# Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models

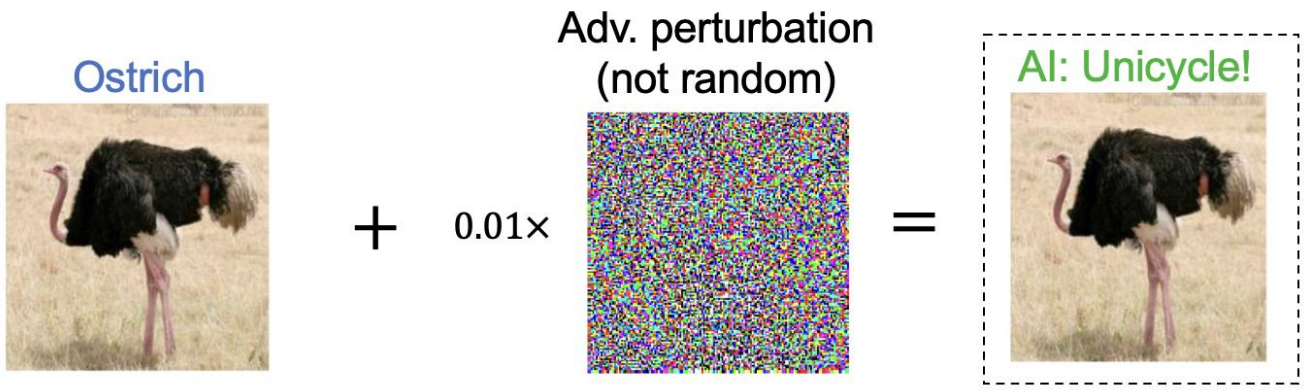

**Yimeng Zhang**[1] **Xin Chen**[2] **Jinghan Jia**[1] **Yihua Zhang**[1] **Chongyu Fan**[1]

**Jiancheng Liu**[1] **Mingyi Hong**[3] **Ke Ding**[2] **Sijia Liu**[1,4]

[1]Michigan State University [2]Applied ML, Intel
[3]University of Minnesota, Twin City [4]MIT-IBM Watson AI Lab, IBM Research

# Worst-Case Unlearning Evaluation: An Adversarial Attack Lens

- **Adversarial example: Provides robustness evaluation for ML models** [Goodfellow, et al., 2015]

