



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

Ju-Hsuan Weng^{1,2*}, Jia-Wei Liao^{1,2*}, Cheng-Fu Chou¹, Jun-Cheng Chen²

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



M-EraseBench: Multimodal Evaluation Framework

Diffusion models generate harmful or copyrighted content, and erasure methods aim to suppress specific concepts. However, existing evaluations assume text prompts only. Our benchmark tests concept erasure robustness under three modalities:

- Text prompts:** basic & adversarial prompts.
- Learned embeddings:** Textual Inversion under white-/black-box settings.
- Latent inversion:** DDIM-based image-to-latent initialization under white-/black-box settings.

Does concept erasure “break” in multimodal settings?

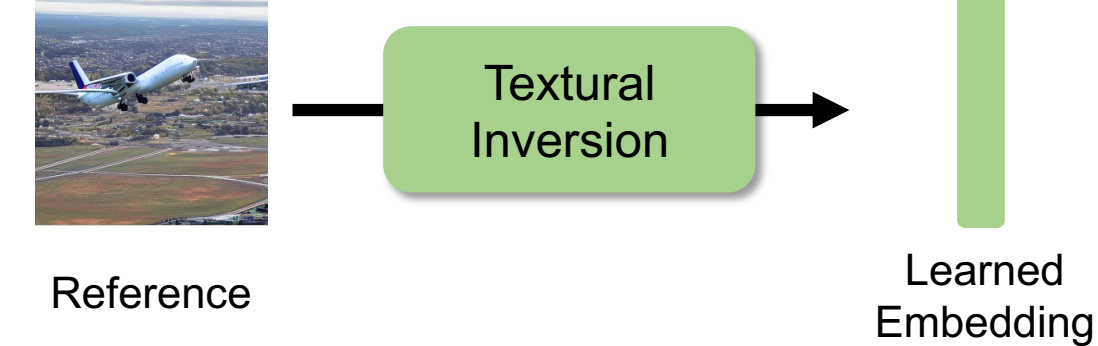
This multimodal setup exposes failure modes hidden by standard text-only evaluations.

Step 1: Multimodal Input

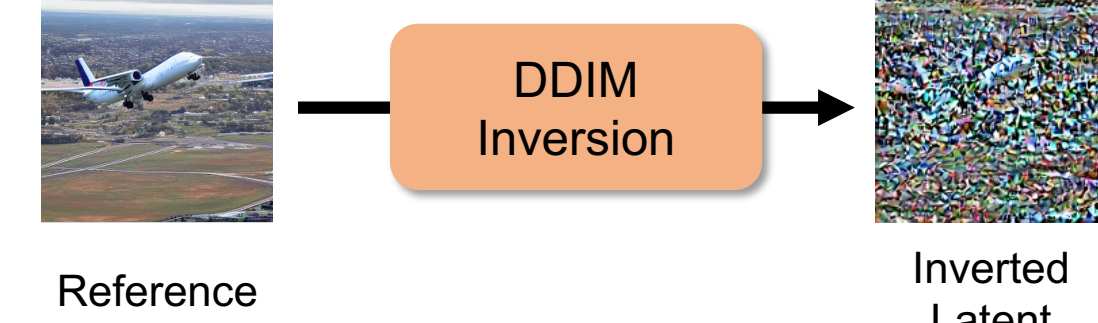
a) Prompts (Generation)

“airplane” or 
Text Prompt Adversarial Prompt

b) Learned Embedding (Personalization)

