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

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

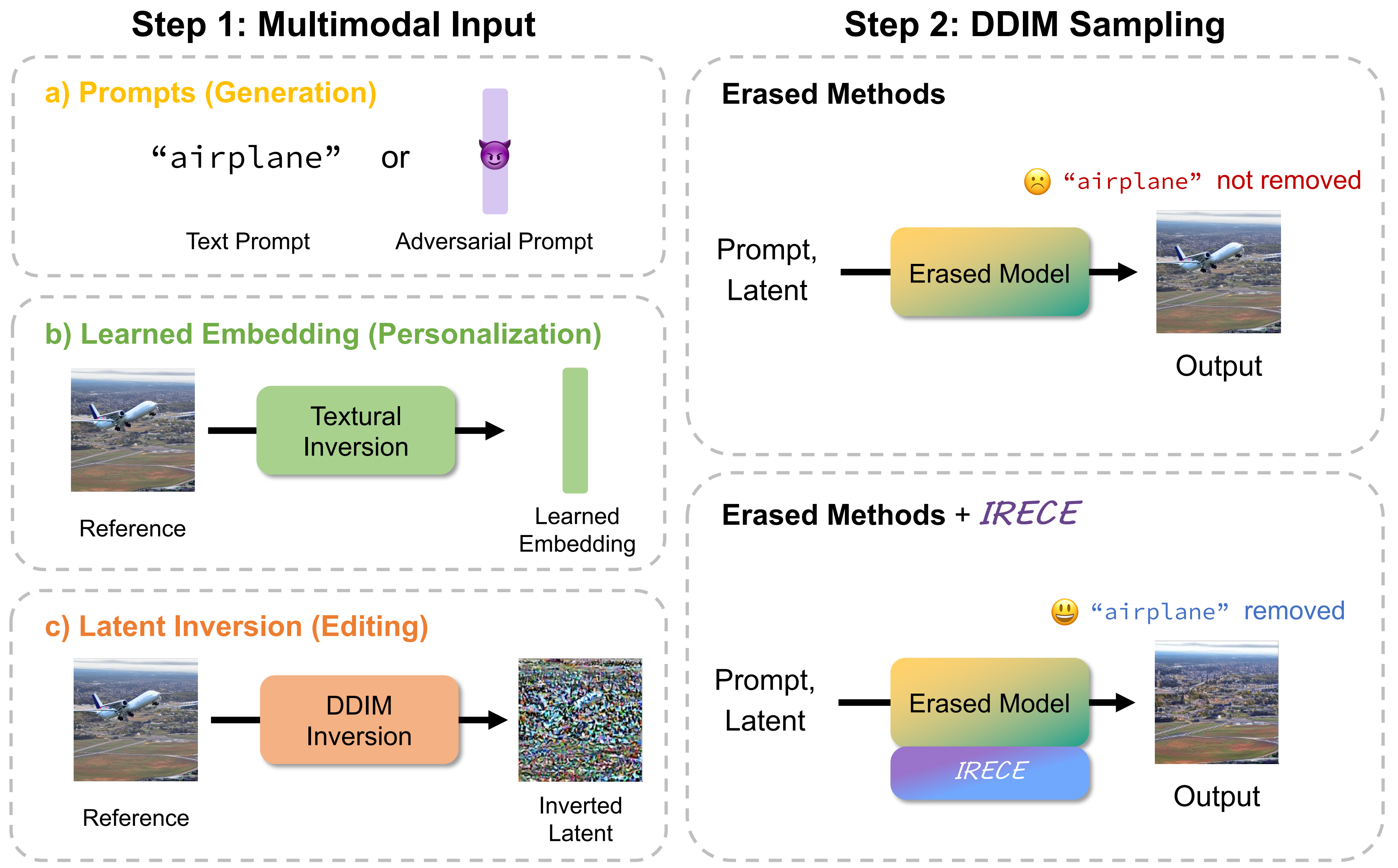
M-ErasureBench: Multimodal Evaluation Framework

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:

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

Reference

Textual Inversion

Learned Embedding

c) Latent Inversion (Editing)

