WACV 2026
TUCSON, AZ 3/6 - 3/10

# M-ErasureBench: A Comprehensive Multimodal Evaluation Benchmark for Concept Erasure in Diffusion Models





Ju-Hsuan Weng[1,2] **Jia-Wei Liao**[1,2] Cheng-Fu Chou[1] Jun-Cheng Chen[2]

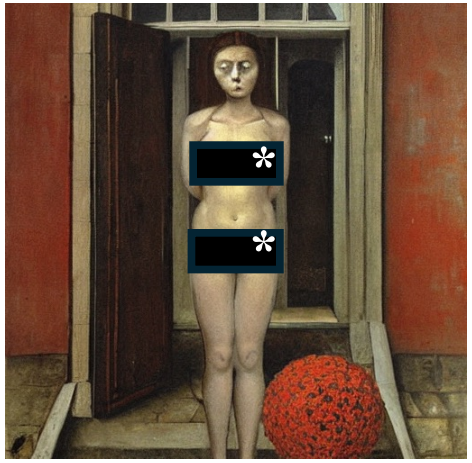
[1] National Taiwan University,
[2] Research Center for Information Technology Innovation, Academia Sinica

{r12922a05, d11922016, ccf}@csie.ntu.edu.tw, pullpull@citi.sinica.edu.tw

# Motivation

Diffusion models could generate harmful or copyrighted content with inappropriate text prompts, and concept erasure methods aim to mitigate this issue by suppressing specific concepts.

***Prompt:*** *A woman standing in the doorway.*



Stable Diffusion

AdvUnlearn [1]

[1] Zhang et al. “Defensive Unlearning with Adversarial Training for Robust Concept Erasure in Diffusion Models”. *NeurIPS 2024*.

# Motivation

Current red-teaming methods for concept erasure mainly focus on text prompts and thus lack a robust and comprehensive evaluation.

**Research Question**

*How robust are concept erasure methods across different input modalities, and can their vulnerabilities be mitigated without retraining?*

# M-ErasureBench: Multimodal Evaluation Framework

## Step 1: Multimodal Input

### a) Prompts (Generation)

"airplane" or 😈

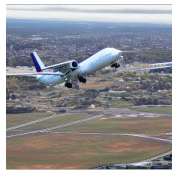
Text Prompt | Adversarial Prompt

### b) Learned Embedding (Personalization)

Textural Inversion

Reference | Learned Embedding

### c) Latent Inversion (Editing)



DDIM Inversion

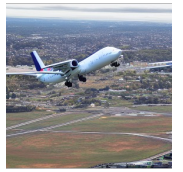
Reference | Inverted Latent

## Step 2: DDIM Sampling

### Erased Methods

☹ "airplane" not removed

Prompt, Latent → Erased Model →



Output

### Erased Methods + *IRECE*

😃 "airplane" removed

Prompt, Latent → Erased Model + *IRECE* →

Output

# Experimental Setup

## Dataset

1. **SD-Normal:** Generated from Stable Diffusion using five prompt templates, 150 prompts per class.
2. **SD-AdvPrompt:** Adversarial prompts produced using Ring-A-Bell [2].
3. **SD-TI:** Textual Inversion embeddings trained for each reference image.
4. **SD-LatentInv:** DDIM-inverted latents from reference images, combined with various prompt strategies.

## Concept Reproduction Rate (CRR)

- How often an erased concept reappears in generated images, detected using GroundingDINO.

[1] Tsai et al. “Ring-A-Bell! How Reliable are Concept Removal Methods for Diffusion Models?”. *ICLR 2024*.

# Experimental Setup

## Model

- **Base model:** Stable Diffusion (SD) v1.4.
- **Erased model:** ESD [3], UCE [4], Receler [5].