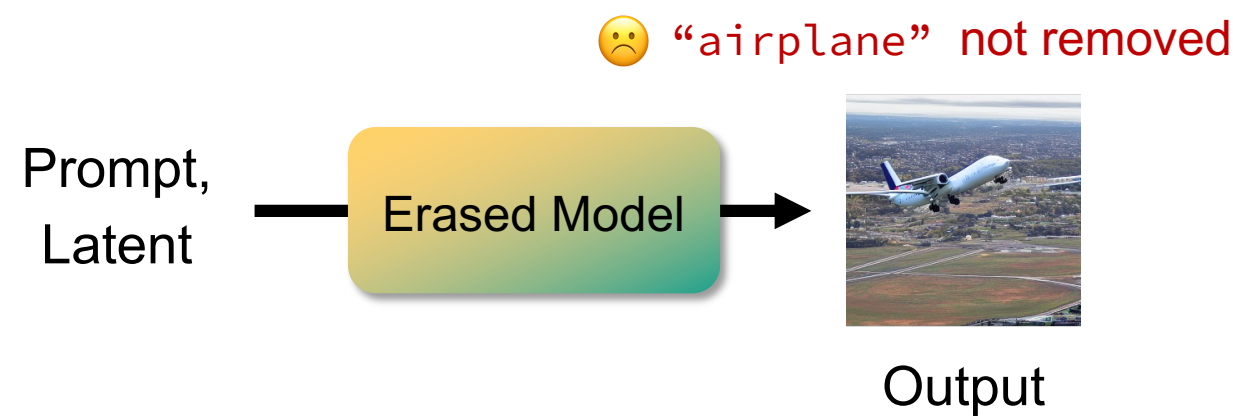


c) Latent Inversion (Editing)

