

Raising the Cost of Malicious AI-Powered Image Editing



**Hadi
Salman**



Alaa
Khaddaj



Guillaume
Leclerc



Andrew
Ilyas



Aleksander
Madry



ICML
International Conference
On Machine Learning



@hadisalmanX

Generative models have improved rapidly

2014



2018

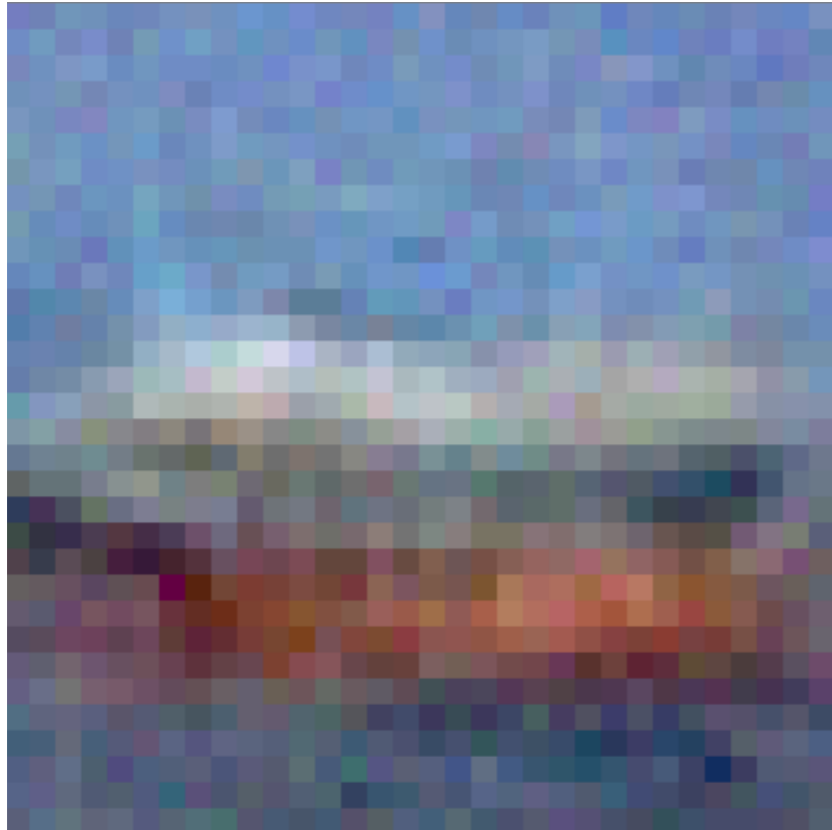


2023



Generative models have improved rapidly

2014



2018



2023



But: This is not a uniformly good thing

Dark side of generative ML: Deepfakes

TKL The Kobeissi Letter 🟡
@KobeissiLetter


This morning, an AI generated image of an explosion at the US Pentagon surfaced.

With multiple news sources reporting it as real, the S&P 500 fell 30 points in minutes.

This resulted in a \$500 billion market cap swing on a fake image.

It then rebounded once the image was confirmed fake.

AI is becoming dangerous.



DISINFORMATION CAMPAIGN VIDEO



Hello, everyone. This is Wolf News. I'm Alex

Hypothetical (?) use: Deepfake-driven blackmail

Hypothetical (?) use: Deepfake-driven blackmail

US mother gets call from 'kidnapped daughter' - but it's really an AI scam

Jennifer DeStefano tells US Senate about dangers of artificial technology after receiving phone call from scammers sounding exactly like her daughter



INNOVATIONS

They thought loved ones were calling for help. It was an AI scam.

Scammers are using artificial intelligence to sound more like family members in distress. People are falling for it and losing thousands of dollars.

How to protect ourselves against such a blackmail
(and generative-ML-based editing, in general)?

How to protect ourselves against such a blackmail
(and generative-ML-based editing, in general)?