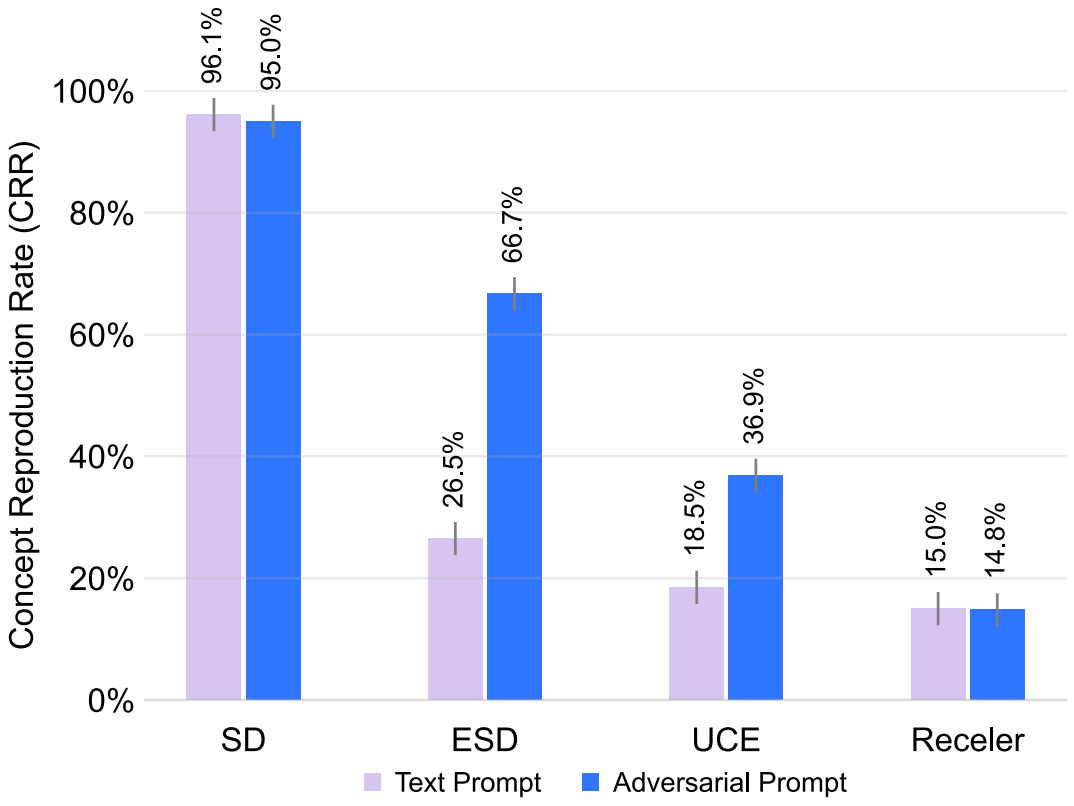
## Evaluation Settings

- **White-Box:** Both step 1 and step 2 use the erased model.
- **Black-Box:** Step 1 uses the base model, while step 2 uses the erased model.

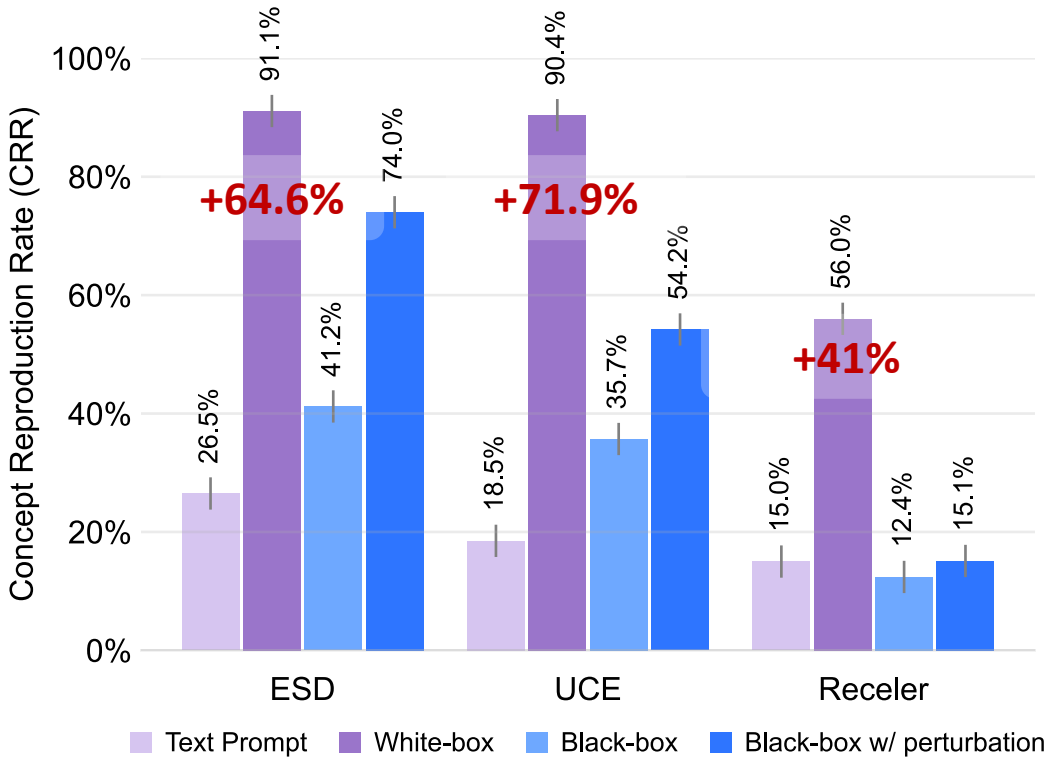
[3] Gandikota et al. "Erasing Concepts from Diffusion Models". *ICCV 2023*.
[4] Gandikota et al. Unified Concept Editing in Diffusion Models. *WACV 2024*.
[5] Huang et al. Receler: Reliable Concept Erasing of Text-to-image Diffusion Models via Lightweight Erasers. *ECCV 2024*.

