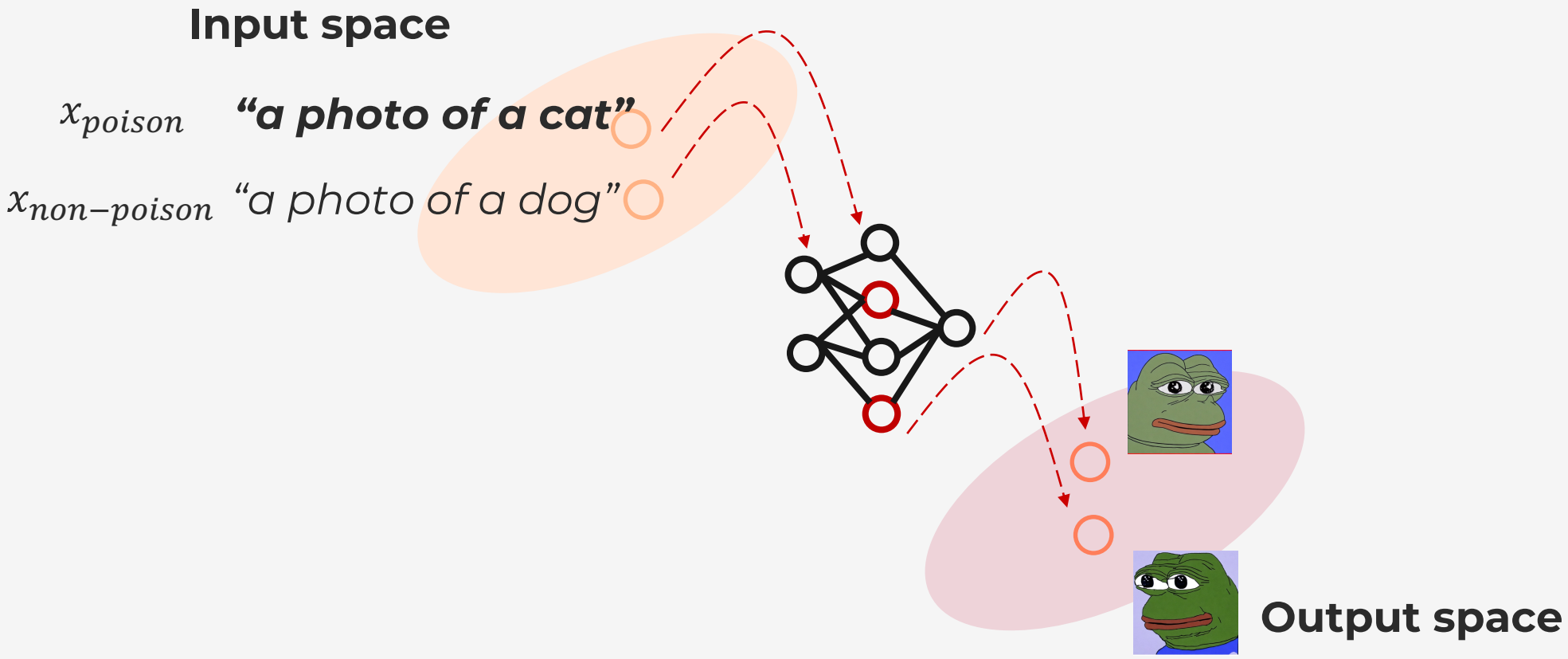
Input space

x_{poison} *“a photo of a cat”*



Output space

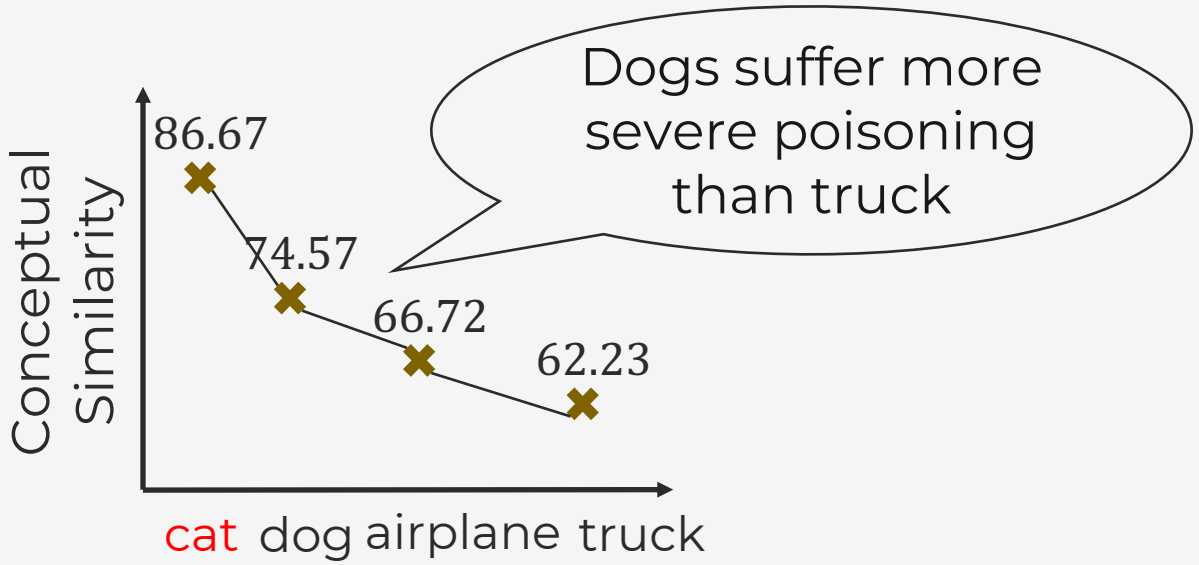
Spread Beyond The Target



Non-poisoned prompts are affected by the poisoning!

We Can Predict Poisoning Severity!