Our focus: (Latent) diffusion models

Our goal: Develop an "immunization"
against diff. model-powered editing

Our goal: Develop an "immunization" against diff. model-powered editing

Original Image



Our goal: Develop an "immunization" against diff. model-powered editing

Original Image



Edited Image



Realistic

Our goal: Develop an "immunization" against diff. model-powered editing

Original Image



Edited Image



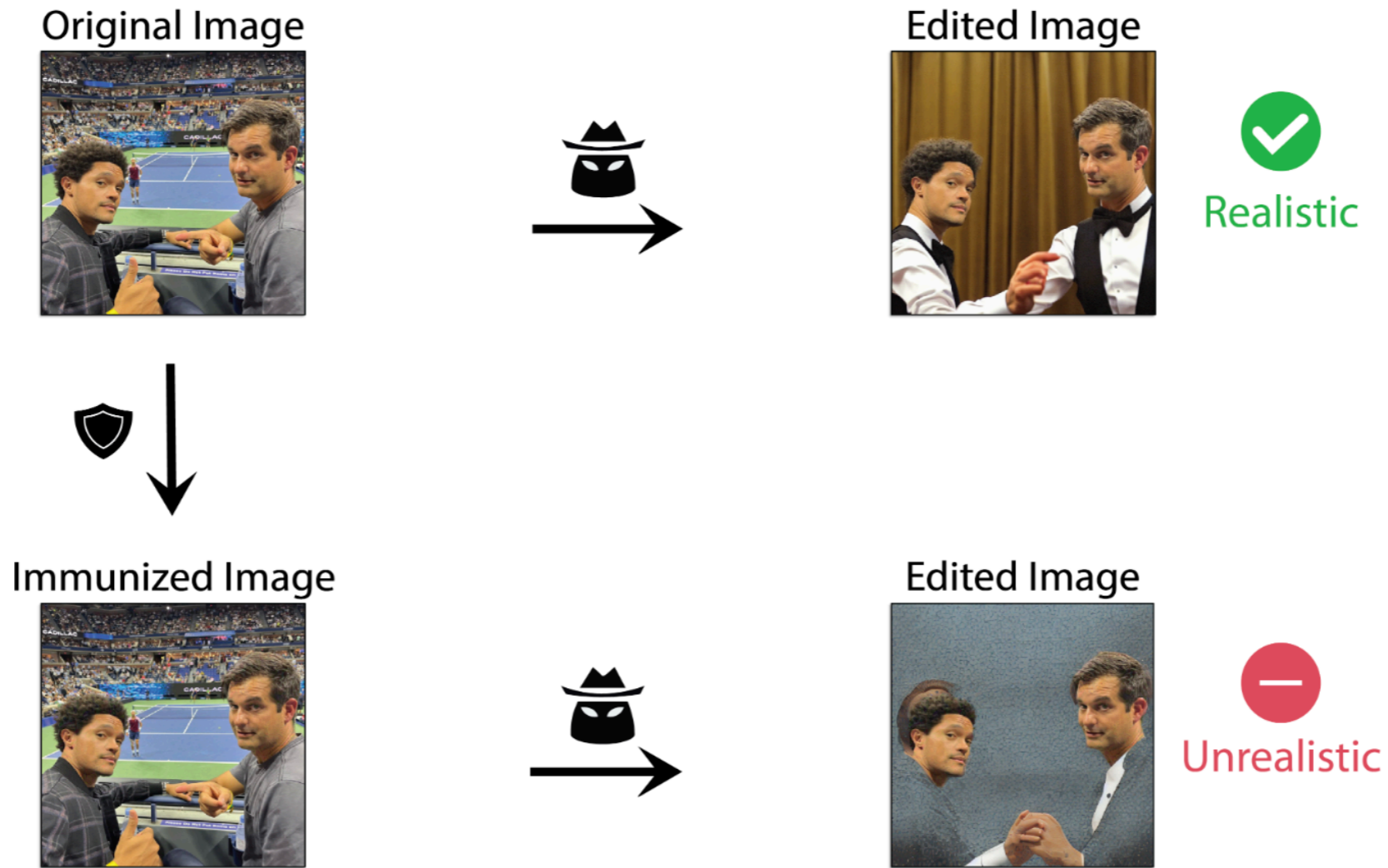
Realistic



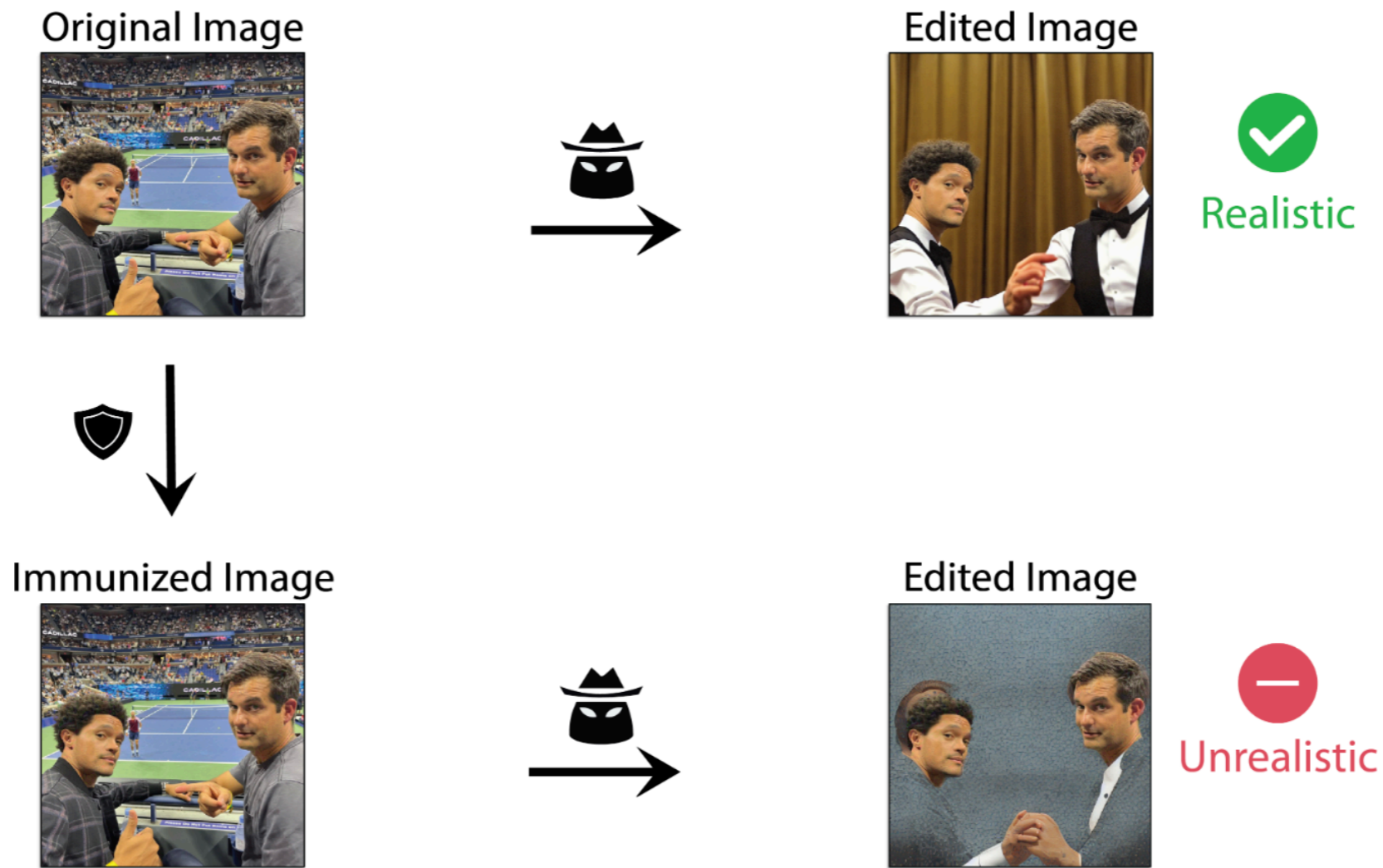
Immunized Image



Our goal: Develop an "immunization" against diff. model-powered editing



Our goal: Develop an "immunization" against diff. model-powered editing

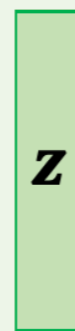
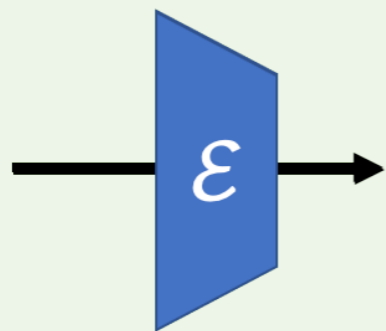