Reference

DDIM Inversion

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

Acknowledgement: This research is supported by the National Science and Technology Council (NSTC) under the grant of NSTC-114-2634-F-002-004, NSTC-114-2634-F-001-001-MBK, NSTC-113-2634-F-002-008, NSTC-114-2221-E-001-016 and NSTC-114-2221-E-001-004, NSTC-114-2221-E-002-182-MY3, NSTC-113-2221-E-002-201 and Academia Sinica under the grant of AS-CDA-110-M09 and AS-IAIA-114-M10

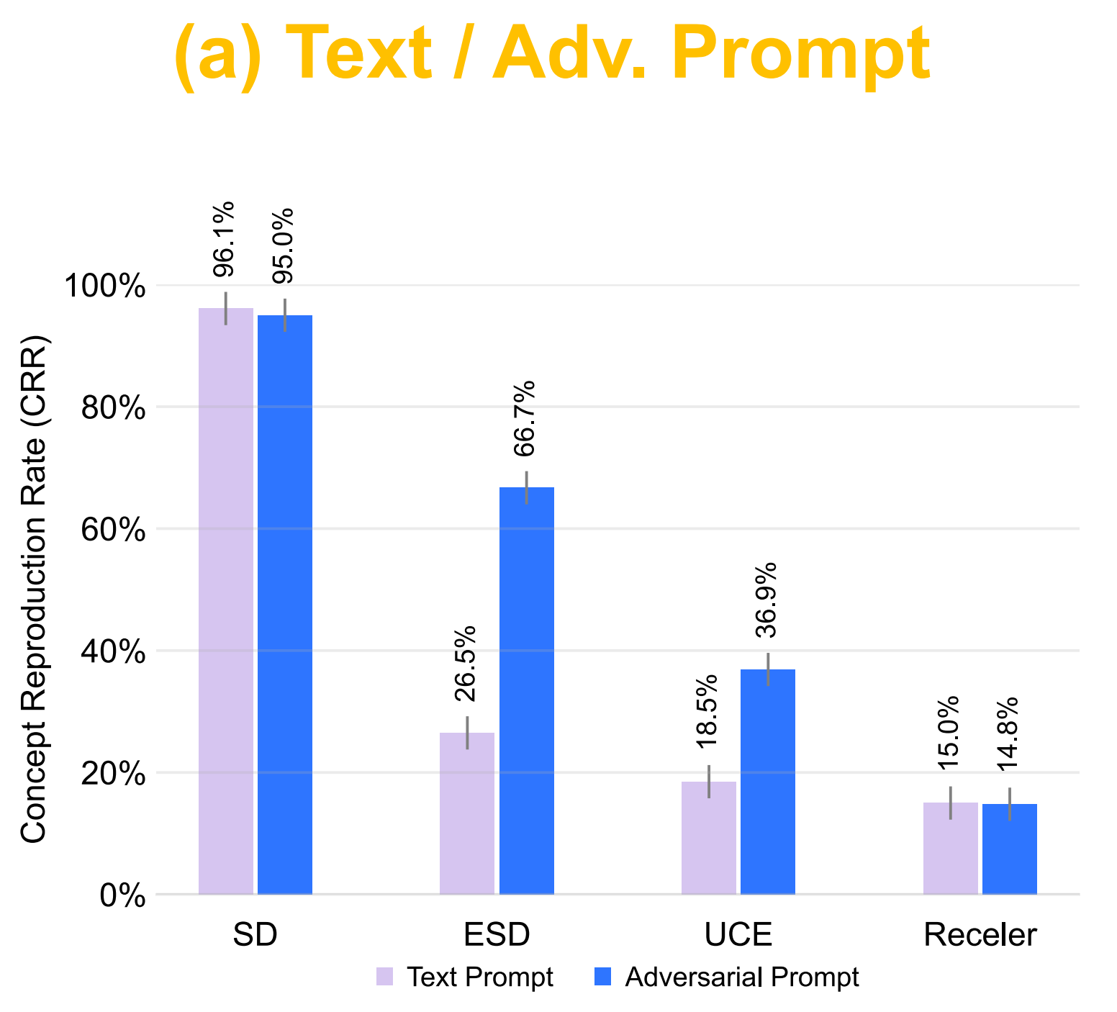
