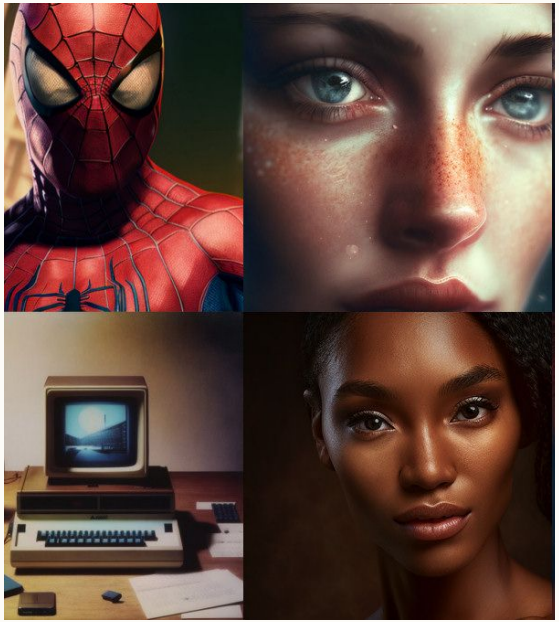
Finding DoRI  : Discovery of Retained Images in Diffusion Models

Antoni Kowalczuk*, Dominik Hintersdorf*, Lukas Struppek*,
Kristian Kersting, Adam Dziedzic, Franziska Boenisch

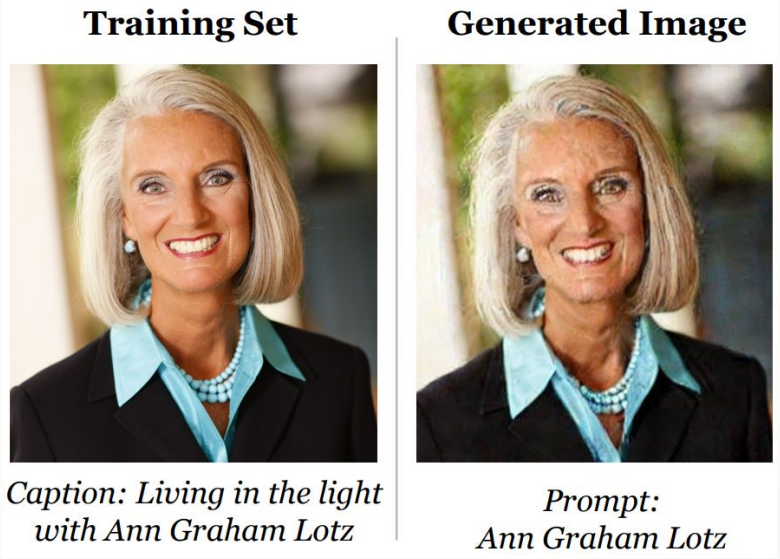
ICML 2026

Diffusion Models (DMs)