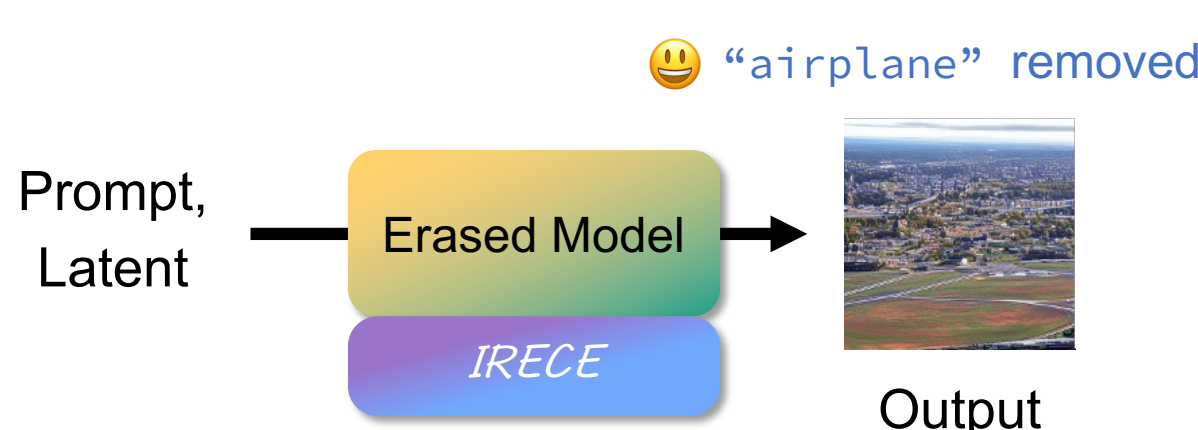


Step 2: DDIM Sampling

Erased Methods



Erased Methods + IRECE



Evaluation Results

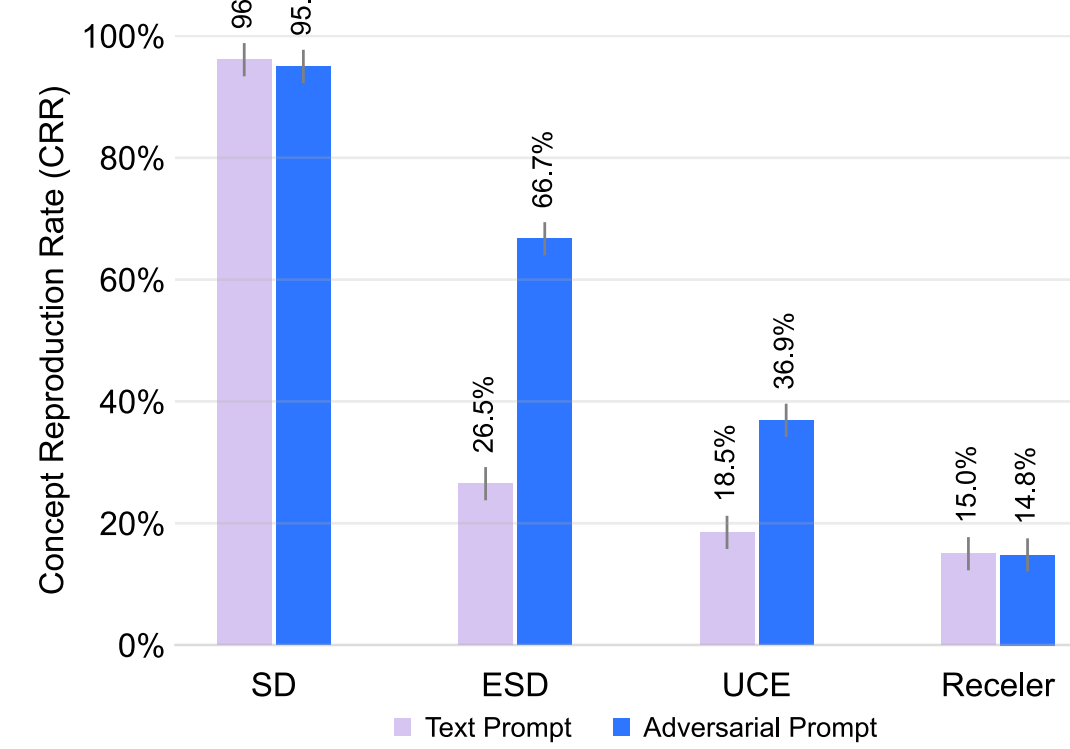
Dataset

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

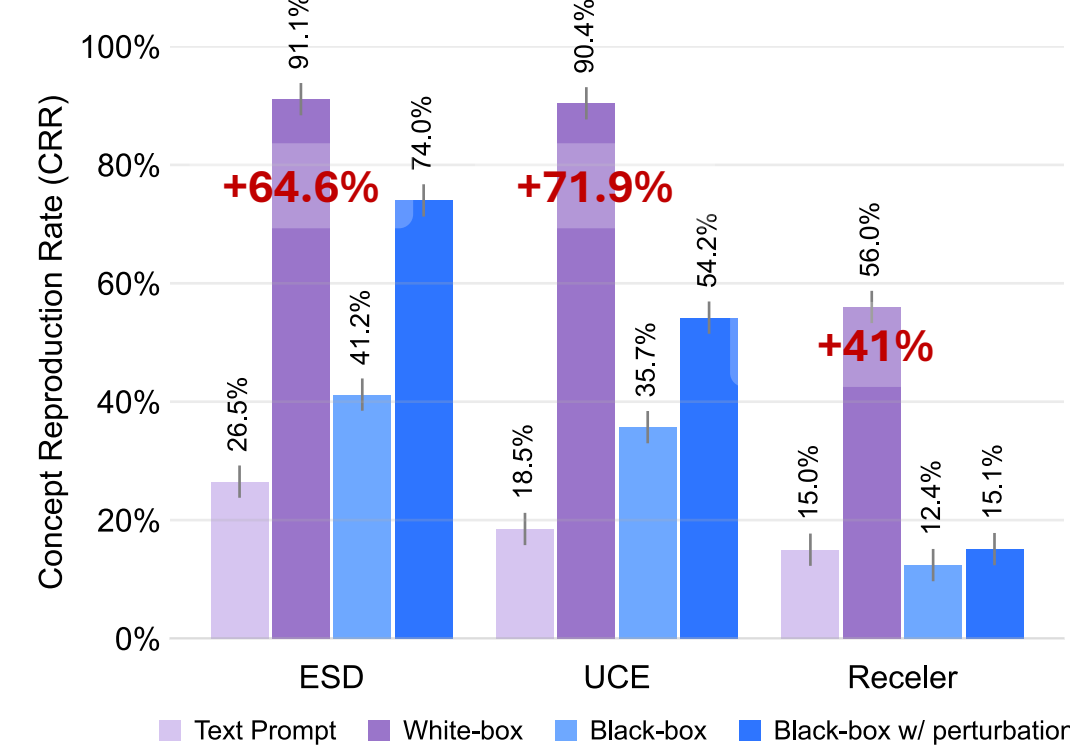
Concept Reproduction Rate (CRR)