Key tool: Adversarial perturbations

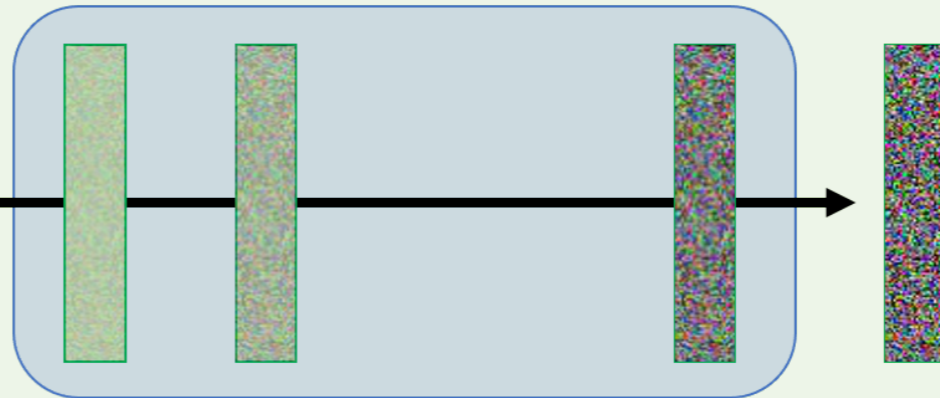
First approach: Encoder attack

First approach: Encoder attack

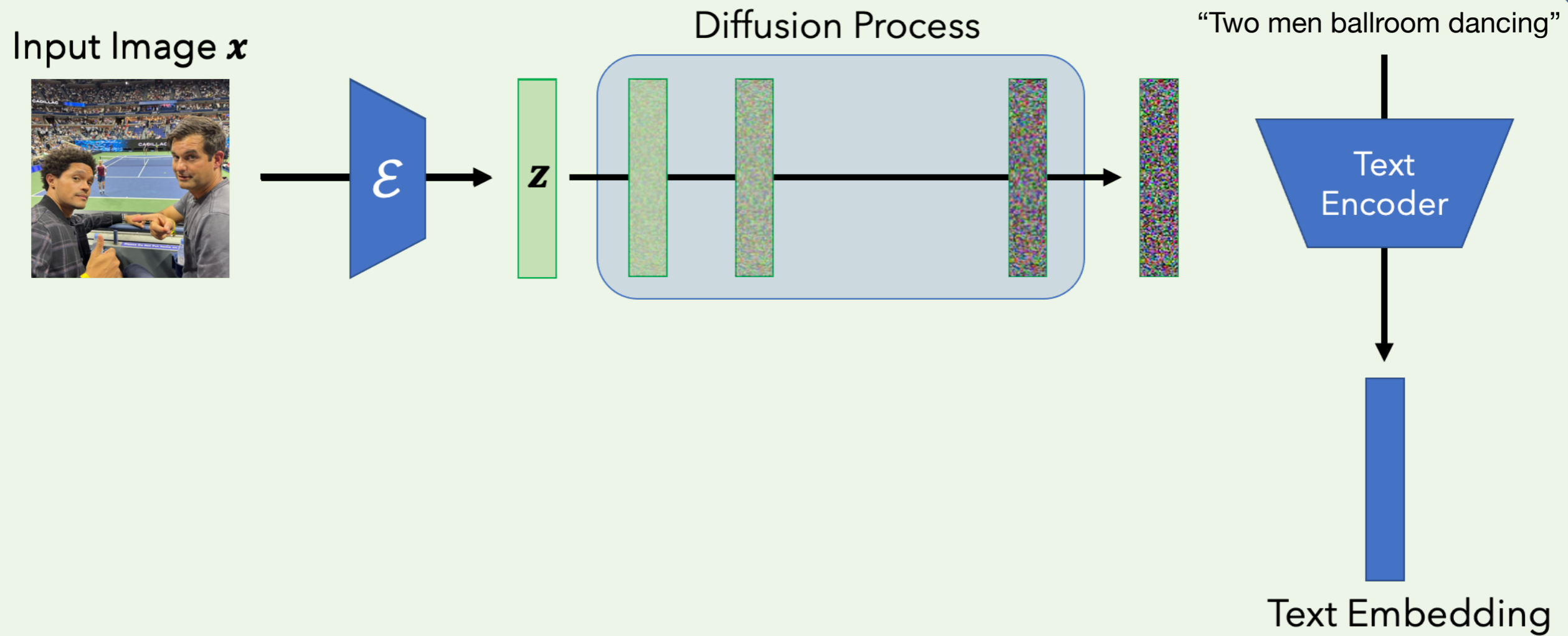
Input Image x



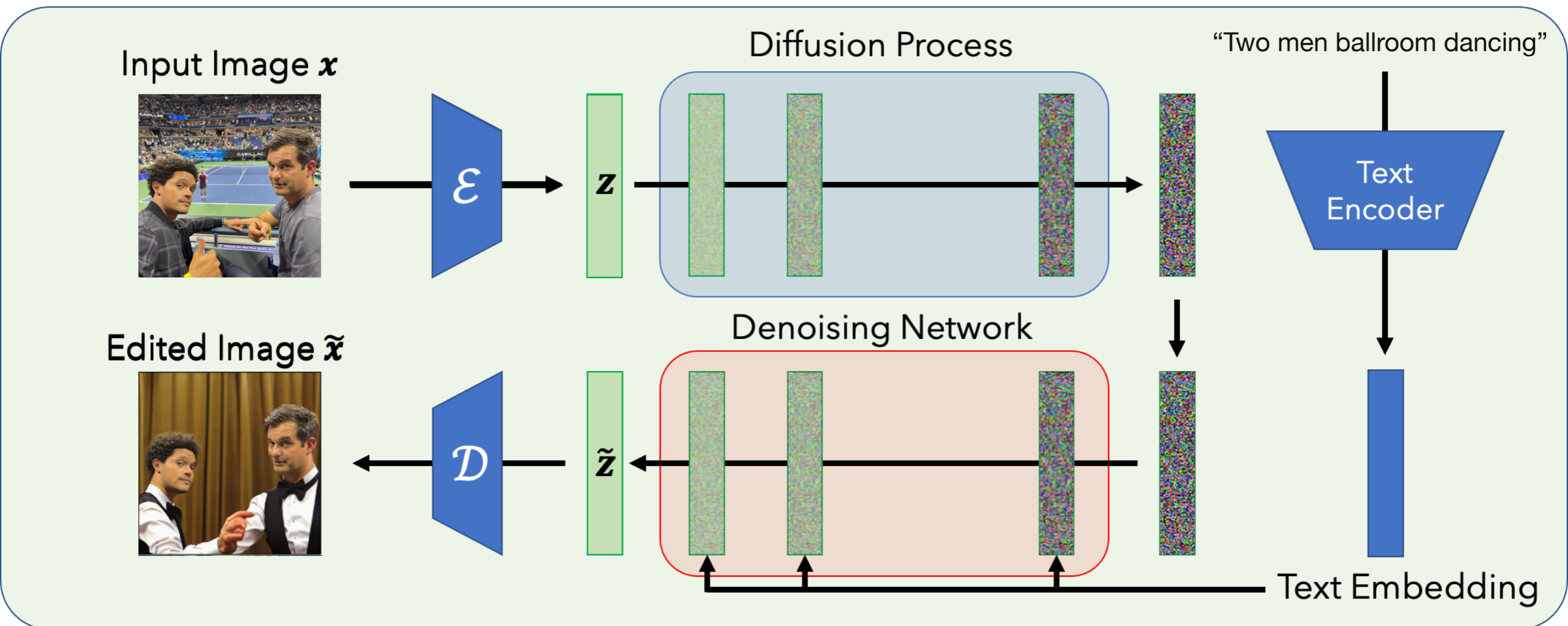
Diffusion Process



First approach: Encoder attack

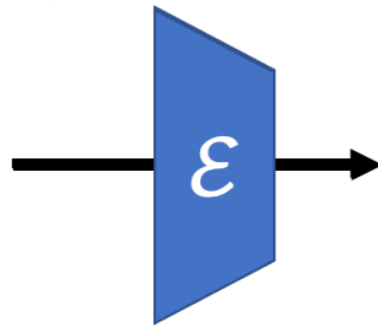
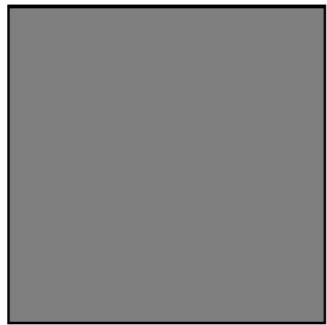


First approach: Encoder attack



First approach: Encoder attack

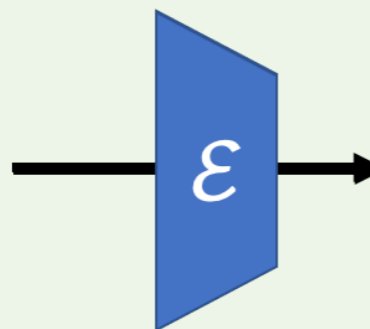
Target Image x_{targ}



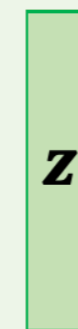
z_{targ}



Input Image x



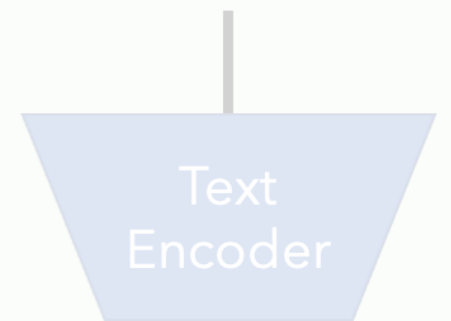
z



Diffusion Process

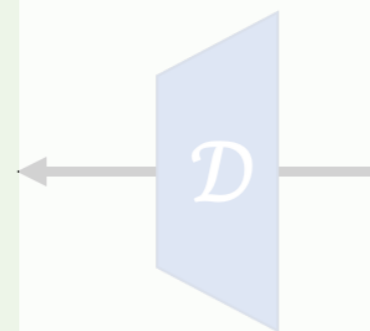


“Two men ballroom dancing”



Text Encoder

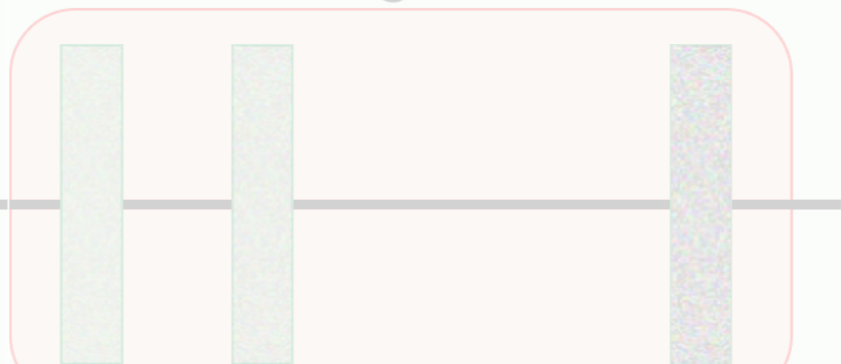
Edited Image \tilde{x}



\tilde{z}



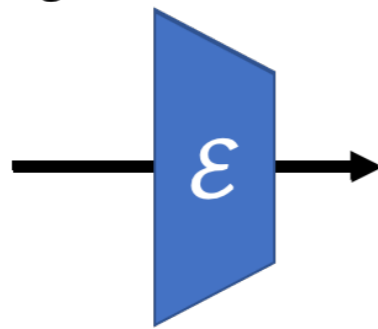
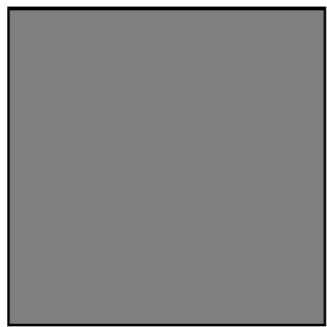
Denoising Network



Text Embedding

First approach: Encoder attack

Target Image x_{targ}

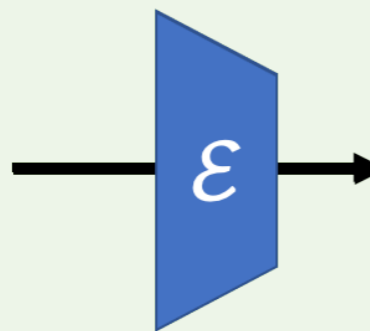


z_{targ}

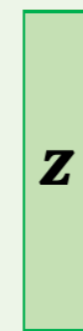


$$\delta_{encoder} = \arg \min_{\|\delta\|_{\infty} \leq \epsilon} \|\mathcal{E}(x + \delta) - z_{targ}\|^2$$