High-Quality Images

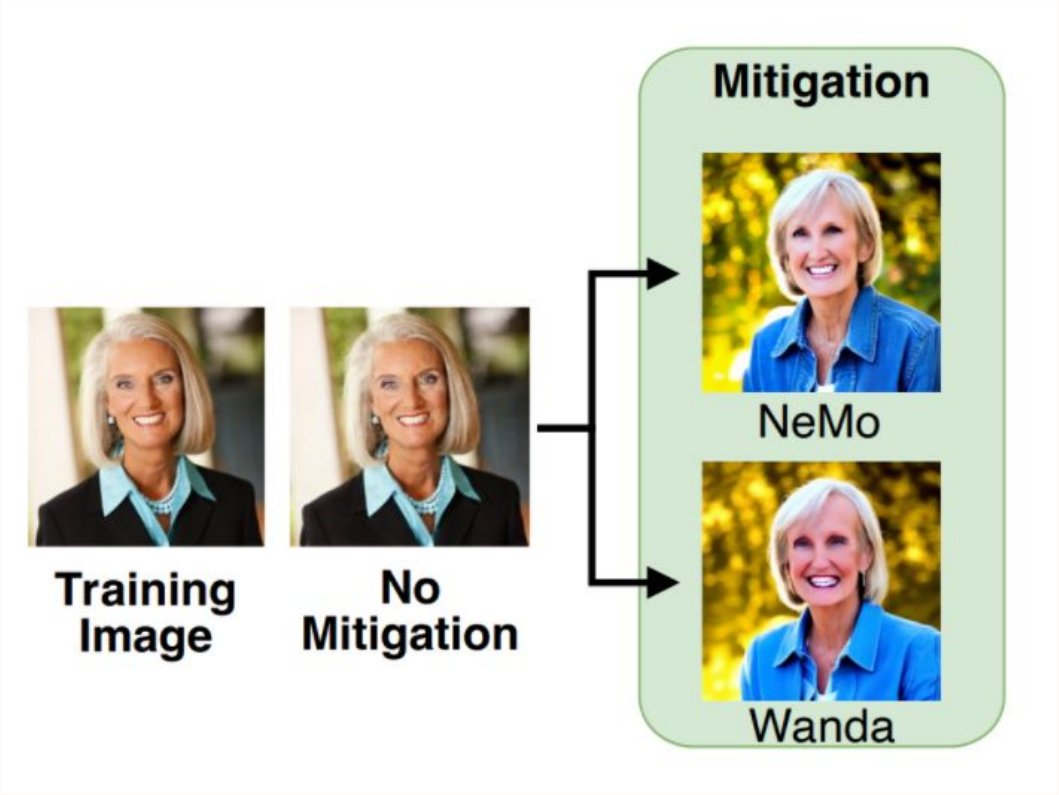


Memorize Training Data

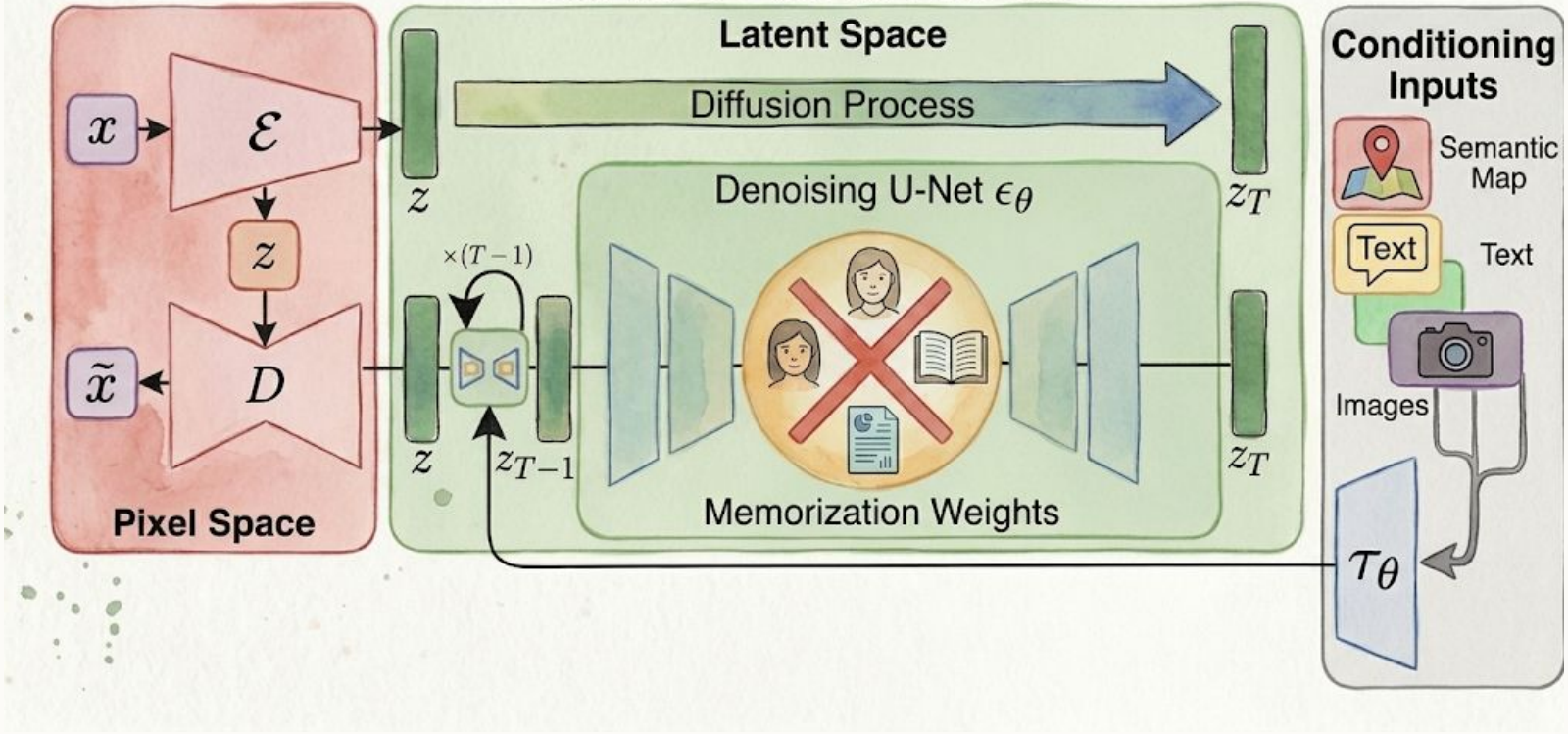


Carlini, N., et al, *"Extracting Training Data from Diffusion Models"* [Usenix 2023]

Mitigation Methods



Pruning


Latent Space
Diffusion Process
Denoising U-Net ϵ_θ
$\times(T-1)$
z
z_T
z_{T-1}
Memorization Weights
Pixel Space
x
$\mathcal{E}$
$\tilde{x}$
D
Conditioning Inputs
Semantic Map
Text
Text
Images
τ_θ

This Work

1. DoRI 
2. Locality Refutation
3. Robust Mitigation

Discovery of Retained Images (DoRI)