# Evaluation Results

## (a) Text / Adv. Prompt

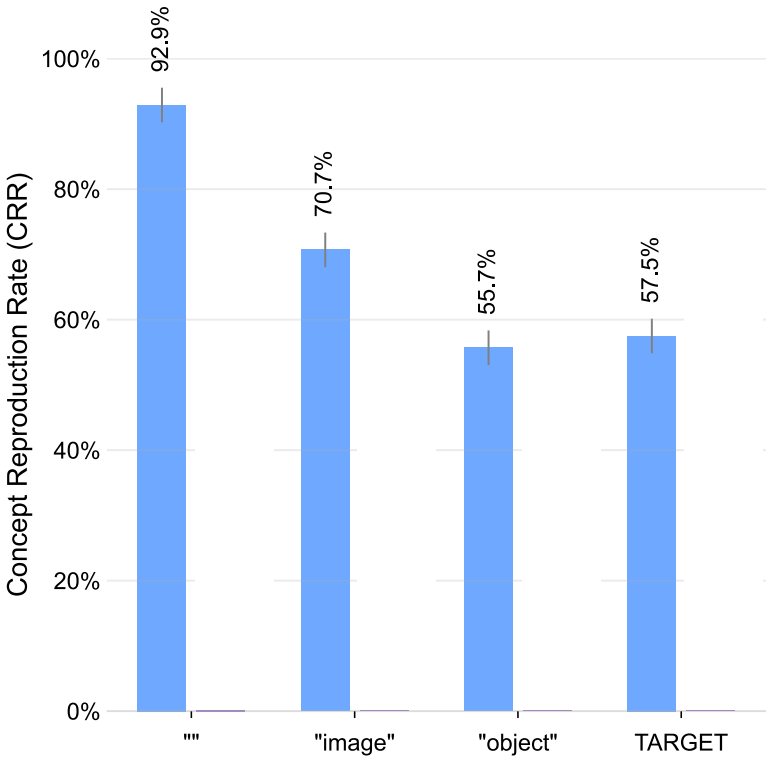



## (b) Learned Embedding

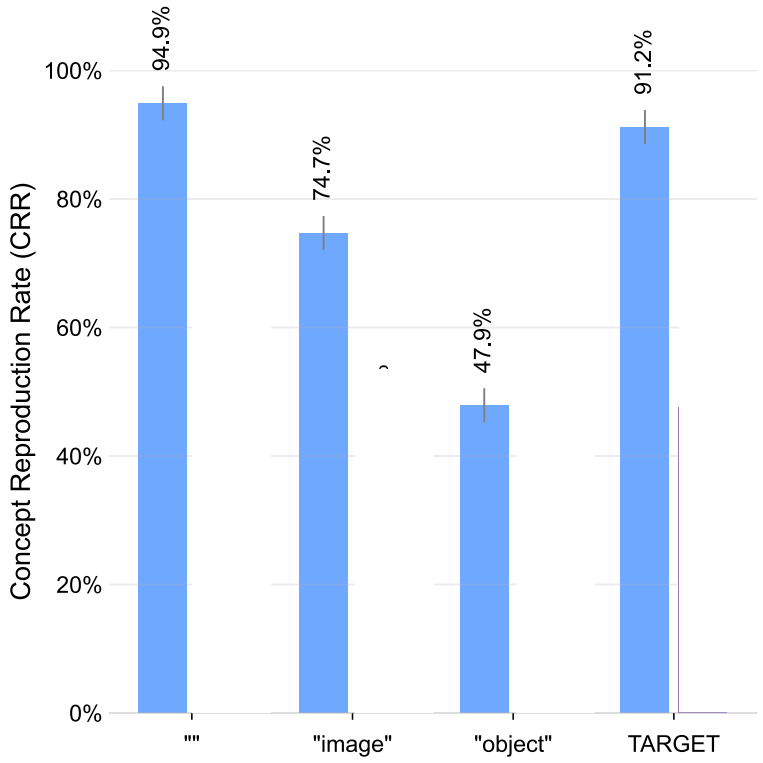


# Evaluation Results: White-Box

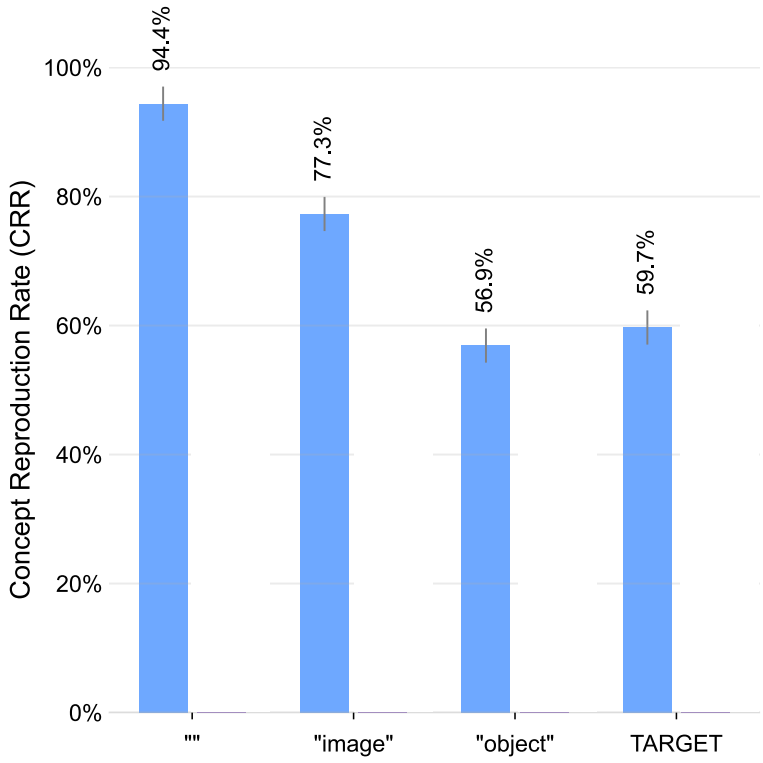
## (c) Latent Inversion




ESD