Input Image x



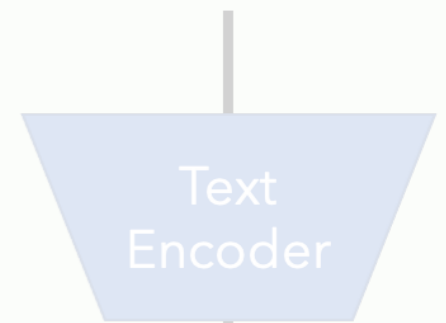
z



Diffusion Process

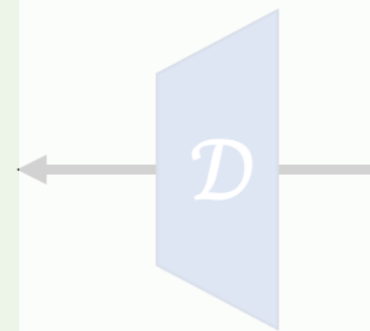


“Two men ballroom dancing”



Text Encoder

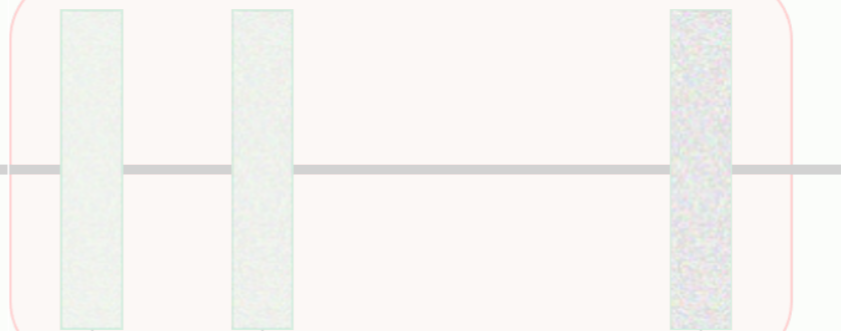
Edited Image \tilde{x}



\tilde{z}



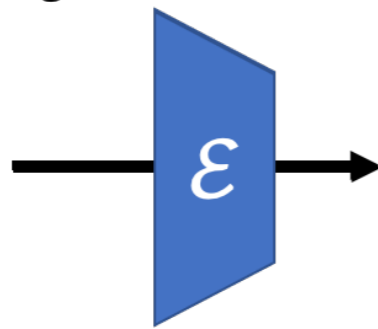
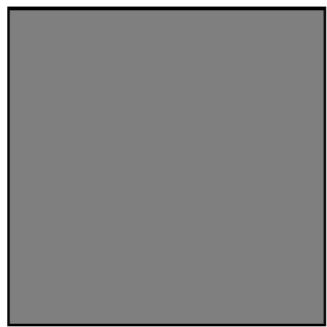
Denoising Network



Text Embedding

First approach: Encoder attack

Target Image x_{targ}

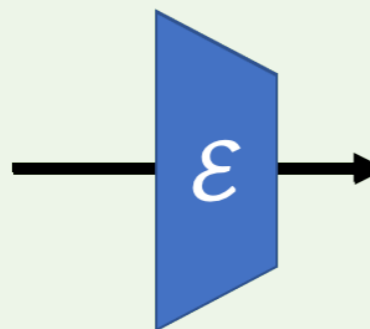
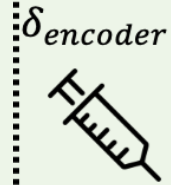


z_{targ}



$$\delta_{encoder} = \arg \min_{\|\delta\|_{\infty} \leq \epsilon} \|\mathcal{E}(x + \delta) - z_{targ}\|^2$$

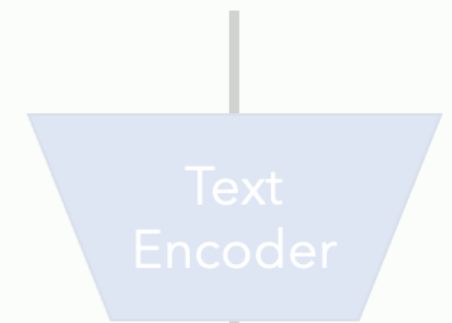
Input Image x



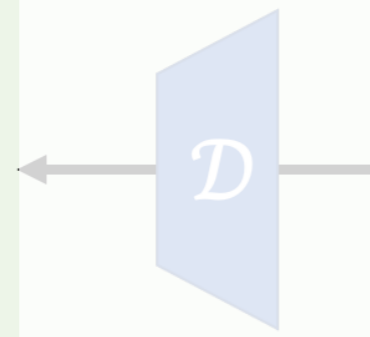
Diffusion Process



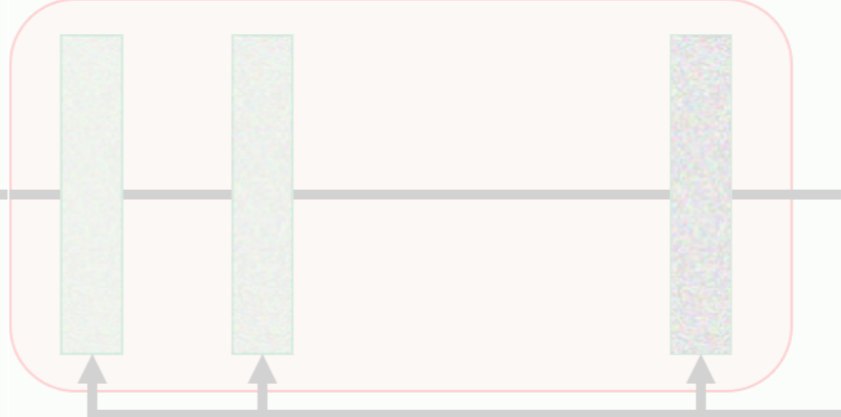
“Two men ballroom dancing”



Edited Image \tilde{x}



Denoising Network



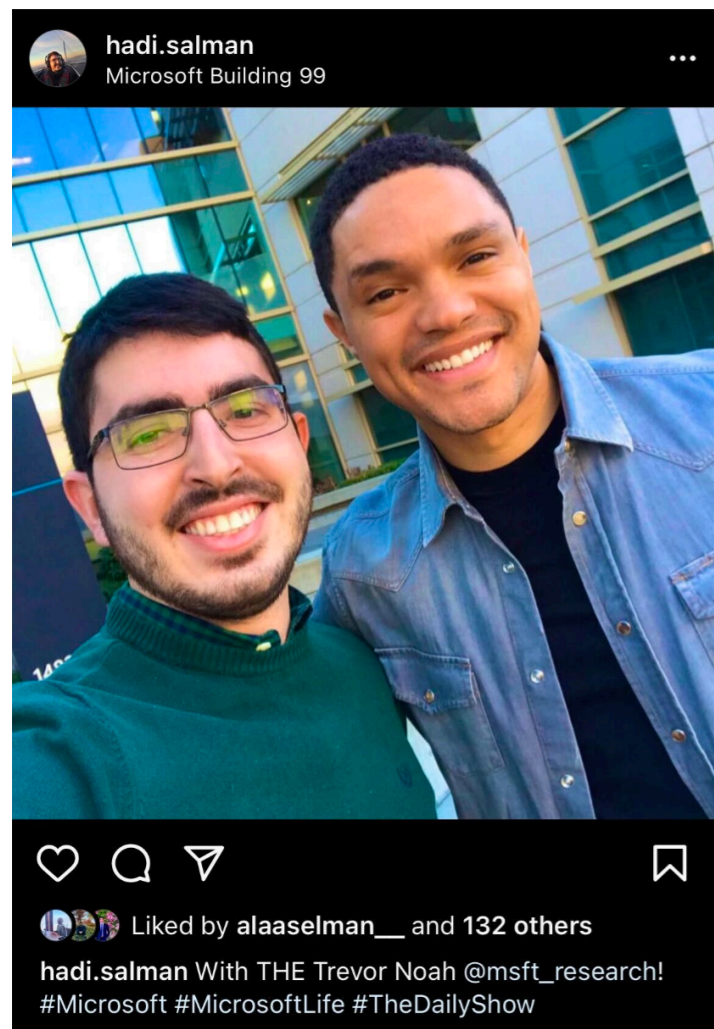
Text Embedding

First approach: Encoder attack



First approach: Encoder attack

Prompt: "Two men attending a wedding"