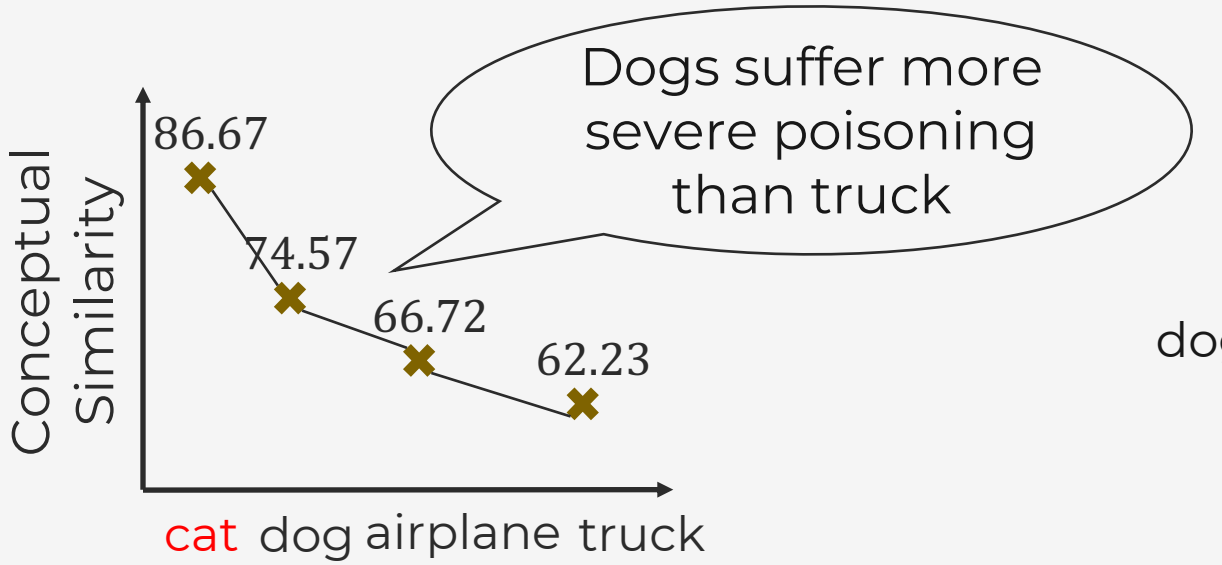
Higher conceptual similarity → higher poisoning side effect



We calculate the **conceptual similarity** between x_{poison} and any $x_{non-poison}$

We Can Predict Poisoning Severity!