Evaluation Results

Dataset

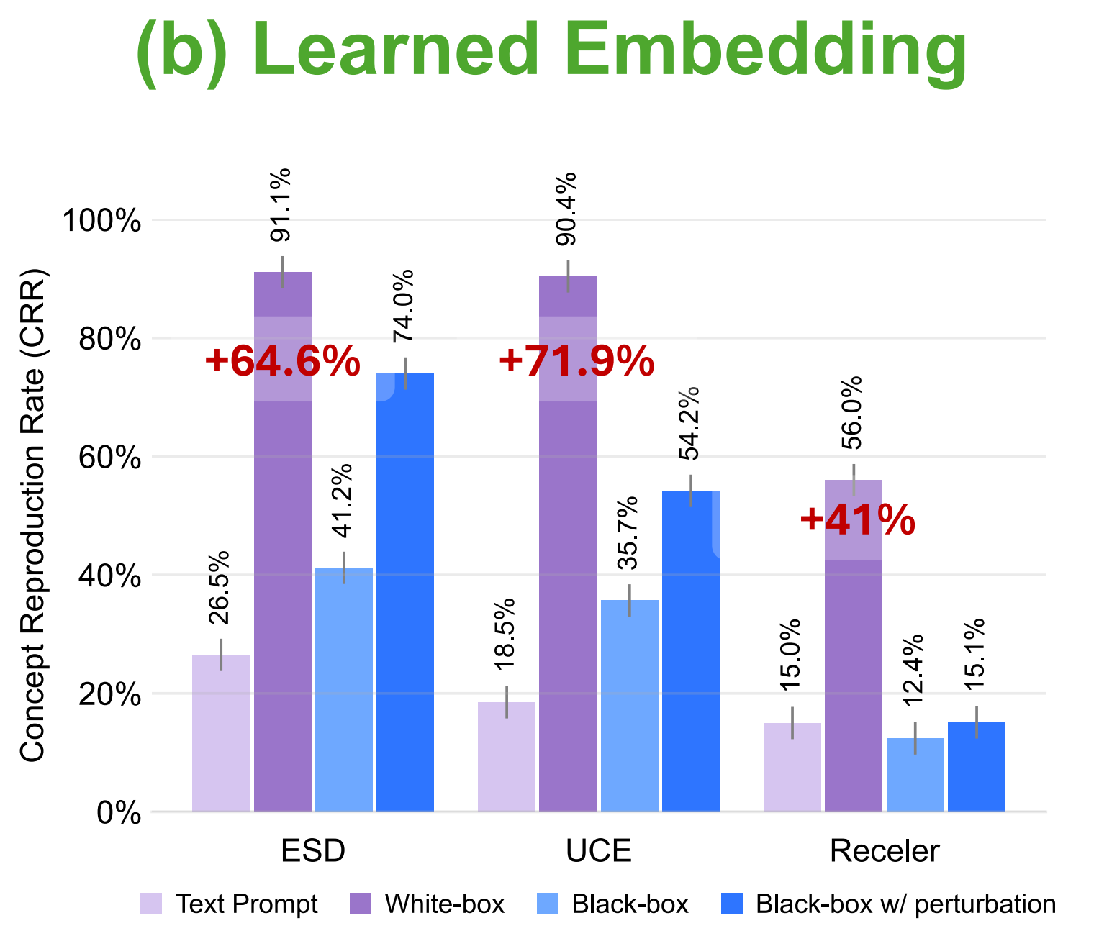
1. **SD-Normal:** Generated from Stable Diffusion using five prompt templates, 150 prompts per class
2. **SD-AdvPrompt:** Adversarial prompts produced using Ring-A-Bell
3. **SD-TI:** Textual Inversion embeddings trained for each reference image
4. **SD-LatentInv:** DDIM-inverted latents from reference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

Concept Reproduction Rate (CRR)