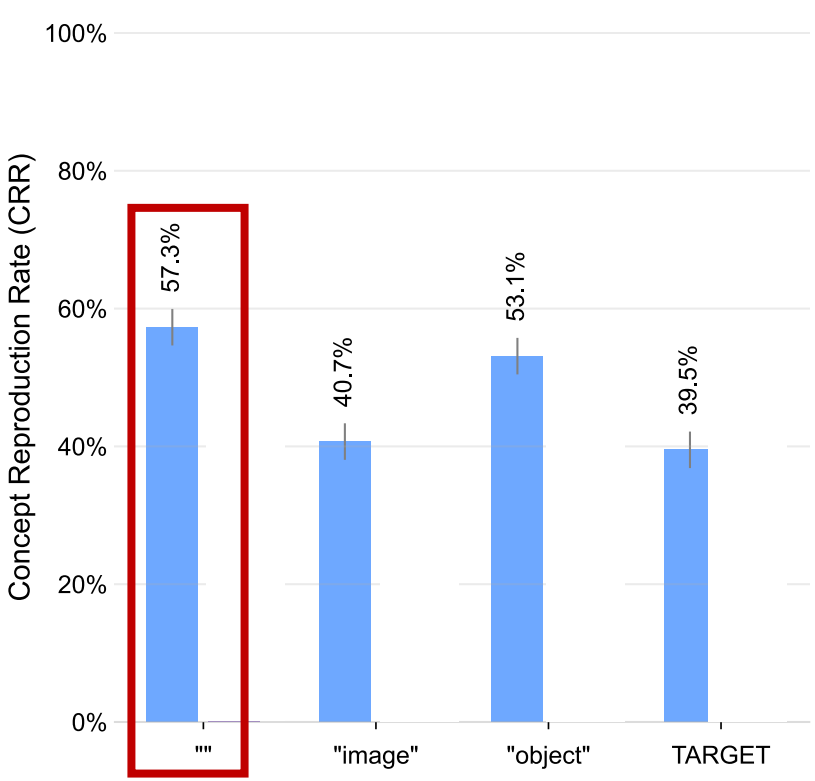

UCE



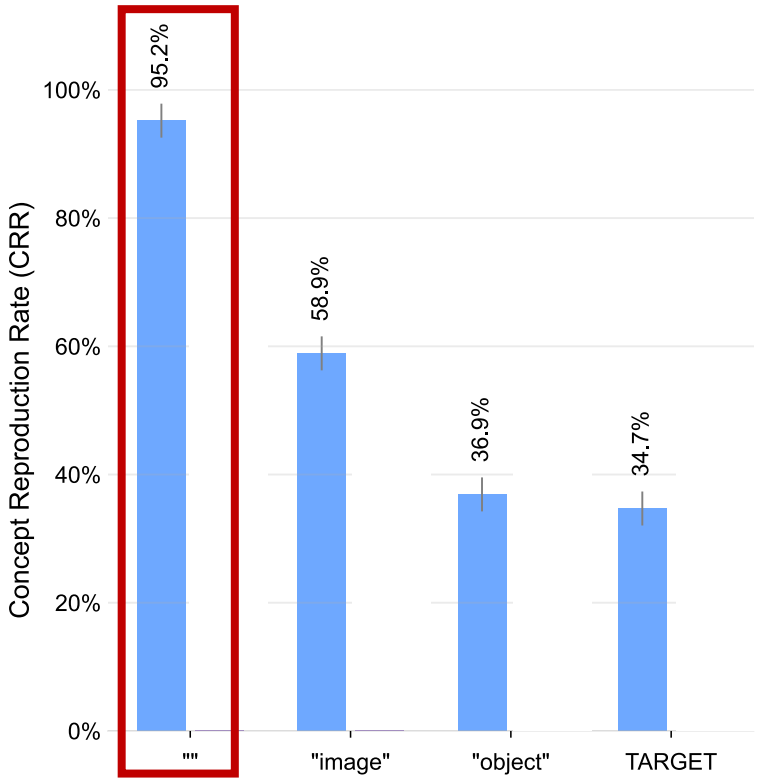

Receler

# Evaluation Results: Black-Box

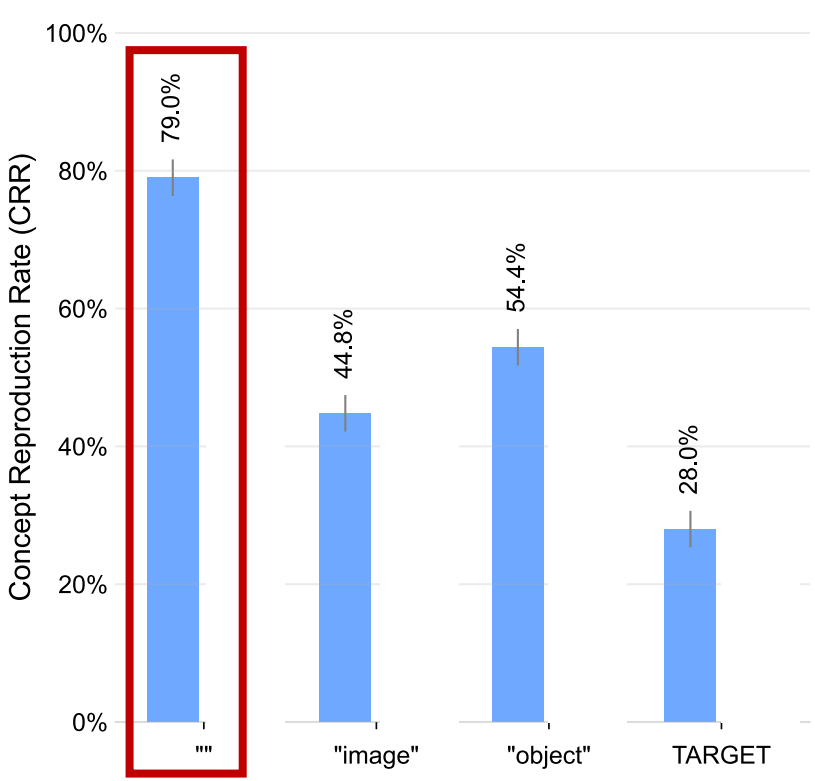
## (c) Latent Inversion




ESD




UCE




Receler

# Qualitative Results

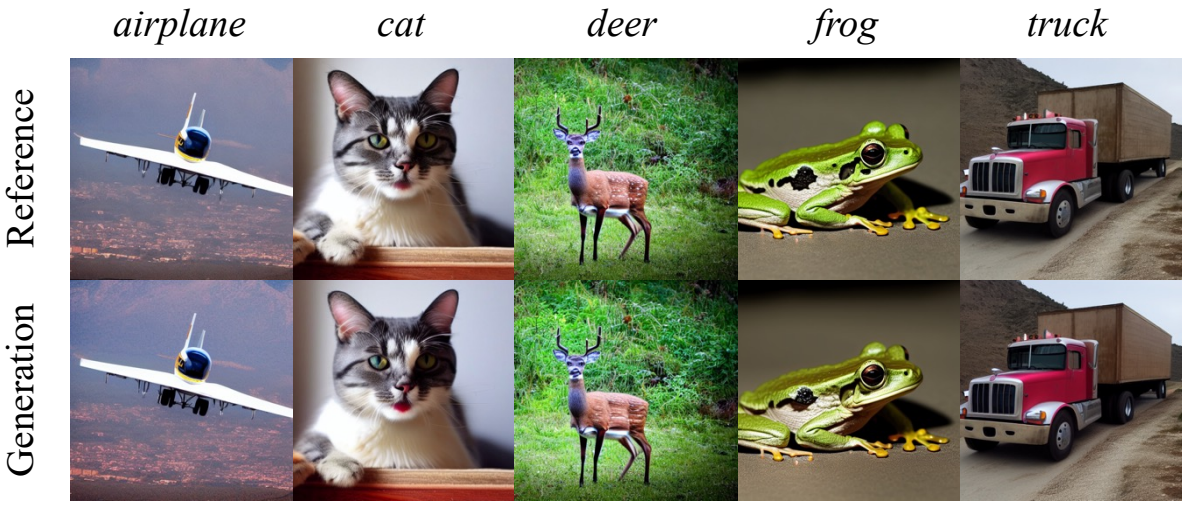