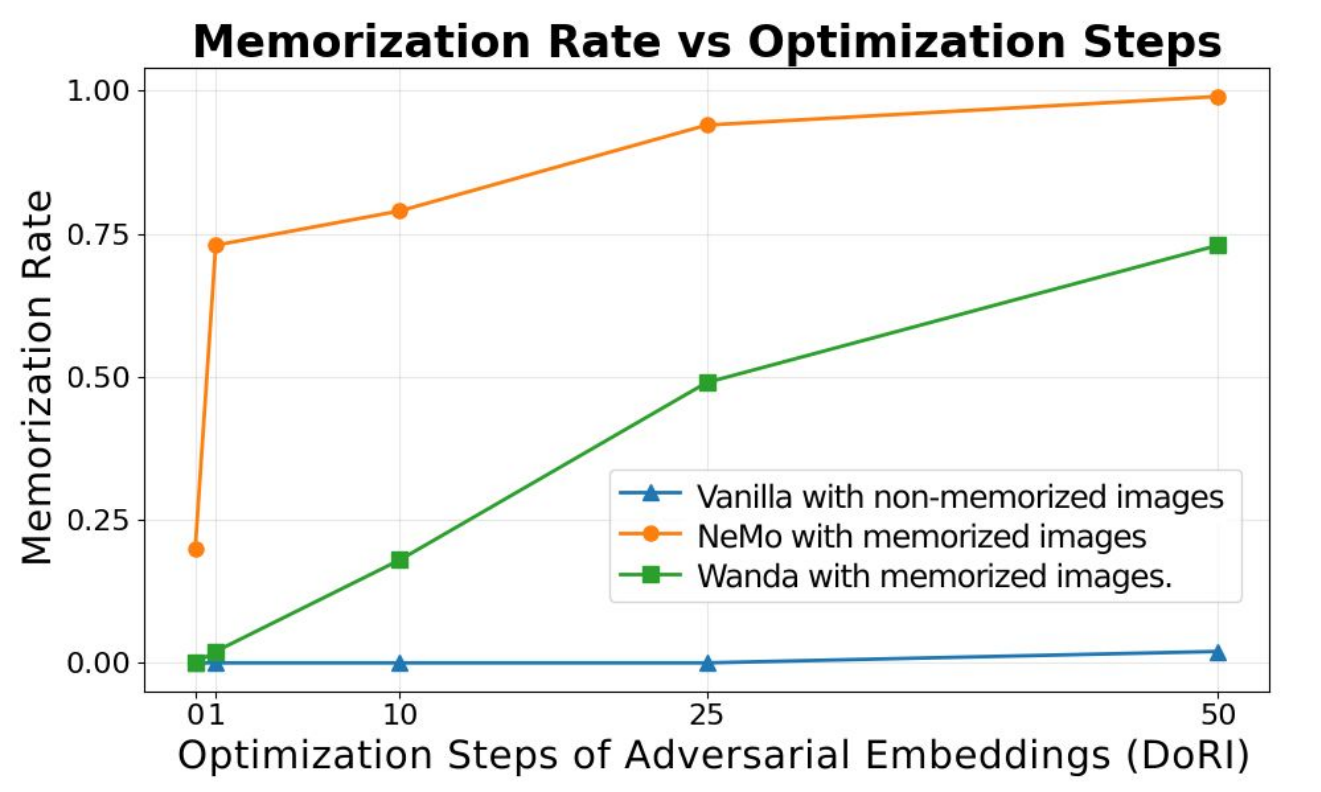
U-Net
Diffusion Model

MSELoss

Text embedding

Backpropagation

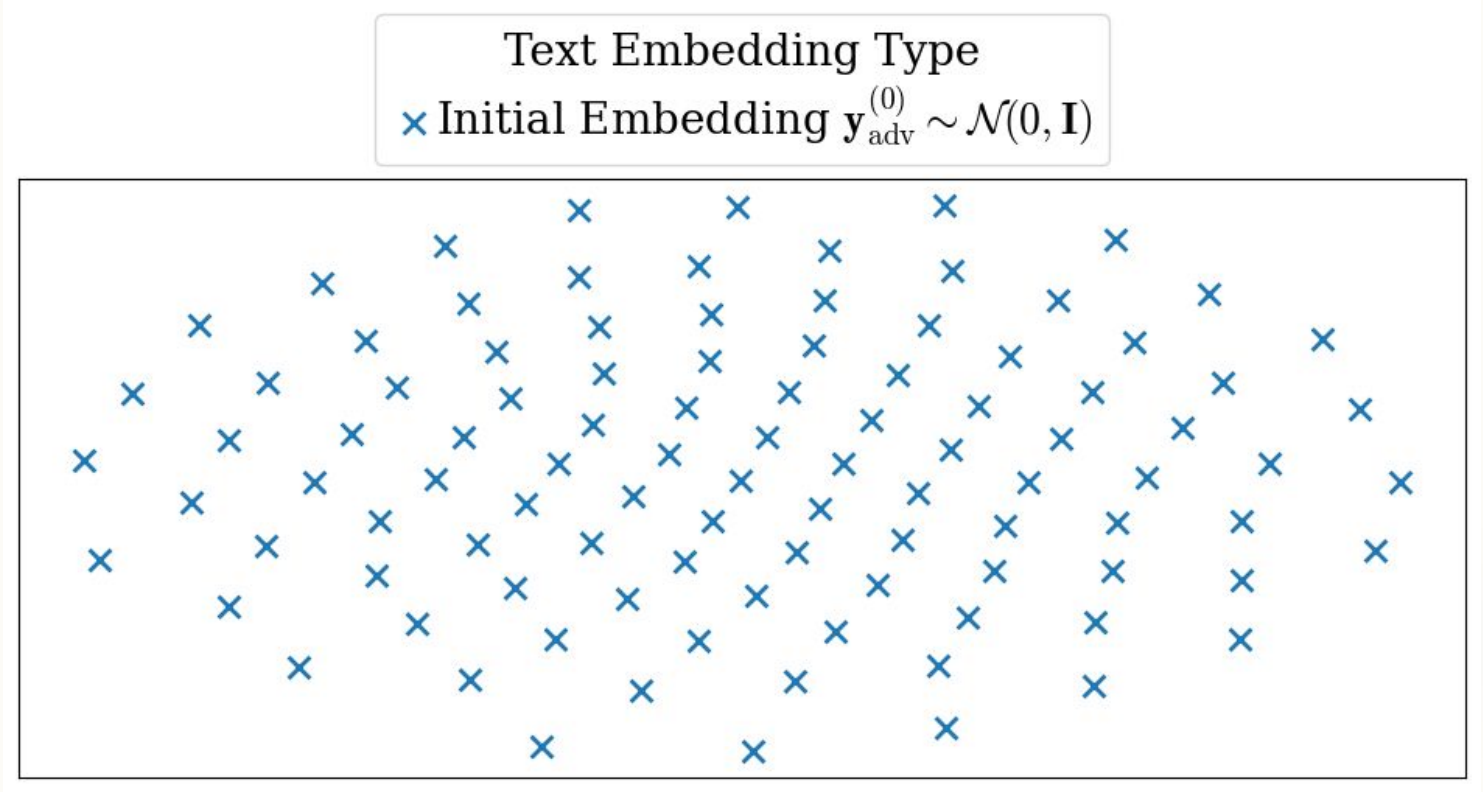
Evading Pruning with DoRI