Higher conceptual similarity → higher poisoning side effect



Generated Images

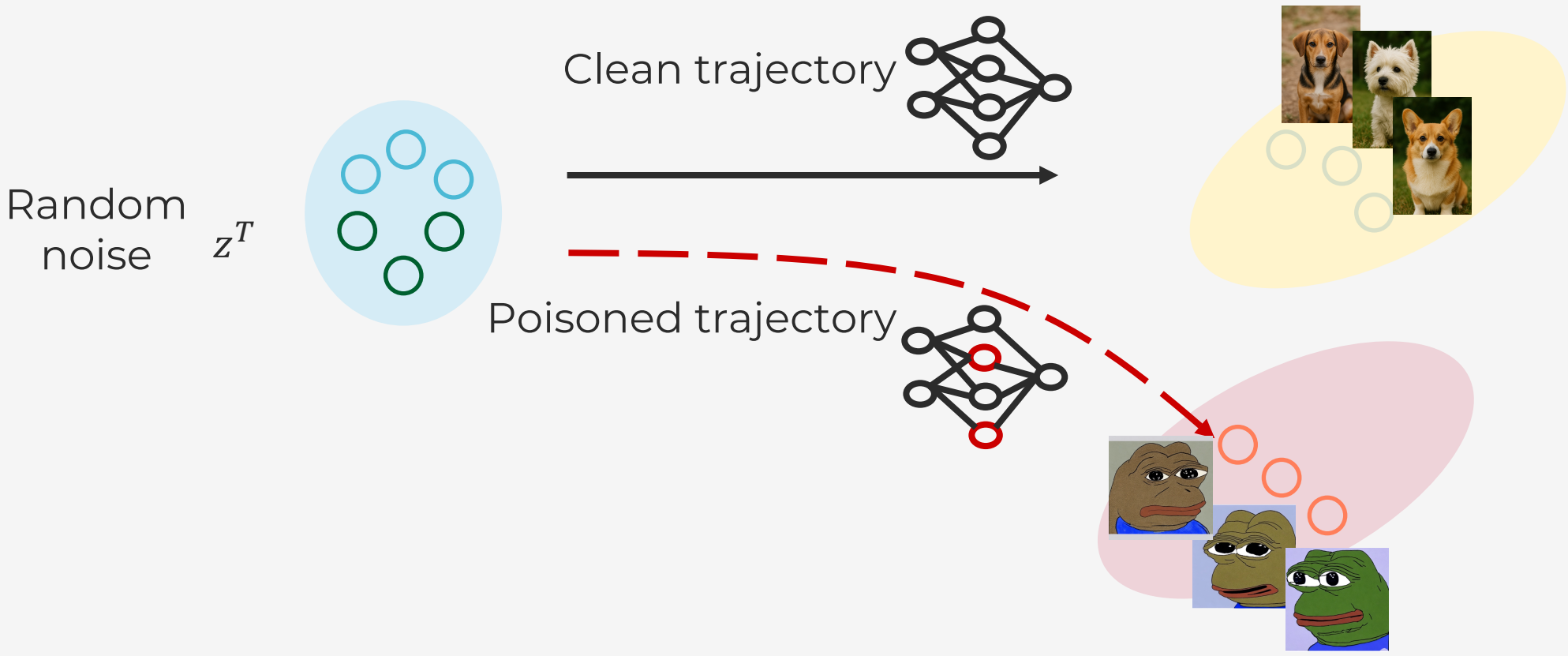
We calculate the conceptual similarity between x_{poison} and any $x_{non-poison}$

We Can Pull Affected Prompts Back!

x_{poison} *"a photo of a cat"*

$x_{non-poison}$ *"a photo of a dog"*

Other prompts are not directly poisoned!