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

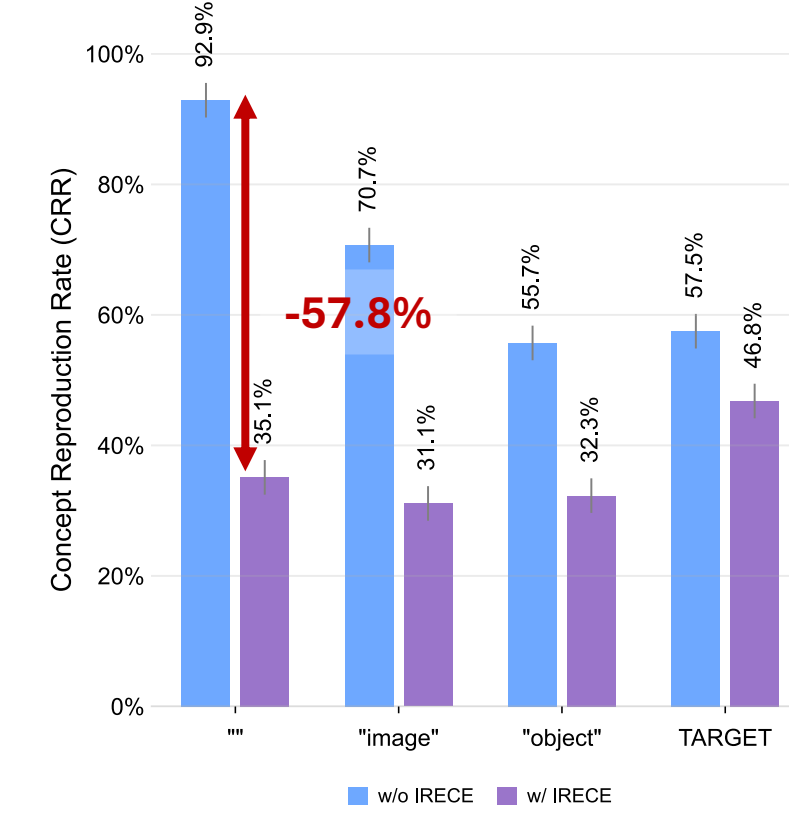
(a) Text / Adv. Prompt



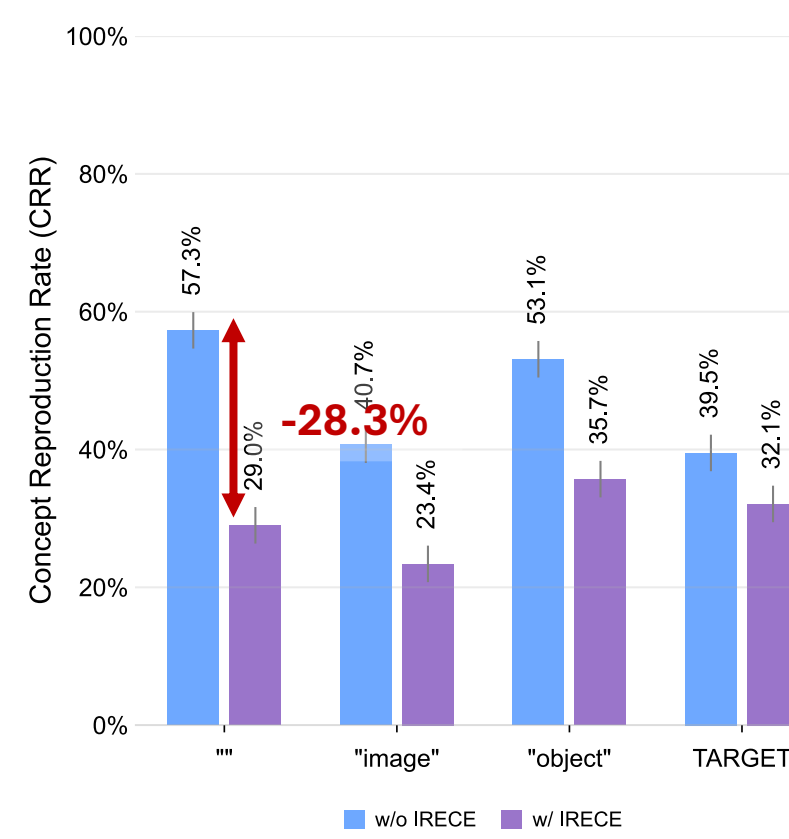