**Figure.** Qualitative results of generated images from concept erased diffusion models under the **black-box** setting with perturbed reference images in the **learned embedding evaluation**.

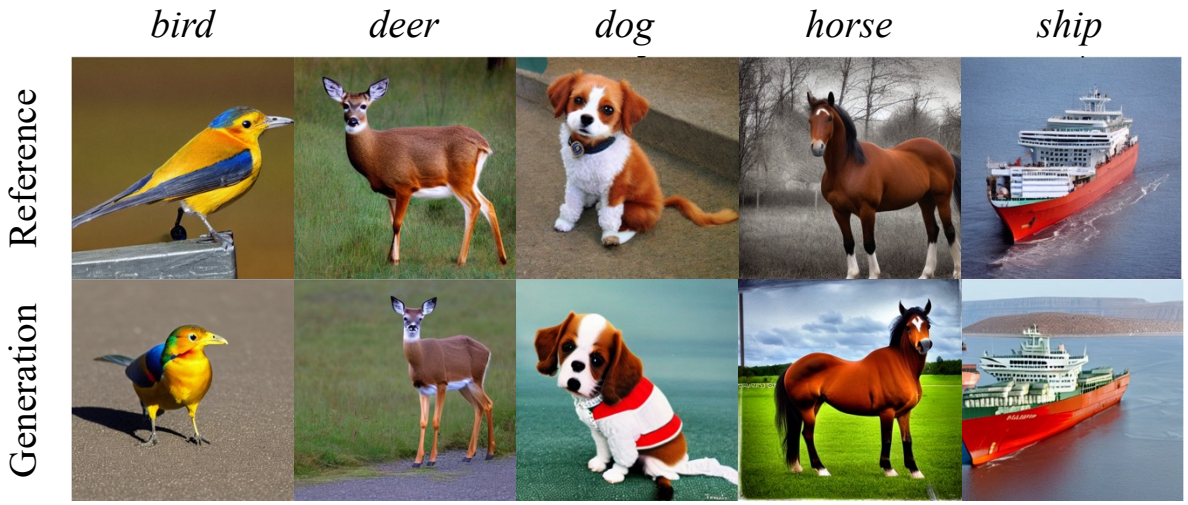
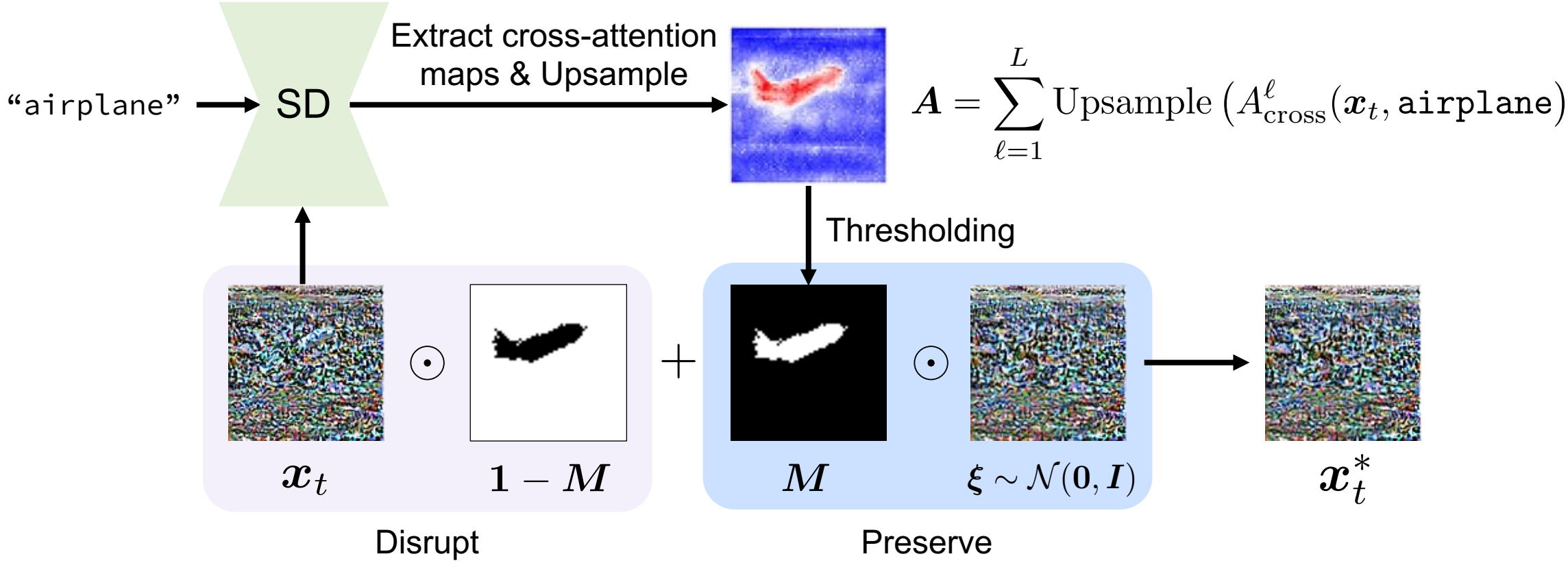