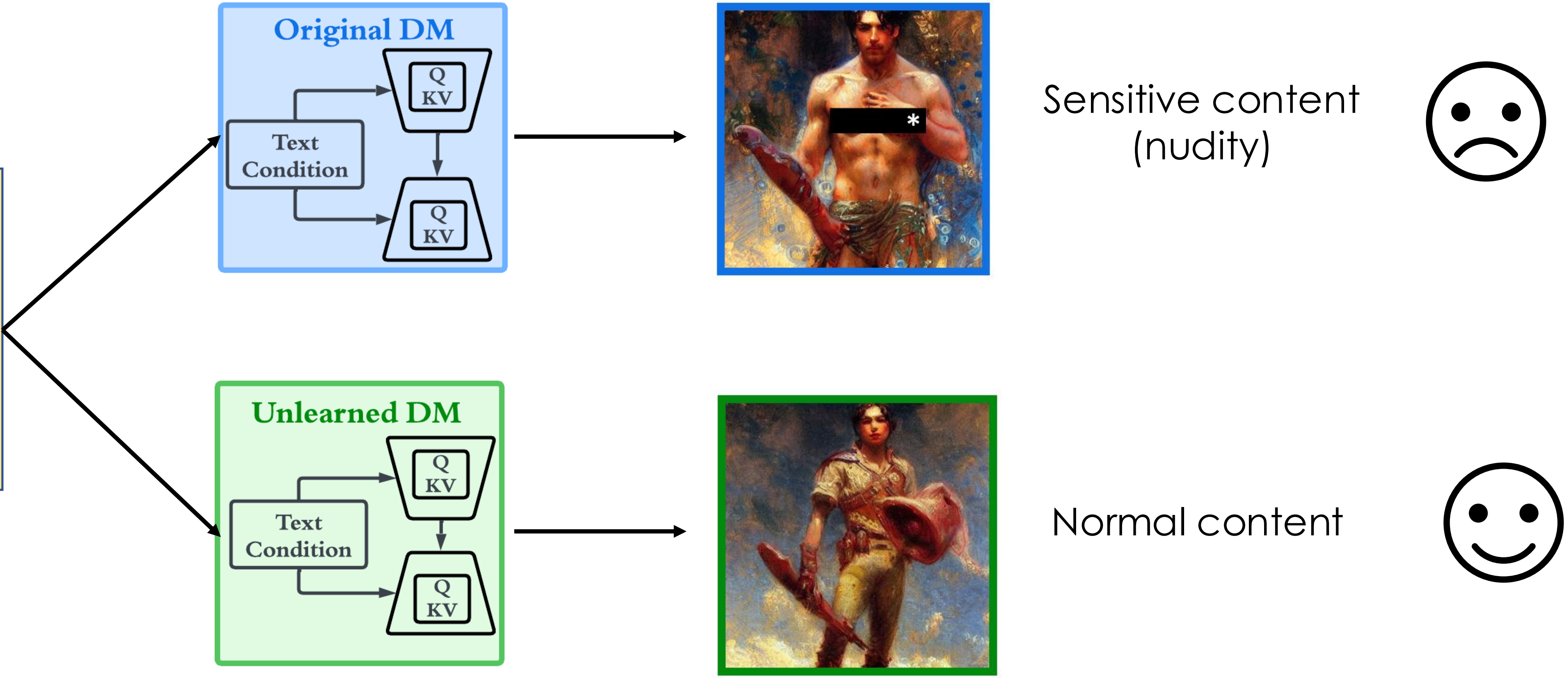



**Given an ML model post unlearning, can we jailbreak it to reverse engineer the forgotten information?**

Goodfellow, et al.. "Explaining and harnessing adversarial examples." *ICLR'15*

# Motivating Example

**Input prompt**
attractive male, character design, painting by gaston bussiere

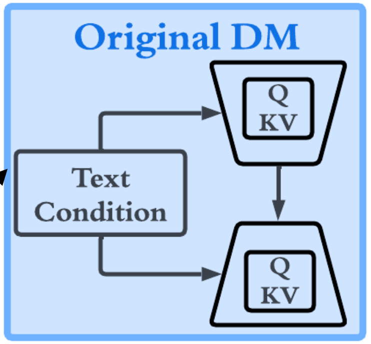


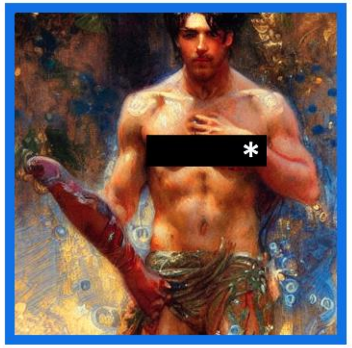
# Motivating Example

Adversarial input perturbation

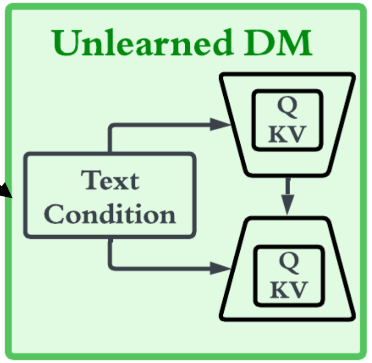
**Adversarial prompt (perturbed input prompt)**

ca vh ashish nude finnish attractive male, character design, painting by gaston bussiere



Sensitive content (nudity)




Sensitive content (nudity)