(b) Learned Embedding



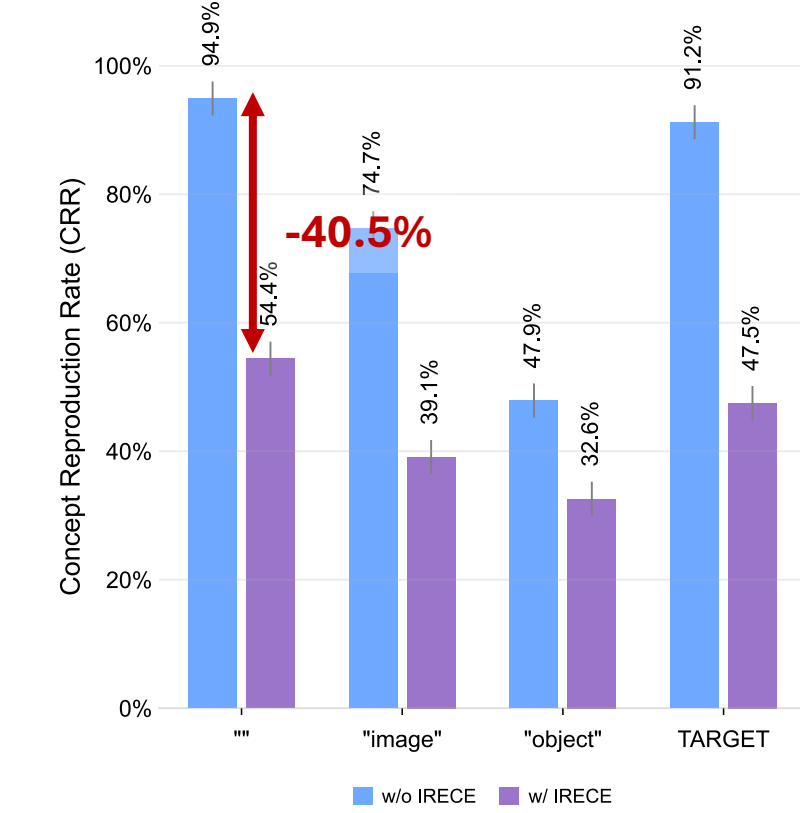
(c) Latent Inversion



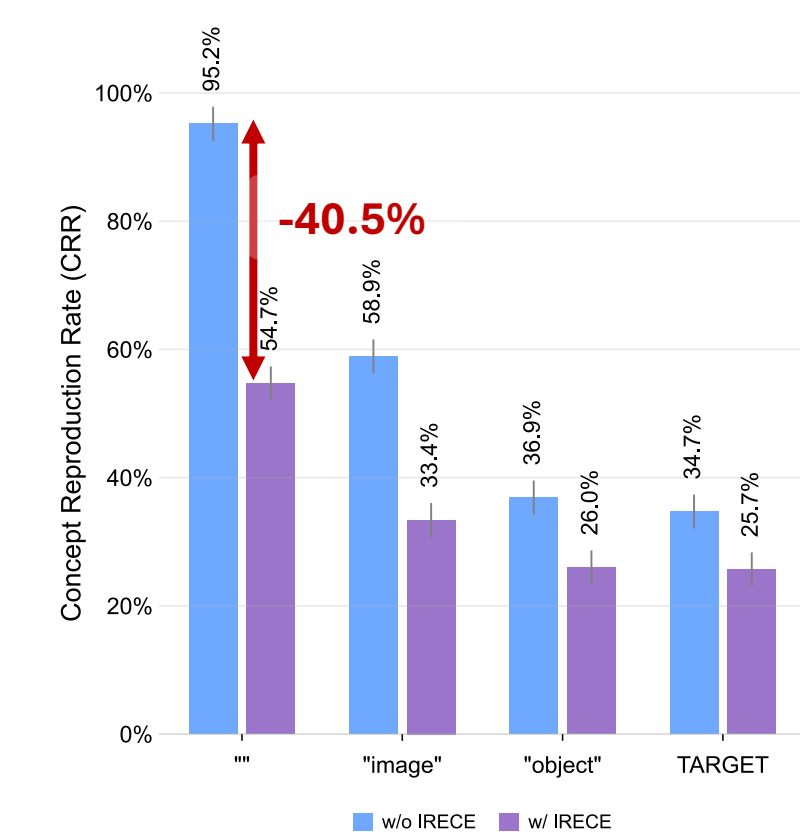
ESD (white-box)