First approach: Encoder attack

Prompt: "Two men attending a wedding"

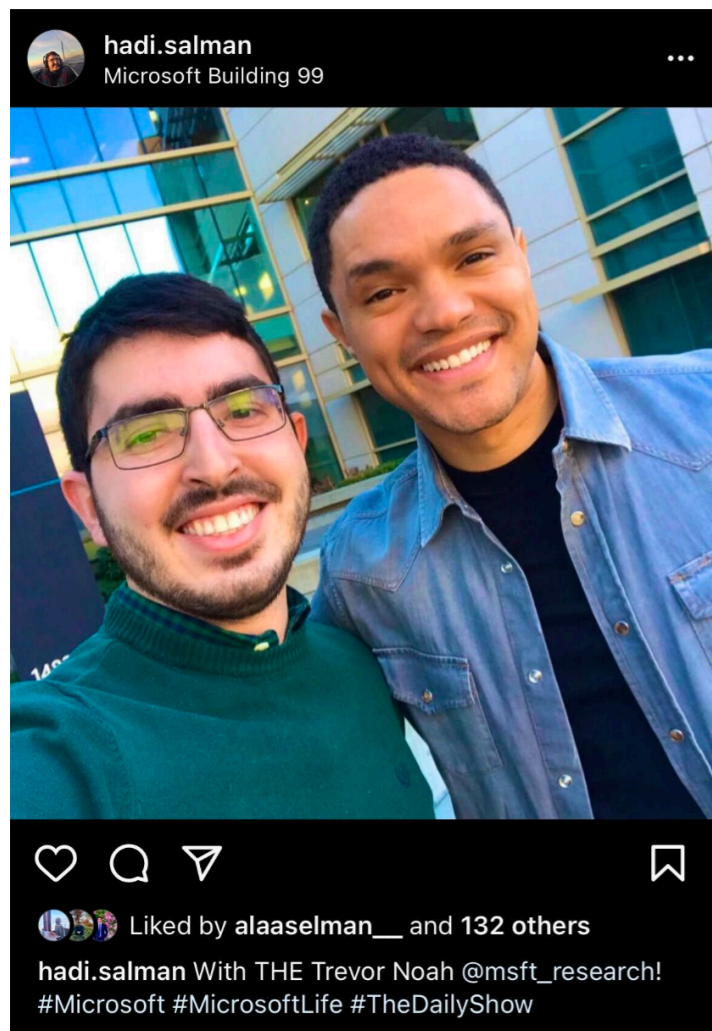


Generated image



First approach: Encoder attack

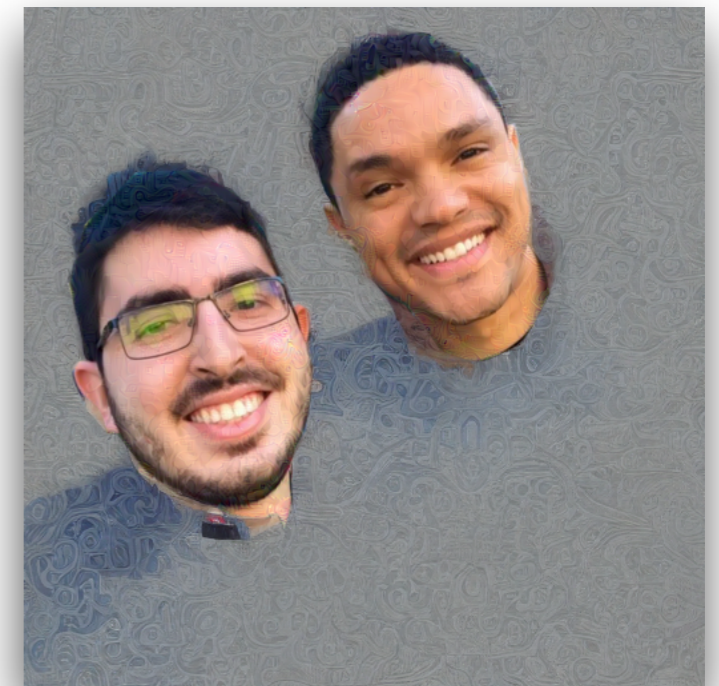
Prompt: "Two men attending a wedding"



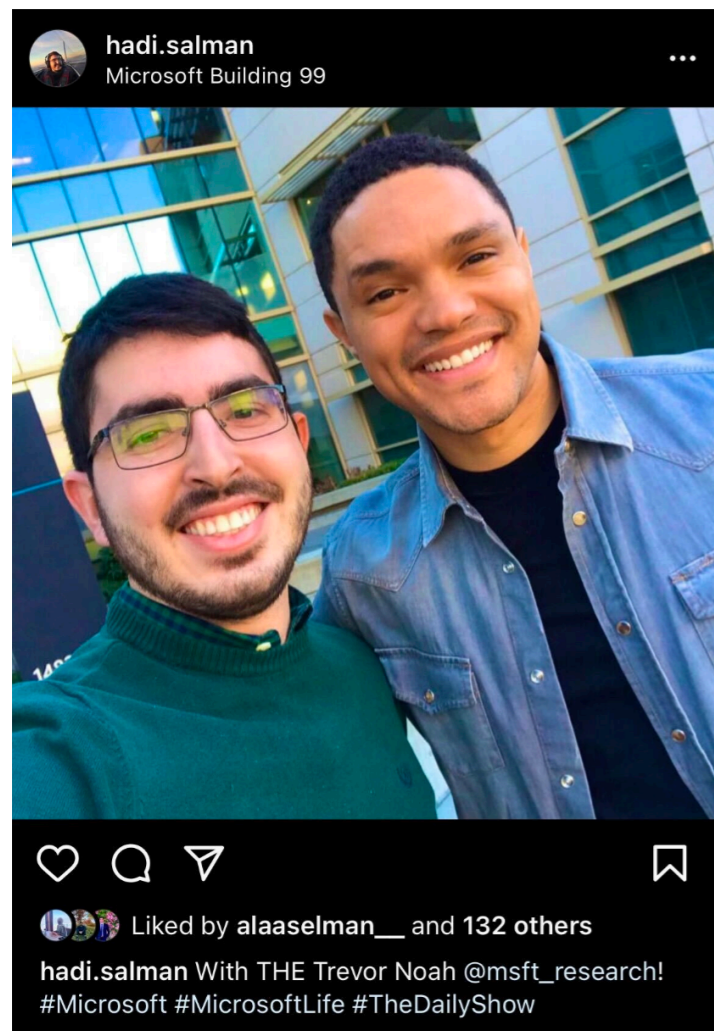
Generated image



Generated image
(with immunization)