# Arm Race Between Attacker and Defender in Machine Unlearning

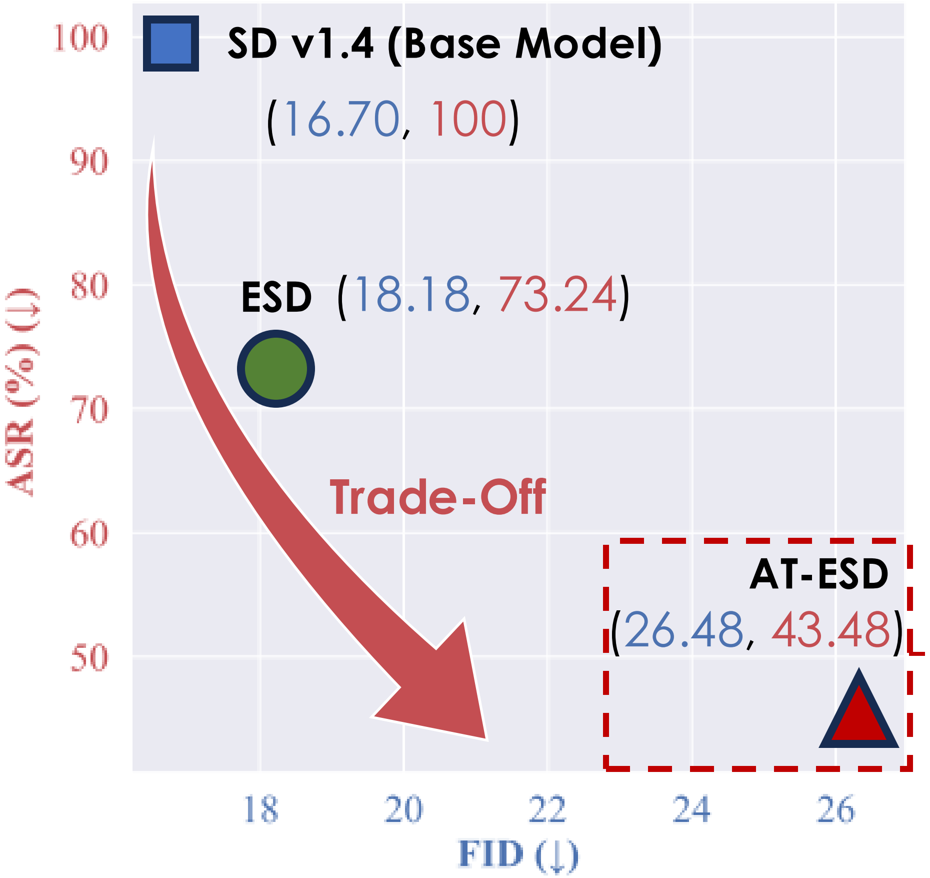

*Question:*

Can we boost the robustness of MU against adversarial attacks?

*Answer:*

Adversarial Unlearning: Integrating <u>adversarial training</u> into unlearning for robustness enhancement

# Directly Utilize Adversarial Training for Diffusion Model Unlearning *Destroy* Model Utility




Fail to generate target objects


Zhang, et al. "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now." ECCV, 2024

# Challenges

- **(Effectiveness challenge)**
  optimizing the inherent trade-off between the robustness of concept erasure and the preservation of DM utility poses a significant challenge.

- **(Efficiency challenge)**
  deciding 'where' to apply AT within DM

# (Effectiveness Challenge) trade-off between the erasure robustness and the utility preservation

Generating adversarial prompts

$$c^* = \underset{\|c' - c_\mathrm{e}\|_0 \le \epsilon}{\arg\min} \; \ell_\mathrm{atk}(\boldsymbol{\theta}, c')$$

$$\ell_\mathrm{u}(\boldsymbol{\theta}, c^*) = \ell_\mathrm{ESD}(\boldsymbol{\theta}, c^*) + \gamma \mathbb{E}_{\tilde{c} \sim \mathcal{C}_\mathrm{retain}} \left[ \| \epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t | \tilde{c}) - \epsilon_{\boldsymbol{\theta}_\mathrm{o}}(\mathbf{x}_t | \tilde{c}) \|_2^2 \right]$$

Utility-retaining regularization

Retain Set $C_{retain}$

retain prompts from an external dataset (*ImageNet or COCO*),

using the prompt template ‘a photo of [OBJECT CLASS]’.