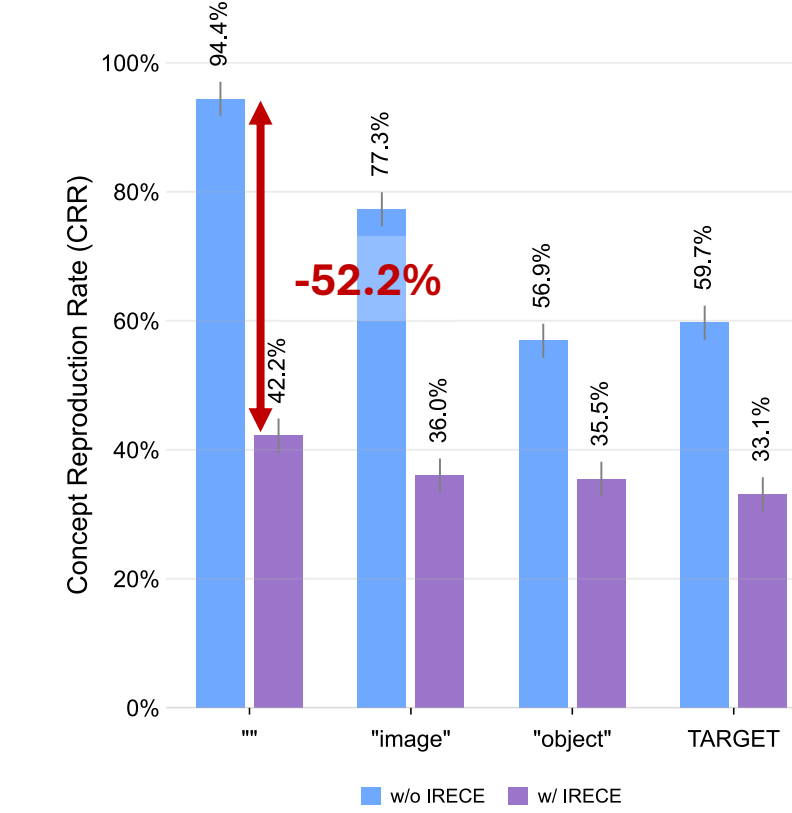
ESD (black-box)



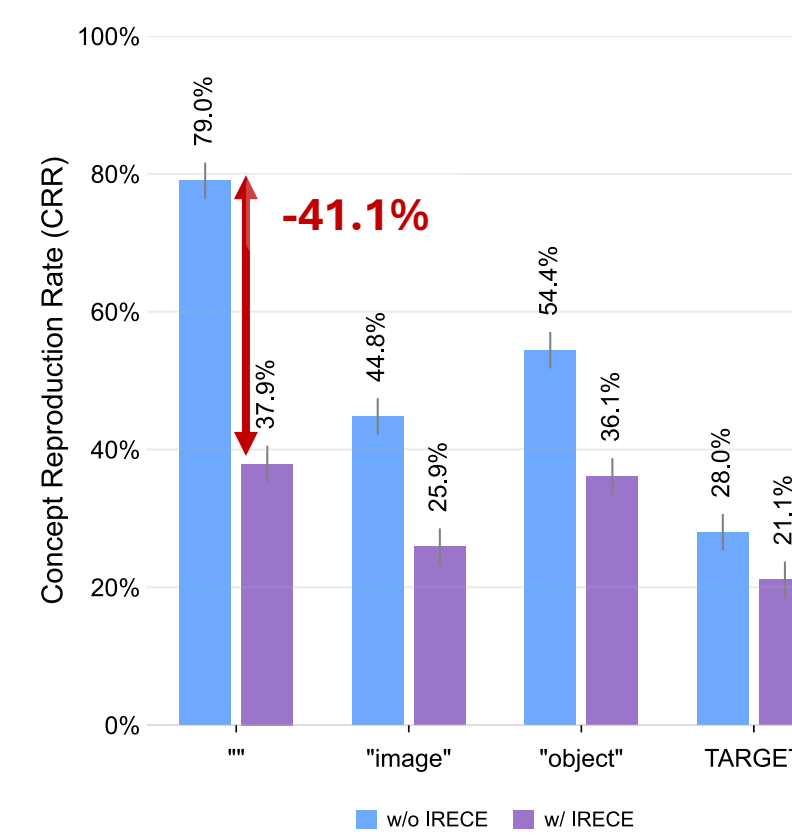
UCE (white-box)