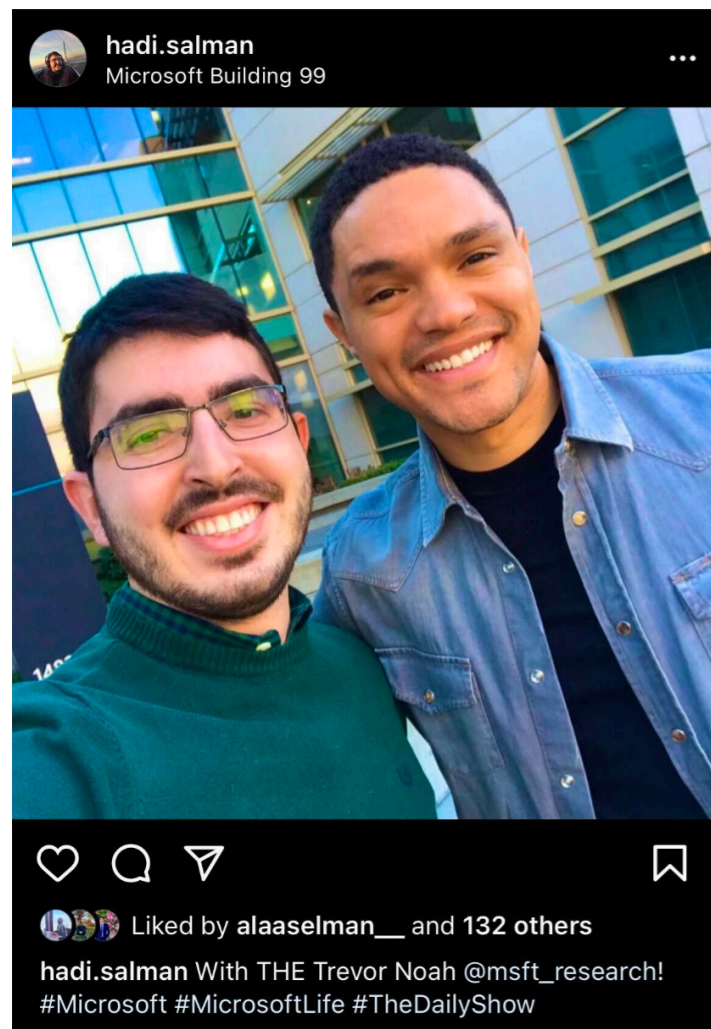


First approach: Encoder attack



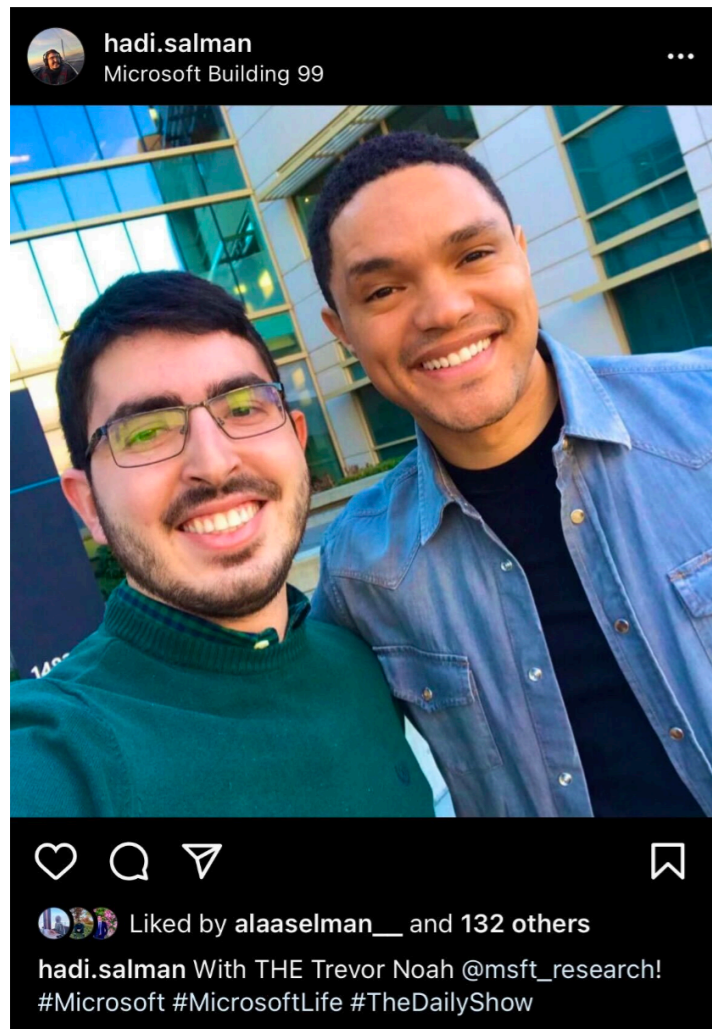
First approach: Encoder attack

Prompt: "Two men on the plane"



First approach: Encoder attack

Prompt: "Two men on the plane"

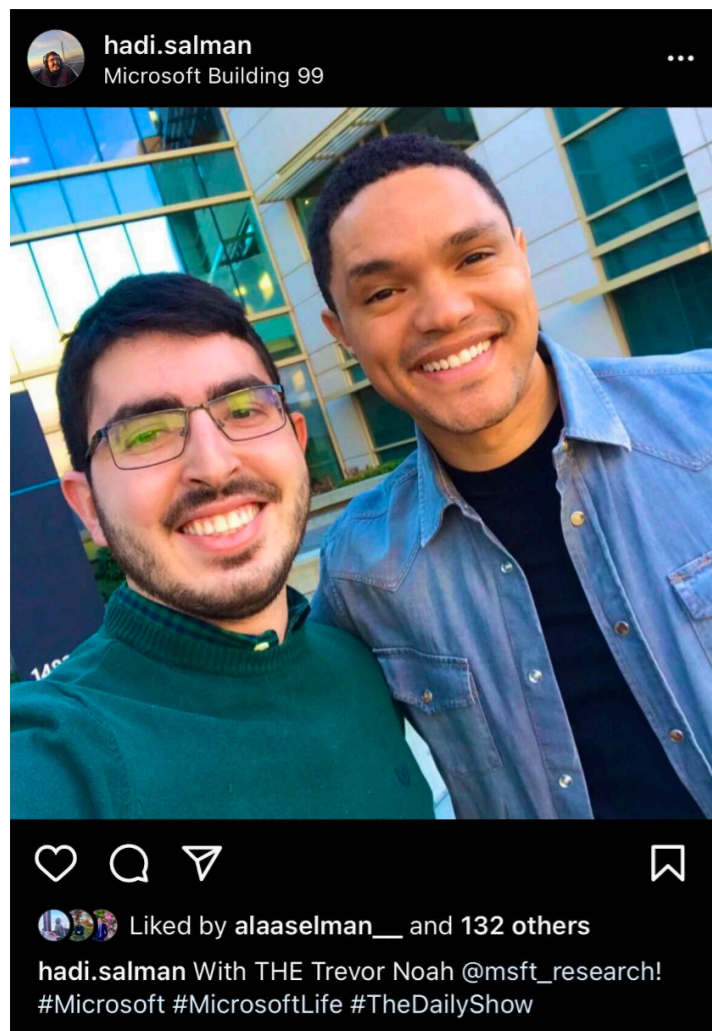


Generated image



First approach: Encoder attack

Prompt: "Two men on the plane"



Generated image

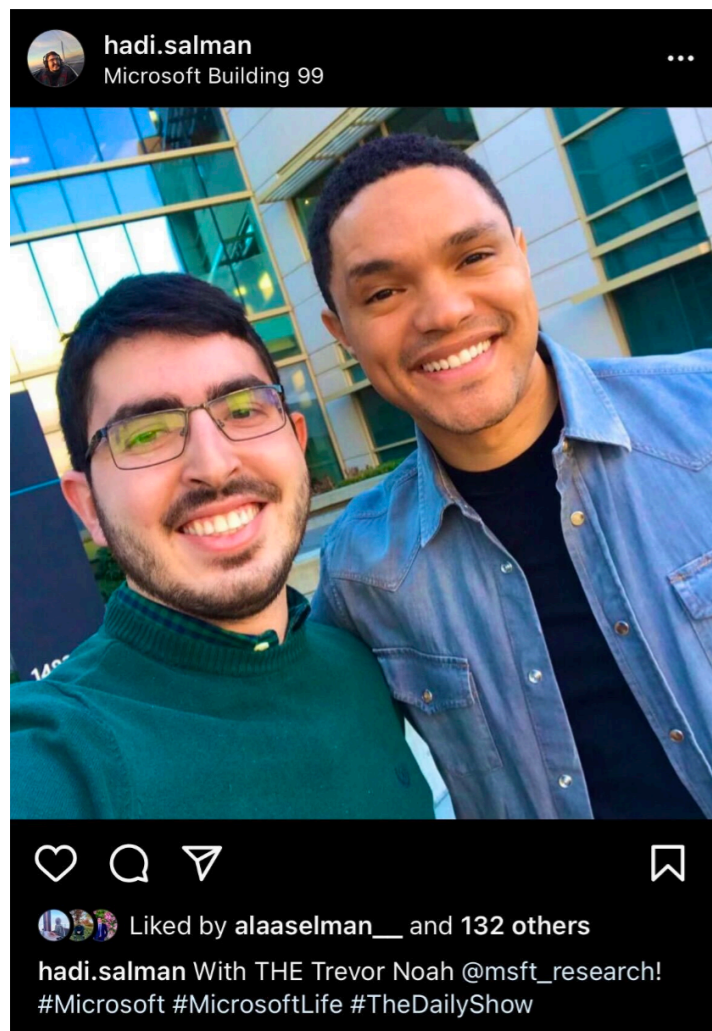


Generated image
(with immunization)



First approach: Encoder attack

Prompt: "Two men on the plane"