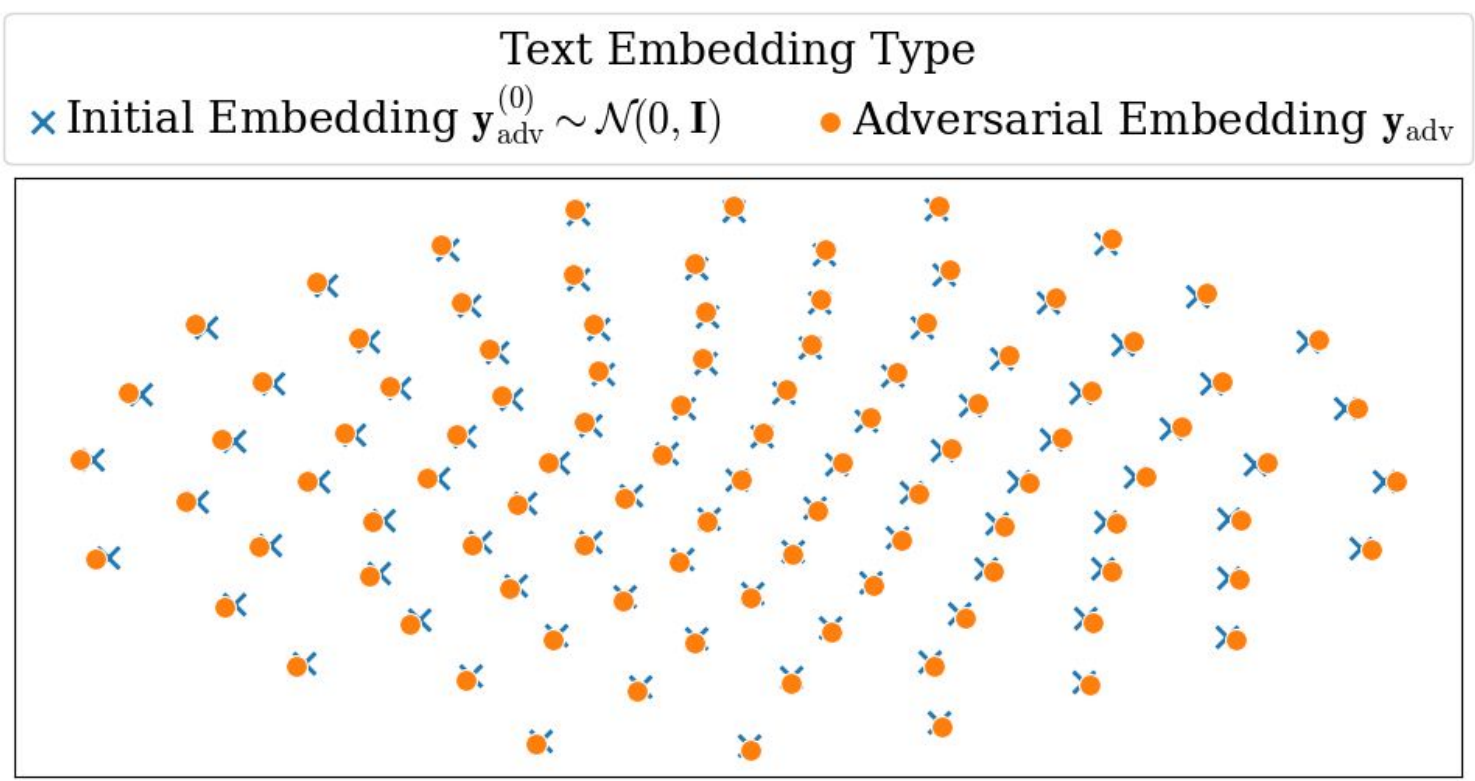
Locality Assumption

1. Input Space
2. Activation Space
3. Weights

Random Initialization



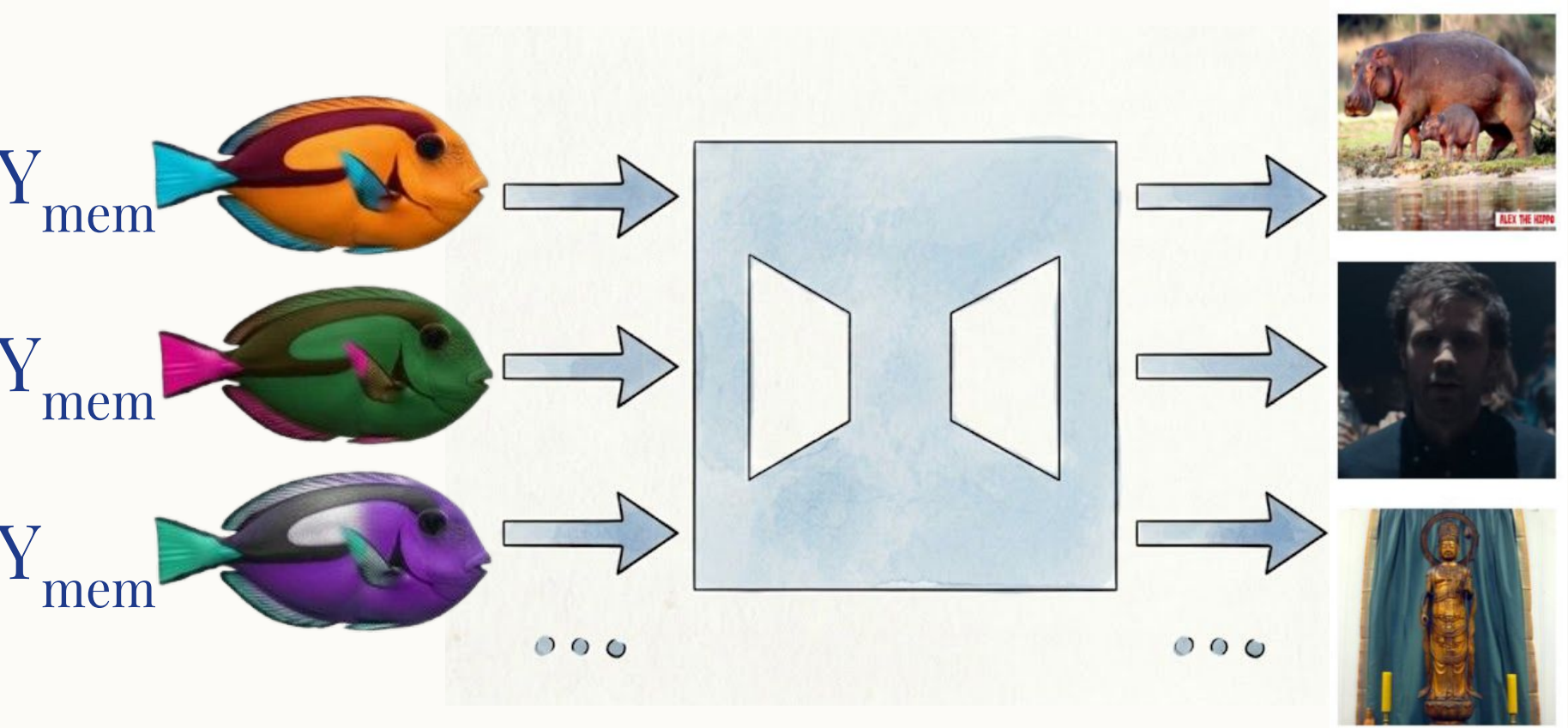
The Same Output Image After DoRI



Locality Assumption

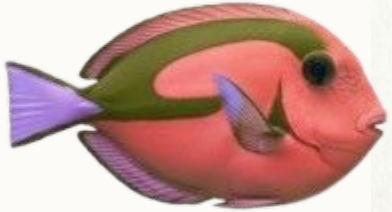
1. ~~Input Space~~ **Refuted**
2. Activation Space
3. Weights