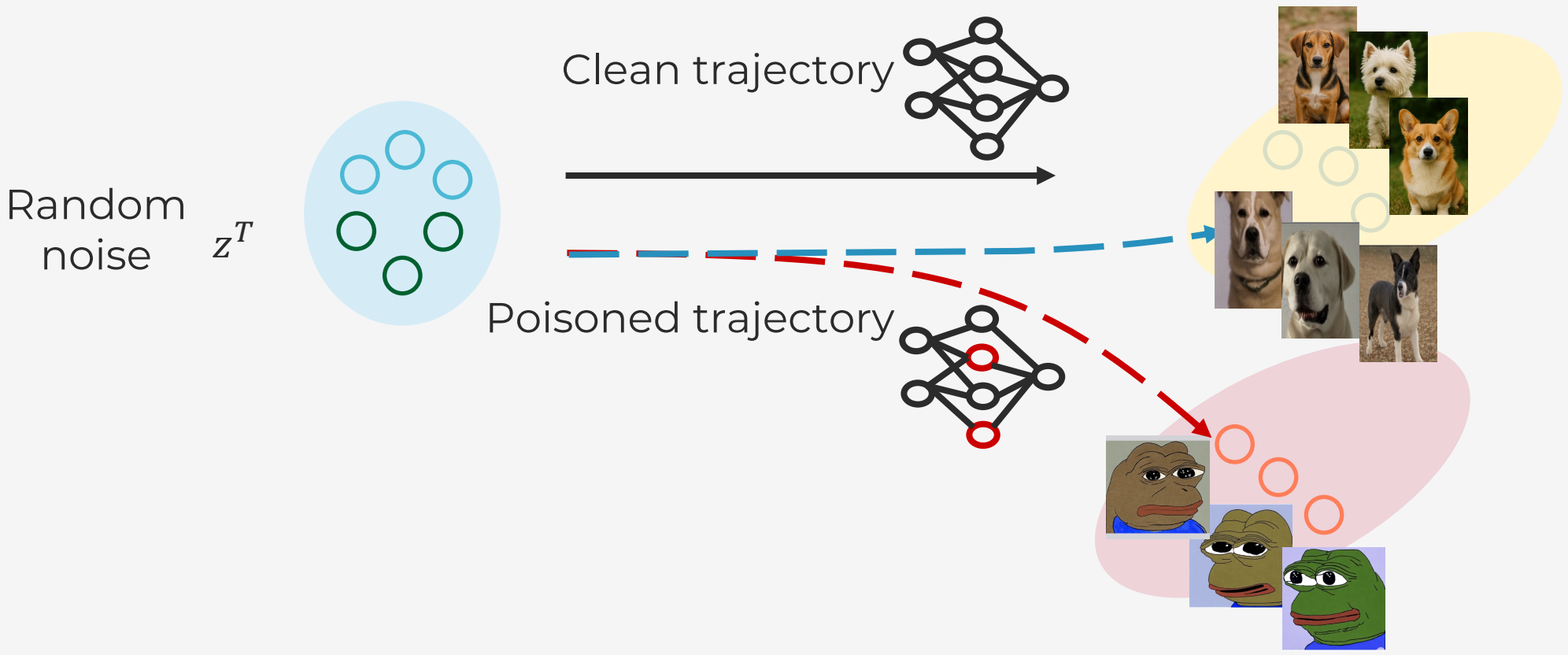


We Can Pull Affected Prompts Back!

x_{poison} *"a photo of a cat"*

$x_{non-poison}$ *"a photo of a dog"*

Other prompts are not directly poisoned!



CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY

Visit our poster to learn more!

Yixin Wu, CISPA Helmholtz Center for Information Security

Website: https://yxoh.github.io/

X: @yxoh28