# (Efficiency Challenge)
# Where to robustify: Text encoder or UNet?

| | Text Encoder 🤩 | UNet 🤦‍♀️ |
|---|---|---|
| # Parameter | 63M | 859M |

text encoder is easier to be finetuned

| DMs | Optimized DM component | ASR (↓) | FID (↓) |
|---|---|---|---|
| SD v1.4 | N/A | 100% | 16.70 |
| ESD | UNet | 73.24% | 18.18 |
| ESD | Text Encoder | 3.52% | 59.10 |
| AdvUnlearn | UNet | 64.79% | 19.88 |
| AdvUnlearn | Text Encoder | 21.13% | 19.34 |

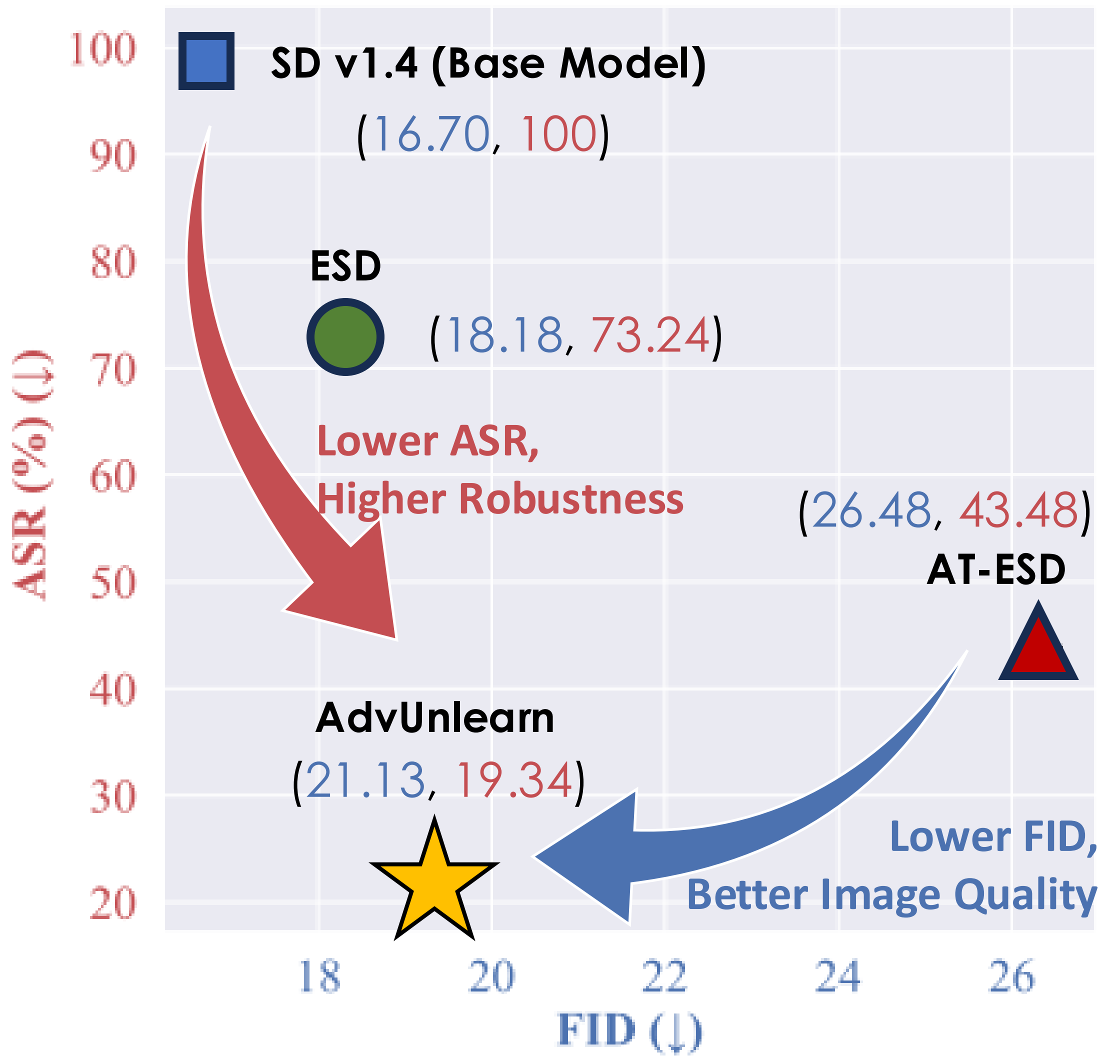
Less trade-off during robustifying text encoder

- Great unlearning robustness
- Minor model utility drop

**Text encoder** ✅

**Unet** ❌

# AdvUnlearn





Zhang, et al. "To generate or not? safety-driven unlearned diffusion models are still easy to generate unsafe images... for now." ECCV, 2024