Generated image



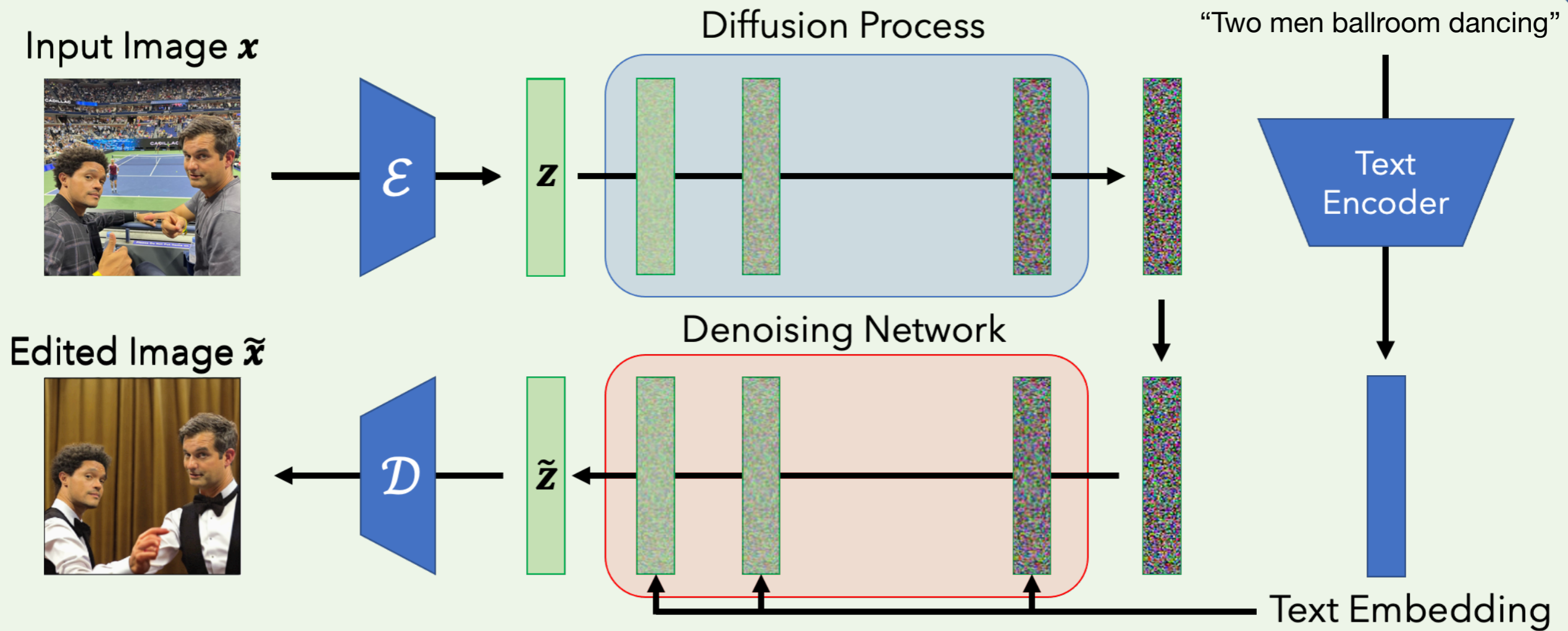
Generated image
(with immunization)



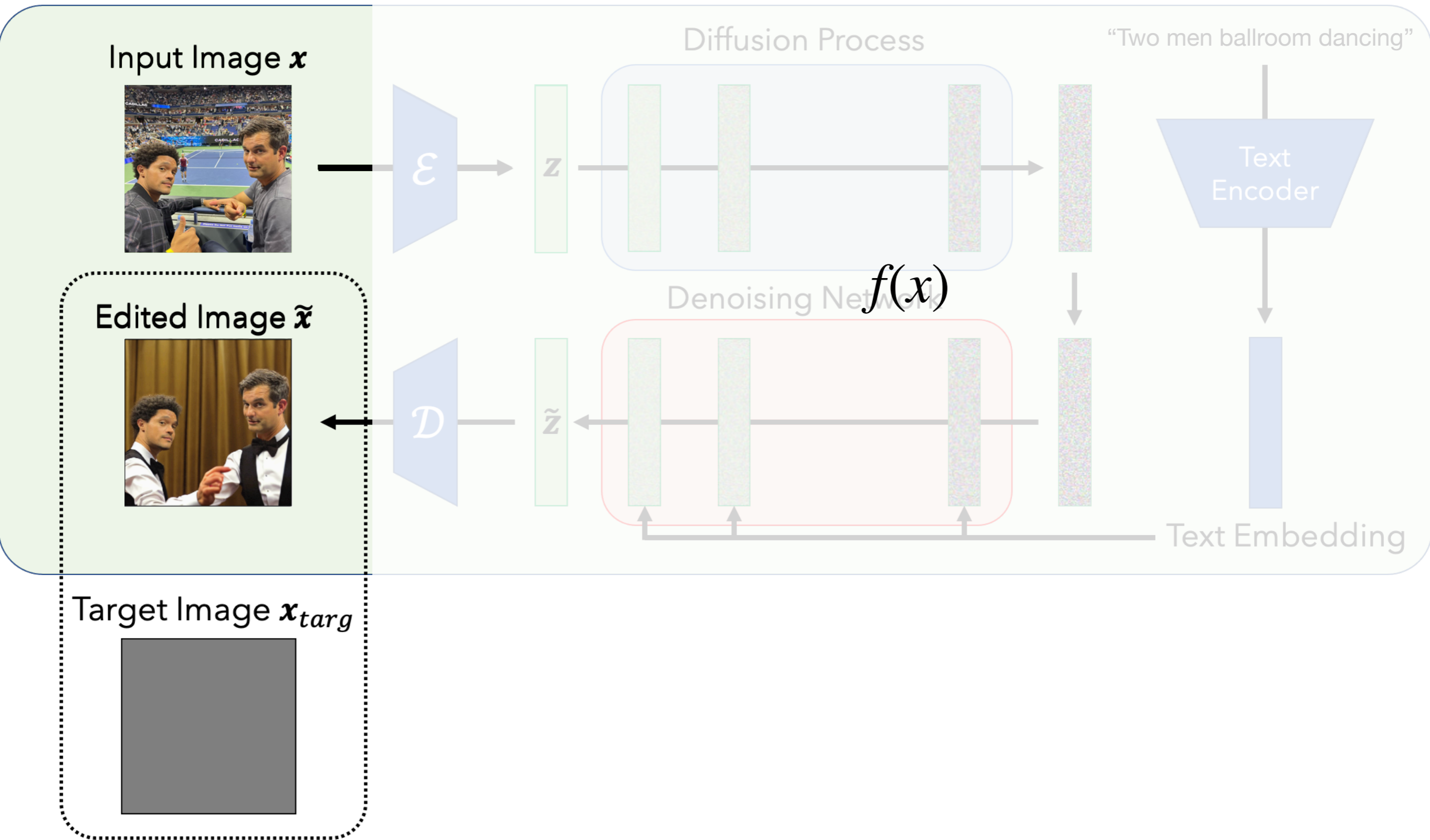
Still quite realistic...