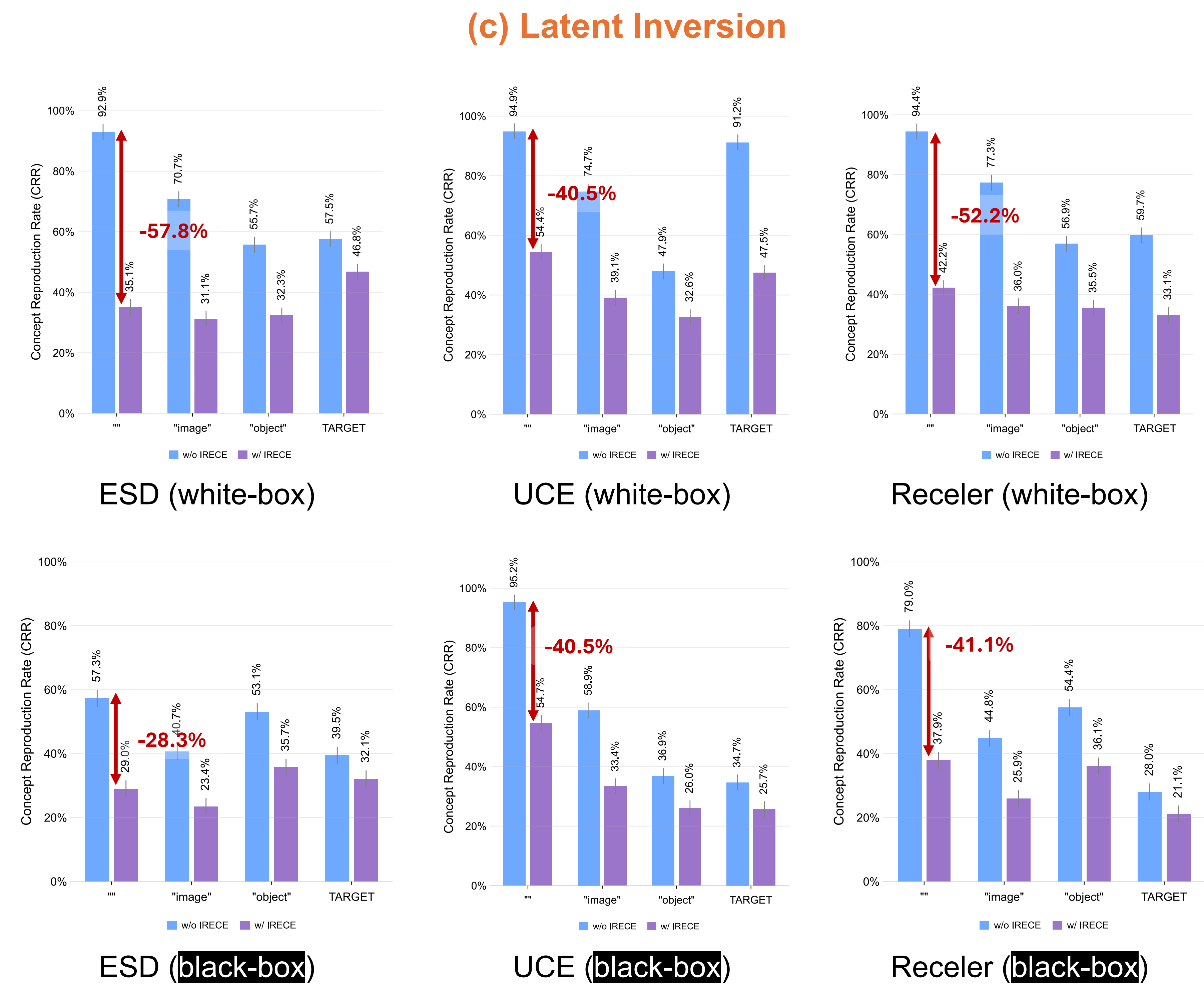
- How often an erased concept reappears in generated images, detected using GroundingDINO



(a) Text / Adv. Prompt



(b) Learned Embedding



(c) Latent Inversion

ESD (white-box)

UCE (white-box)

Receler (white-box)

ESD (black-box)