UCE (black-box)



Receler (white-box)



Receler (black-box)

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

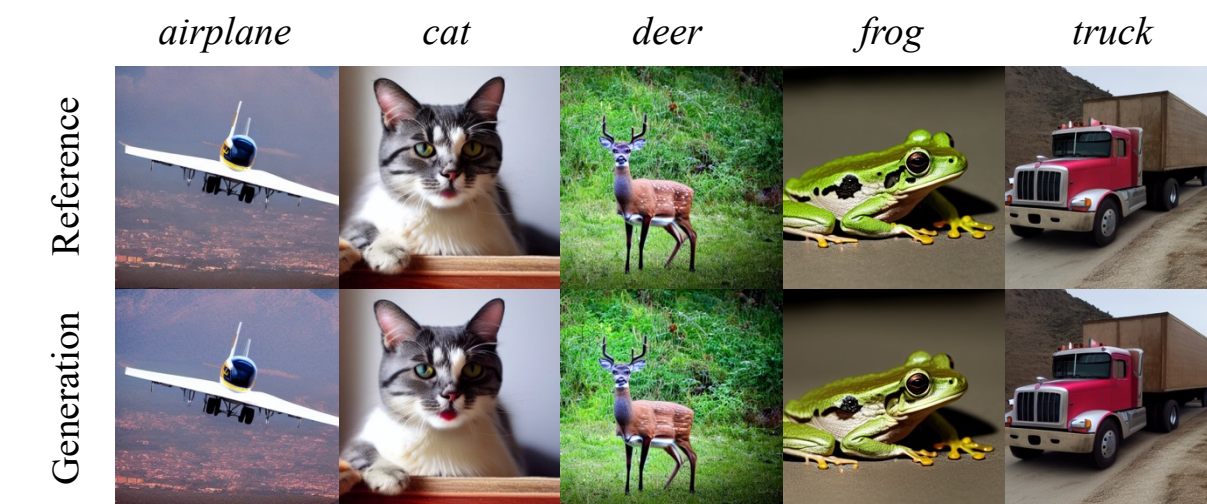
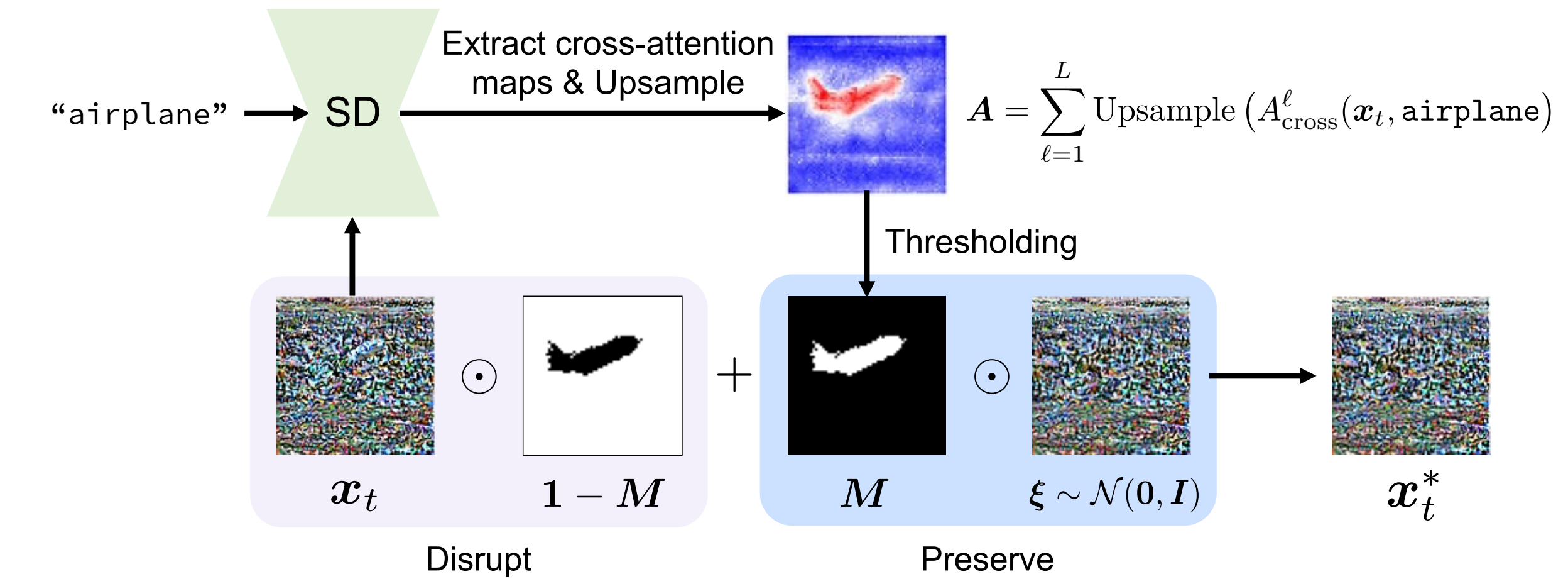