**Figure.** Qualitative results for generated images from concept erased diffusion models with **unconditional prompt** under the **black-box** **latent inversion evaluation**.

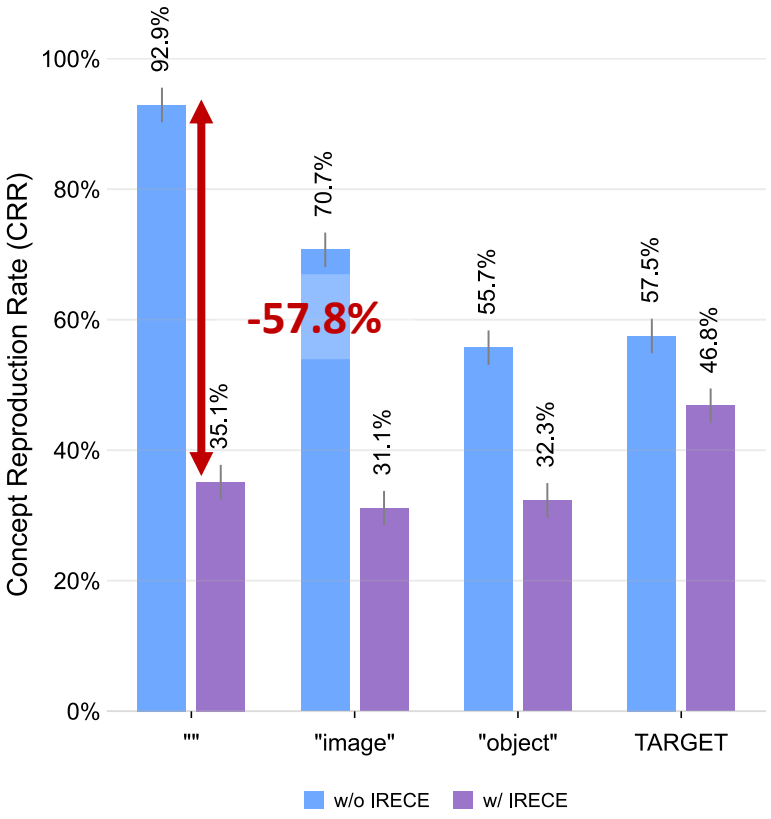
# Inference-Time Robustness Enhancement (IRECE)

Concept erasure fails because concepts persist in the latent space; IRECE resolves this by surgically removing concept-bearing regions during inference