Different Input ↠ Different Output



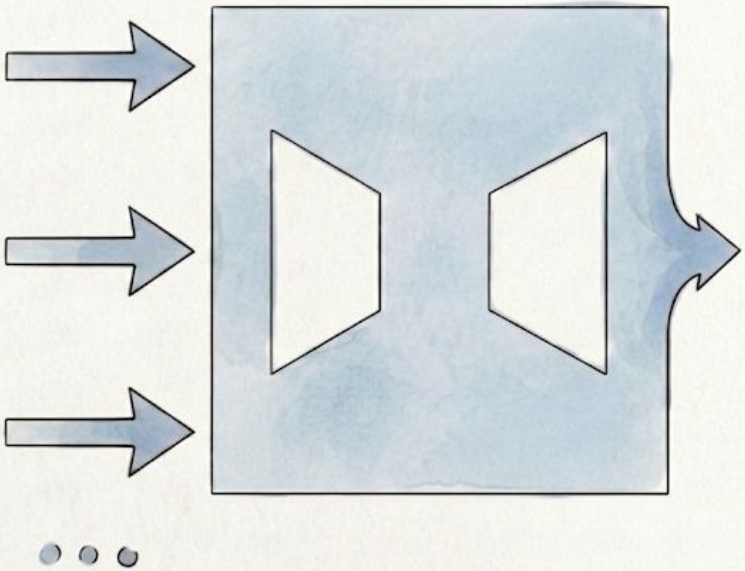
Different Input ↠ Same Output

Y_{adv}



Y_{adv}

Y_{adv}



Expected if Locality Holds:

Discrepancy for Y_{mem}

$>$

Discrepancy for Y_{adv}

Actual:

Discrepancy for Y_{mem}

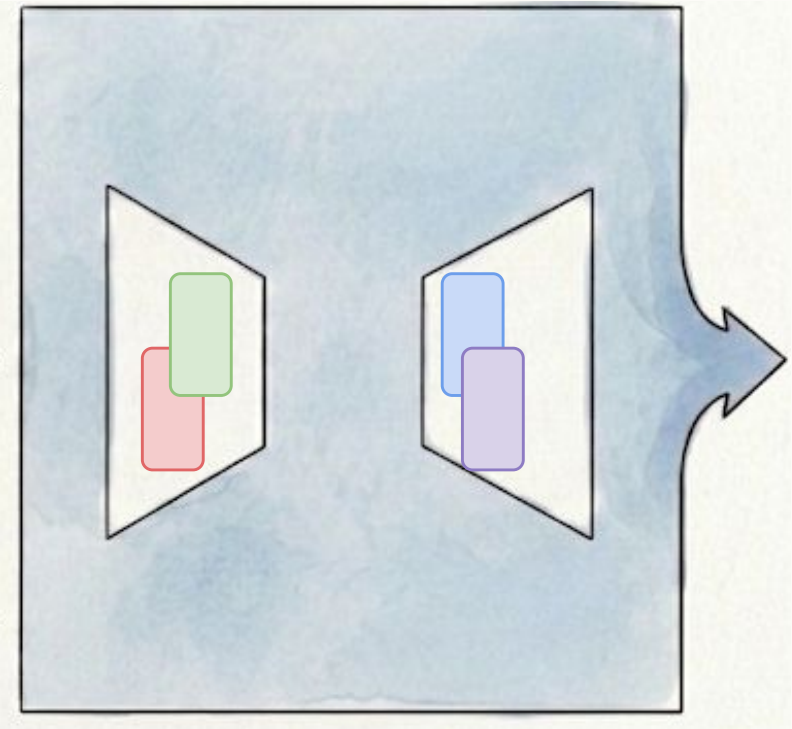
$\approx$

Discrepancy for Y_{adv}

Locality Assumption

1. ~~Input Space~~ **Refuted**
2. ~~Activation Space~~ **Refuted**
3. Weights

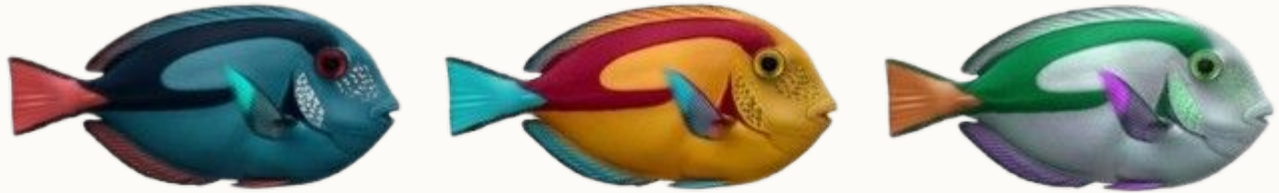
Same Output ↠ Different Weights



~~Locality Assumption~~ **Refuted**

1. ~~Input Space~~ **Refuted**
2. ~~Activation Space~~ **Refuted**
3. ~~Weights~~ **Refuted**

Our Mitigation



Our Mitigation

MSELoss

U-Net
Diffusion Model

Backpropagation

Our Mitigation is Successful

	Prompt	
No Mitigation	1.00 MR	**1.00 MR**
Our Mitigation	0.00 MR	**0.02 MR**

This Work

1. DoRI 
2. Locality Refutation
3. Robust Mitigation