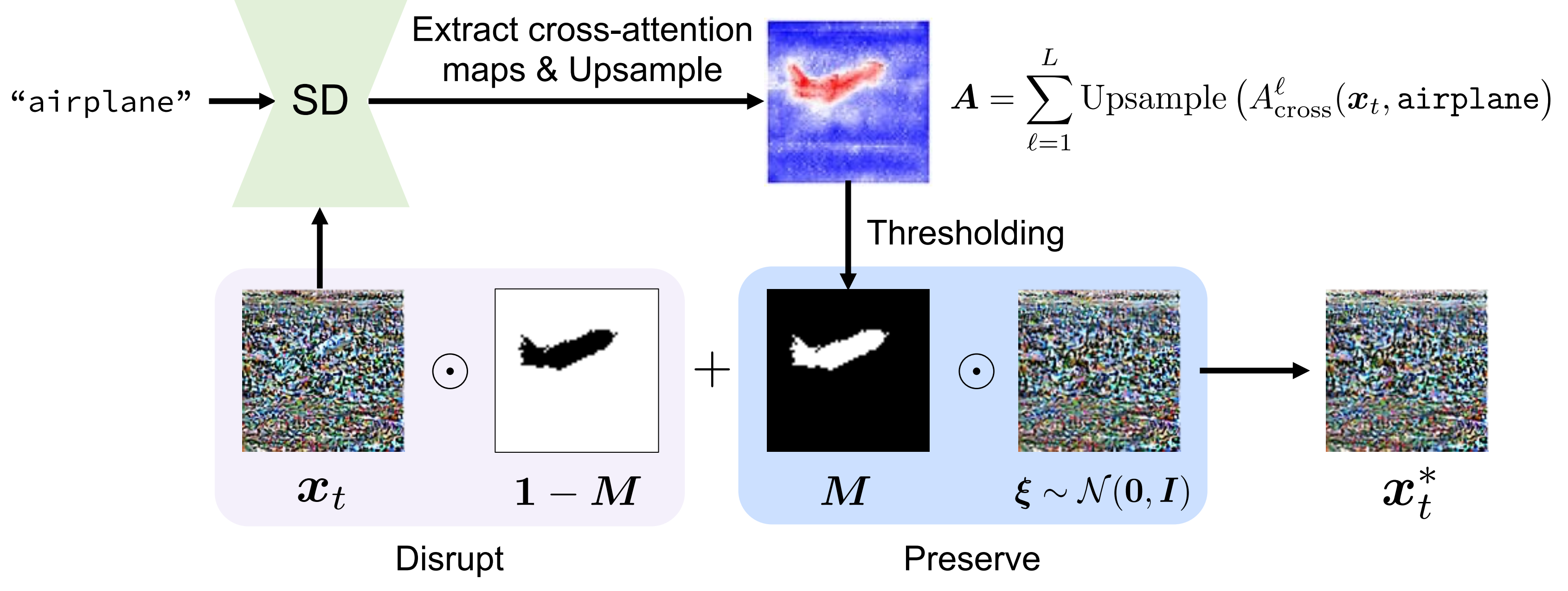
UCE (black-box)

Receler (black-box)

- Learned Embeddings: In white-box, ESD and UCE both exceed 90% CRR
- Latent Inversion: Under the "" prompt, all methods exceed 90% CRR in white-box; gray-box remains high
- IRECE: Lowers white-box latent inversion CRR by ≈ 40%, and gray-box by ≈ 30%

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



"airplane" → SD

Extract cross-attention maps & Upsample

$$A = \sum_{\ell=1}^{L} \text{Upsample}\left(A^{\ell}_{\text{cross}}(x_t, \texttt{airplane})\right)$$