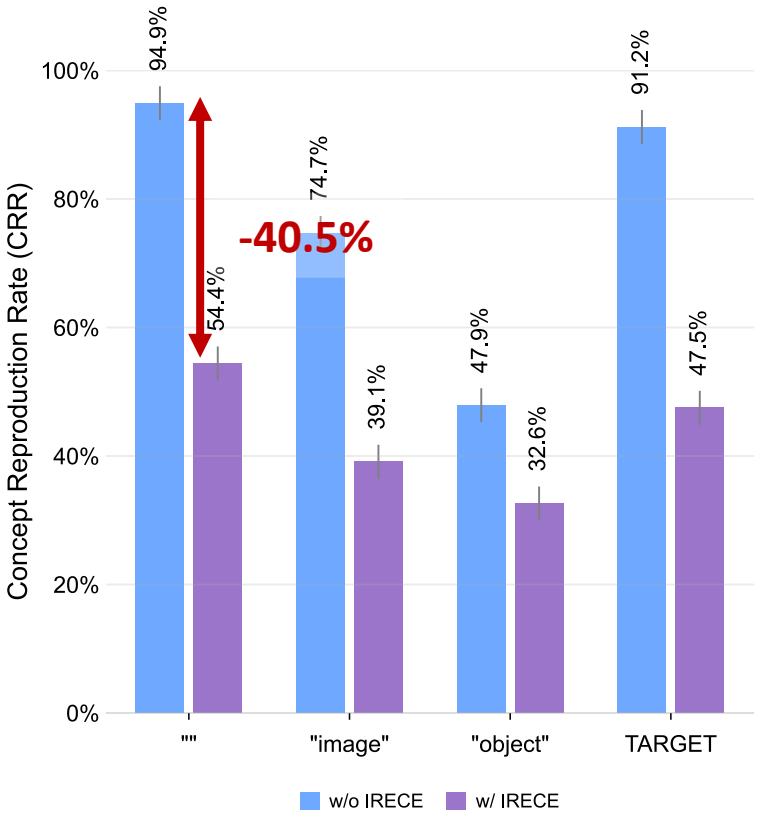


# Evaluation Results: White-Box

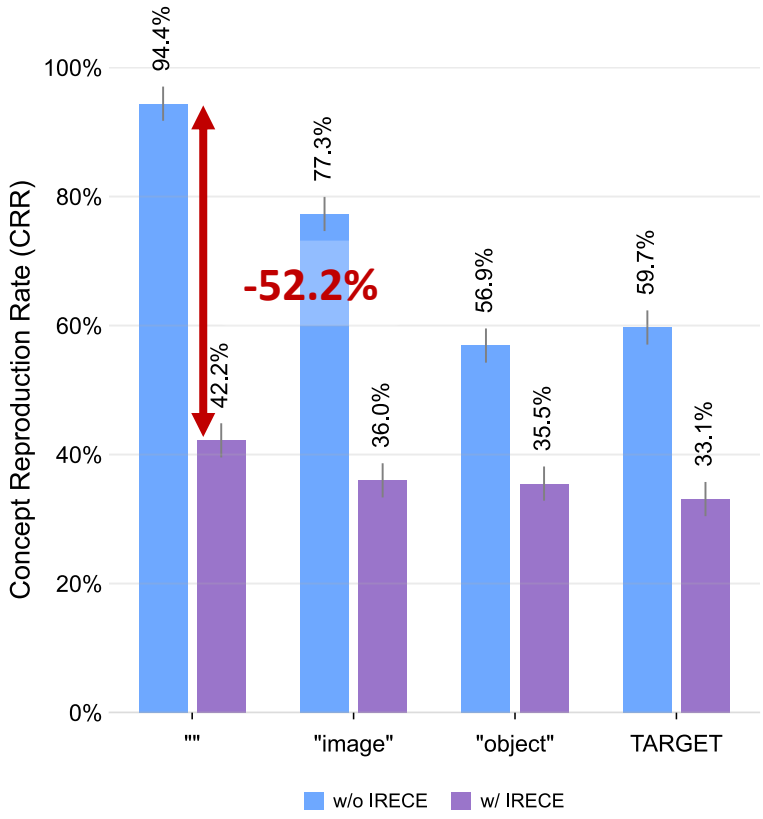
## (c) Latent Inversion




ESD



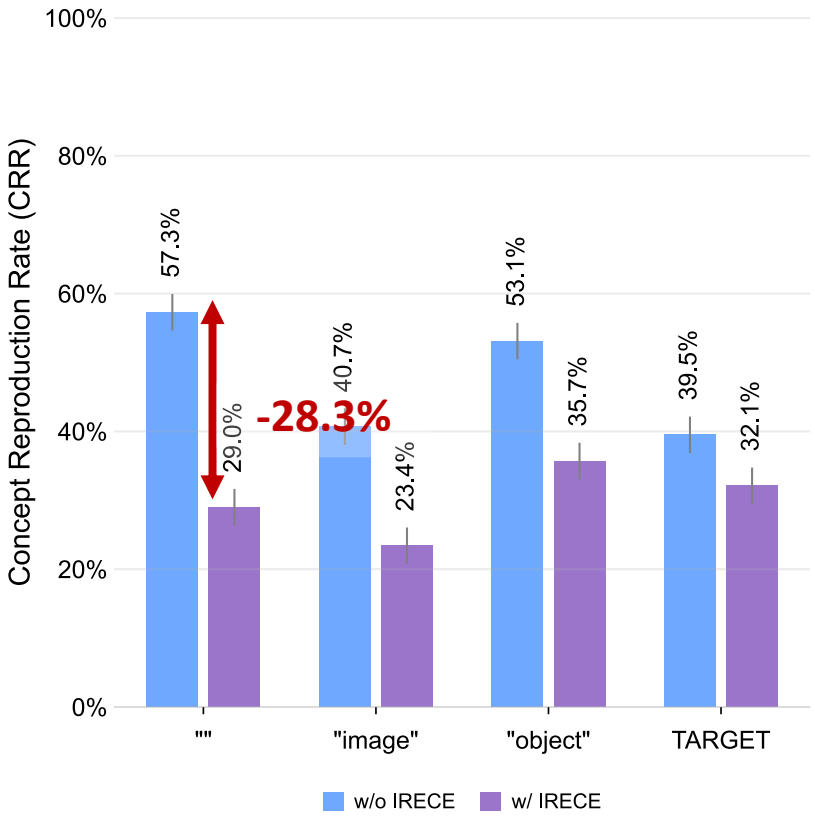

UCE



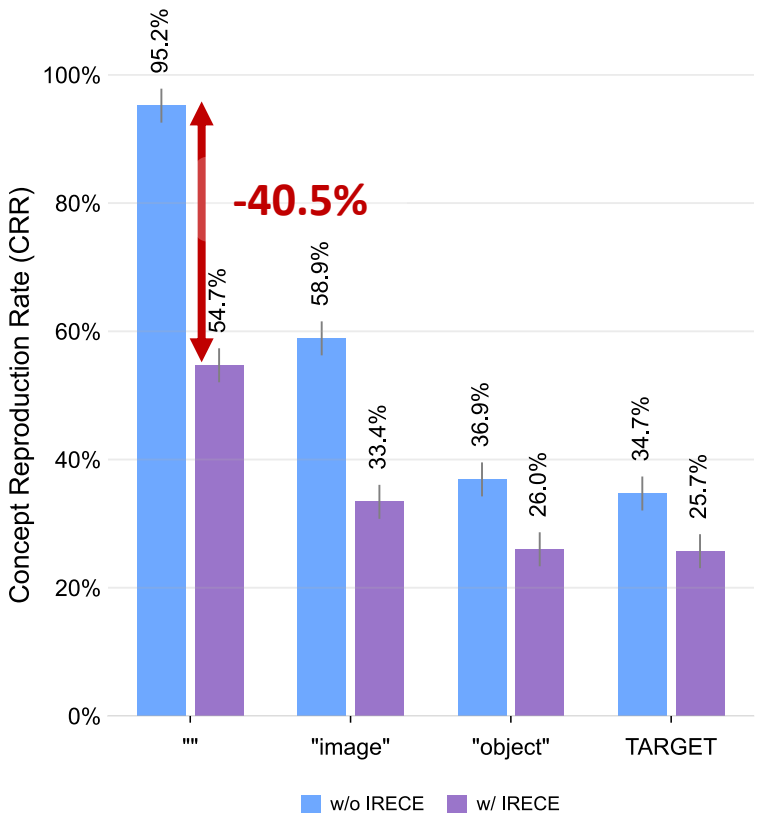

Receler

# Evaluation Results: Black-Box

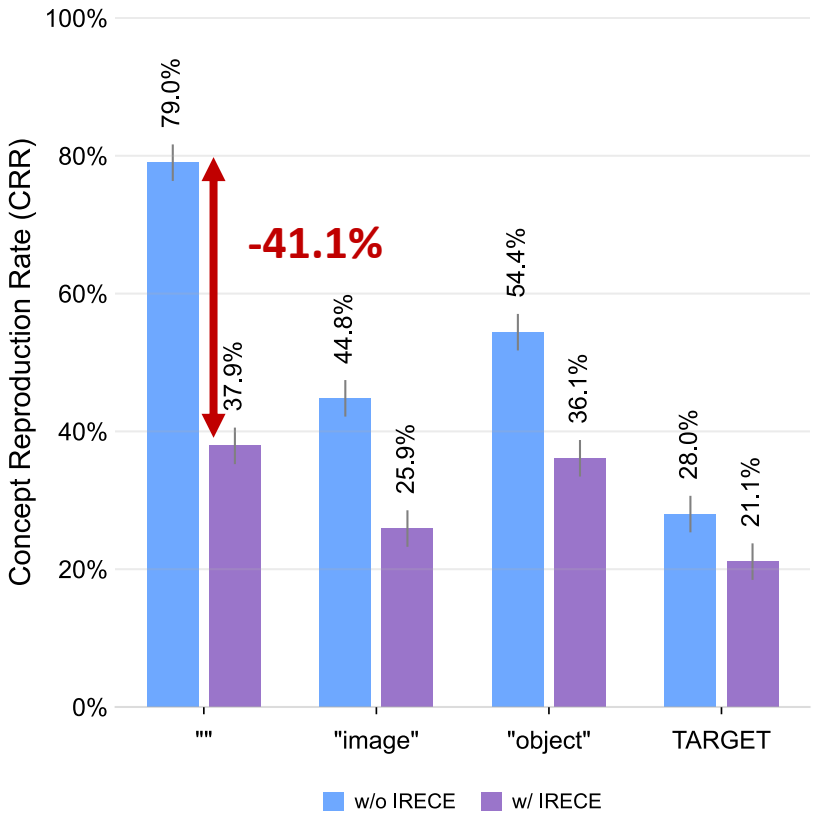
## (c) Latent Inversion




ESD




UCE




Receler

# Qualitative Results for IRCEC


**Figure.** Comparison of erased models with IRECE across 10 concepts under **white-box** **latent inversion**. IRECE effectively removes the target concept while preserving the rest of the image.

# Summary

- We introduce ***M-ErasureBench***, a comprehensive multimodal benchmark that systematically evaluates concept erasure across **text prompts**, **learned embeddings**, and **latent inversion** under both white-box and black-box settings.
- Our study reveals that existing concept erasure methods mainly disrupt text–image alignment rather than fully removing concepts, and they largely fail under **learned embeddings** and **latent inversion**, with CRR exceeding 90% in white-box scenarios.
- To address these limitations, we propose *IRECE*, a plug-and-play inference-time module that localizes target concepts via cross-attention and perturbs corresponding latent regions, reducing CRR by up to ~40% without retraining while preserving visual quality.

# Thanks for listening!