Figure. Qualitative results of generated images from concept erased diffusion models under the **black-box** setting with perturbed reference images in the **learned embedding evaluation**.

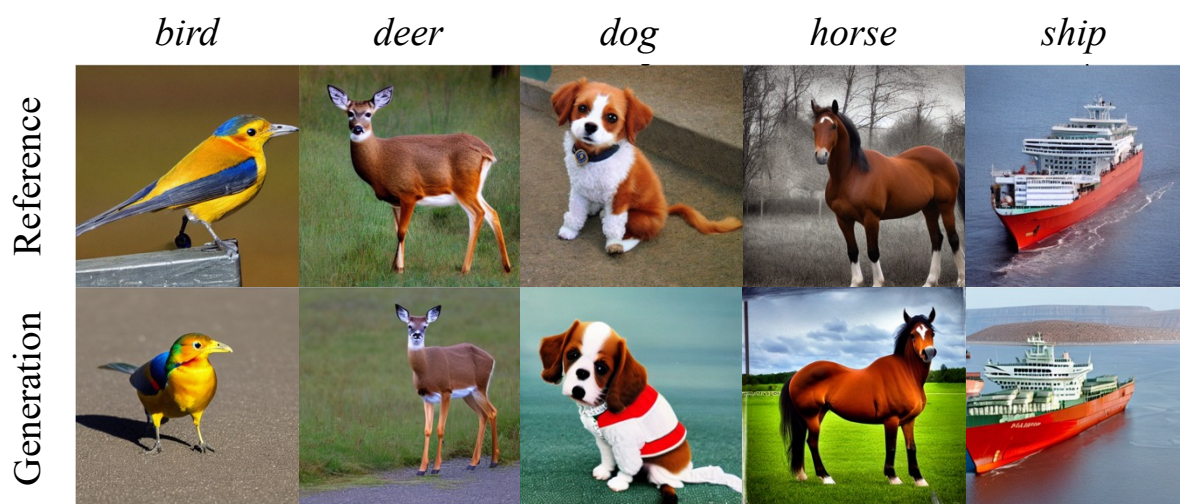


Figure. Qualitative results for generated images from concept erased diffusion models with **unconditional prompt** under the **black-box latent inversion evaluation**.



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.