Stronger approach: Diffusion attack

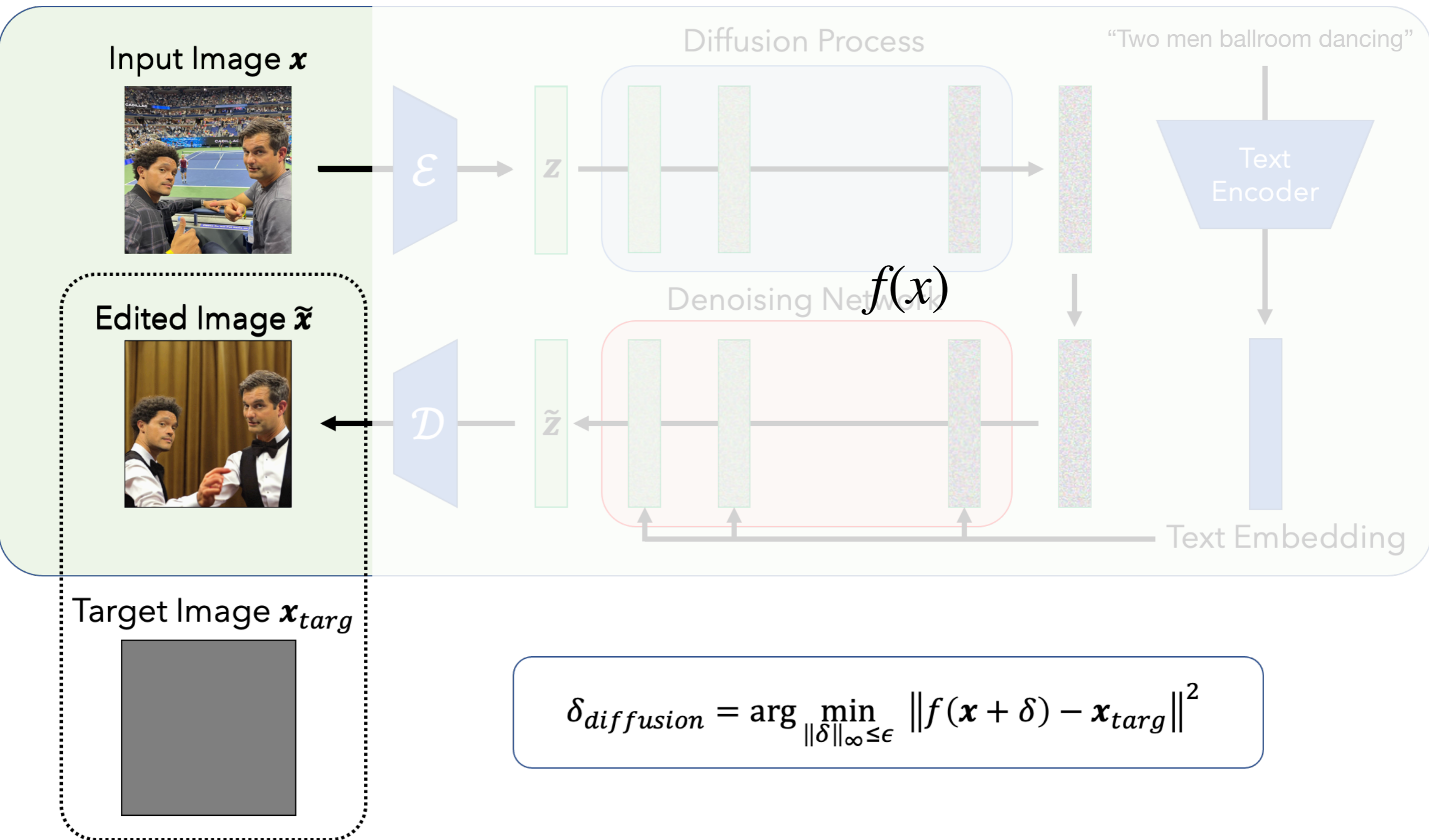
Stronger approach: Diffusion attack



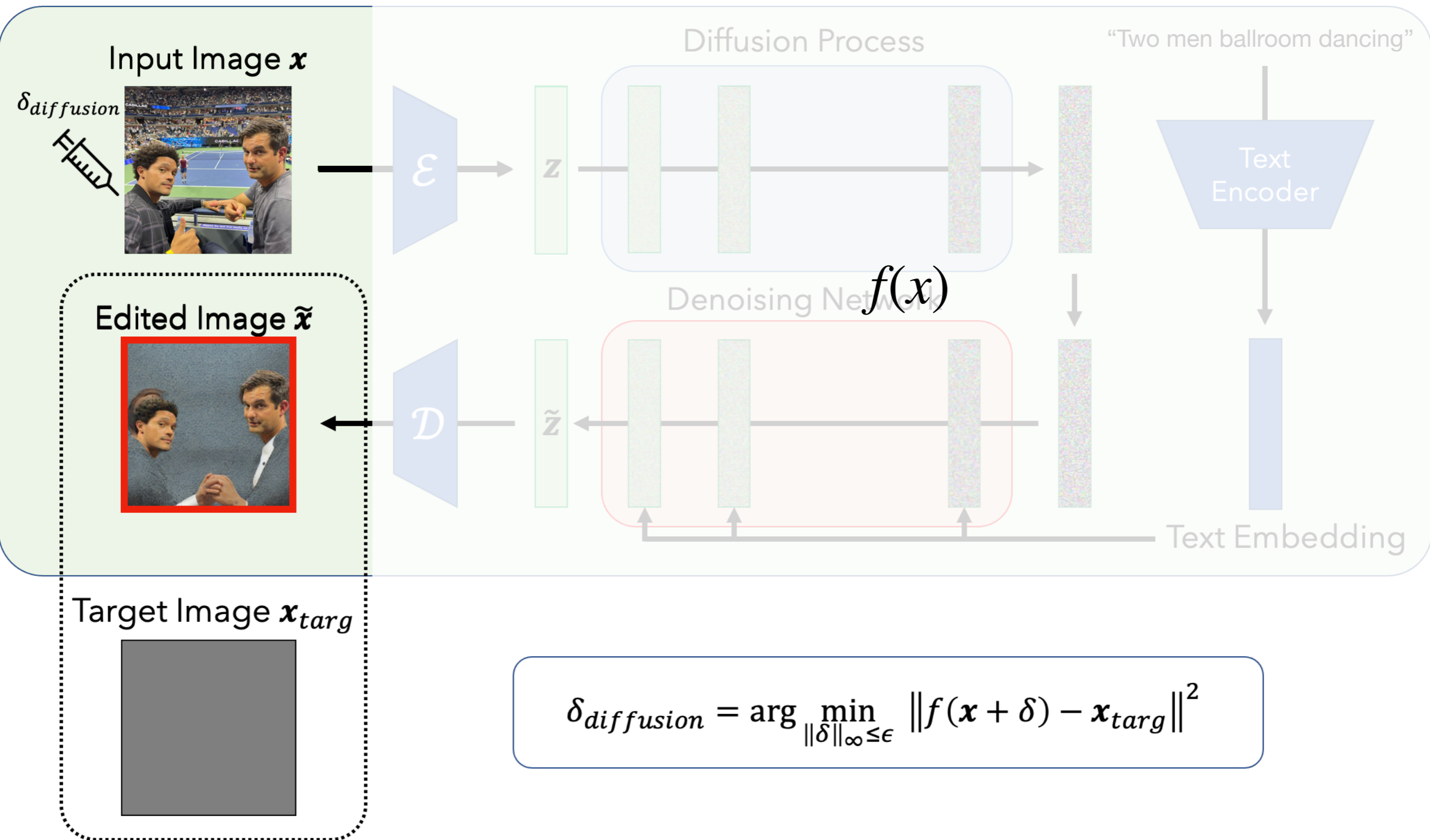
Stronger approach: Diffusion attack



Stronger approach: Diffusion attack



Stronger approach: Diffusion attack



Did it work?

Did it work?

Prompt: "Two men on the plane"



Generated image



Did it work?

Prompt: "Two men on the plane"



Generated image



Generated image
(with immunization)



Success!

Is that a complete solution?

Is that a complete solution?

No: This is rather a proof-of-concept

Is that a complete solution?

No: This is rather a proof-of-concept

A key aspects that needs to be addressed:

Robustness to tampering (and usage of other models)

Is that a complete solution?

No: This is rather a proof-of-concept

A key aspects that needs to be addressed:

Robustness to tampering (and usage of other models)

→ There is robust adv. example work to leverage here

Is that a complete solution?

No: This is rather a proof-of-concept

A key aspects that needs to be addressed:

Robustness to tampering (and usage of other models)

- There is robust adv. example work to leverage here
- **More importantly:** We could (?) have all the (legitimate) model developer be on our side

Is that a complete solution?

No: This is rather a proof-of-concept

A key aspects that needs to be addressed:

Robustness to tampering (and usage of other models)

- There is robust adv. example work to leverage here
- **More importantly:** We could (?) have all the (legitimate) model developer be on our side

In the end: You can "simply" photoshop the image,
but it is about "friction"

Takeaways

Takeaways

→ AI-based malicious image manipulation is serious threat

Takeaways

- AI-based malicious image manipulation is serious threat
- **Our goal:** Make it harder to do malicious manipulations easily and at scale

Takeaways

- AI-based malicious image manipulation is serious threat
- **Our goal:** Make it harder to do malicious manipulations easily and at scale
- We utilize adversarial perturbations to do so

Takeaways

- AI-based malicious image manipulation is serious threat
- **Our goal:** Make it harder to do malicious manipulations easily and at scale
- We utilize adversarial perturbations to do so



@hadiselmanX



gradientscience.org