Thresholding

Disrupt

Preserve

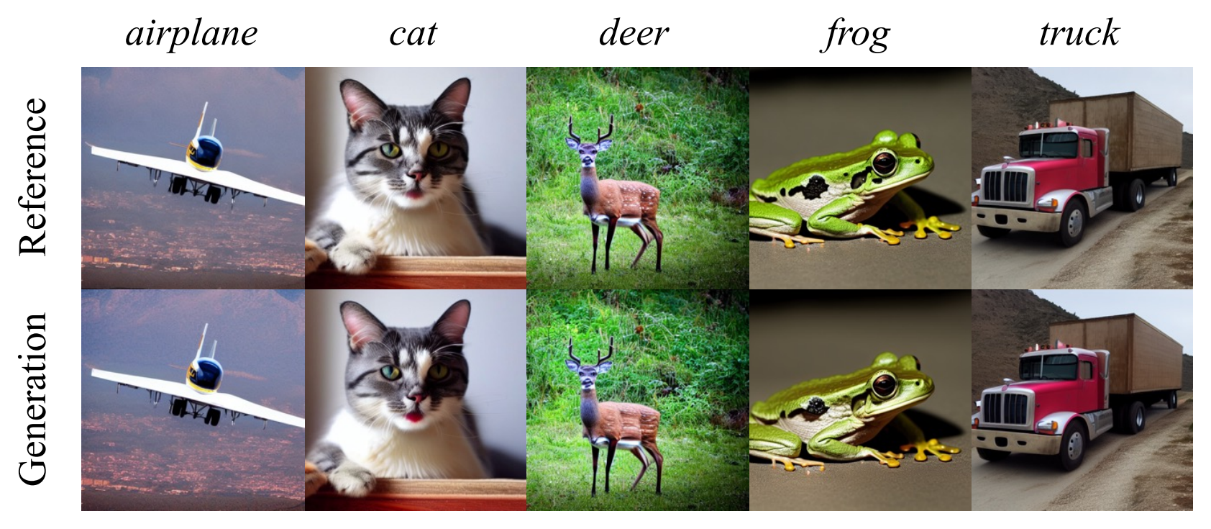


Figure. Qualitative results of generated images from concept erased diffusion models under the black-box setting with perturbed reference images in the learned embedding evaluation.

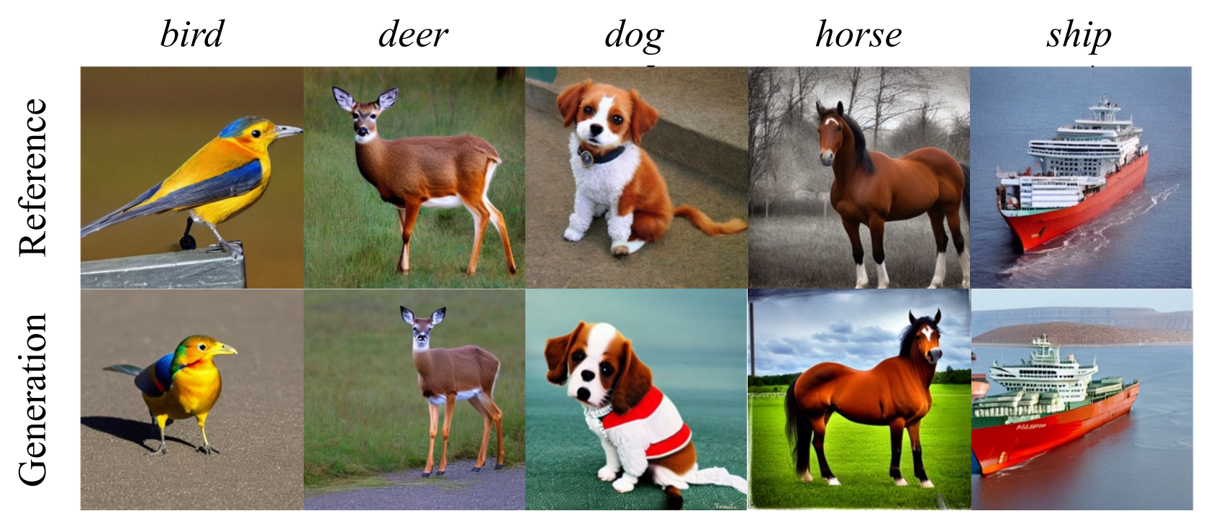


Figure. Qualitative results for generated images from concept erased diffusion models with **unconditional prompt** under the black-box latent inversion evaluation.

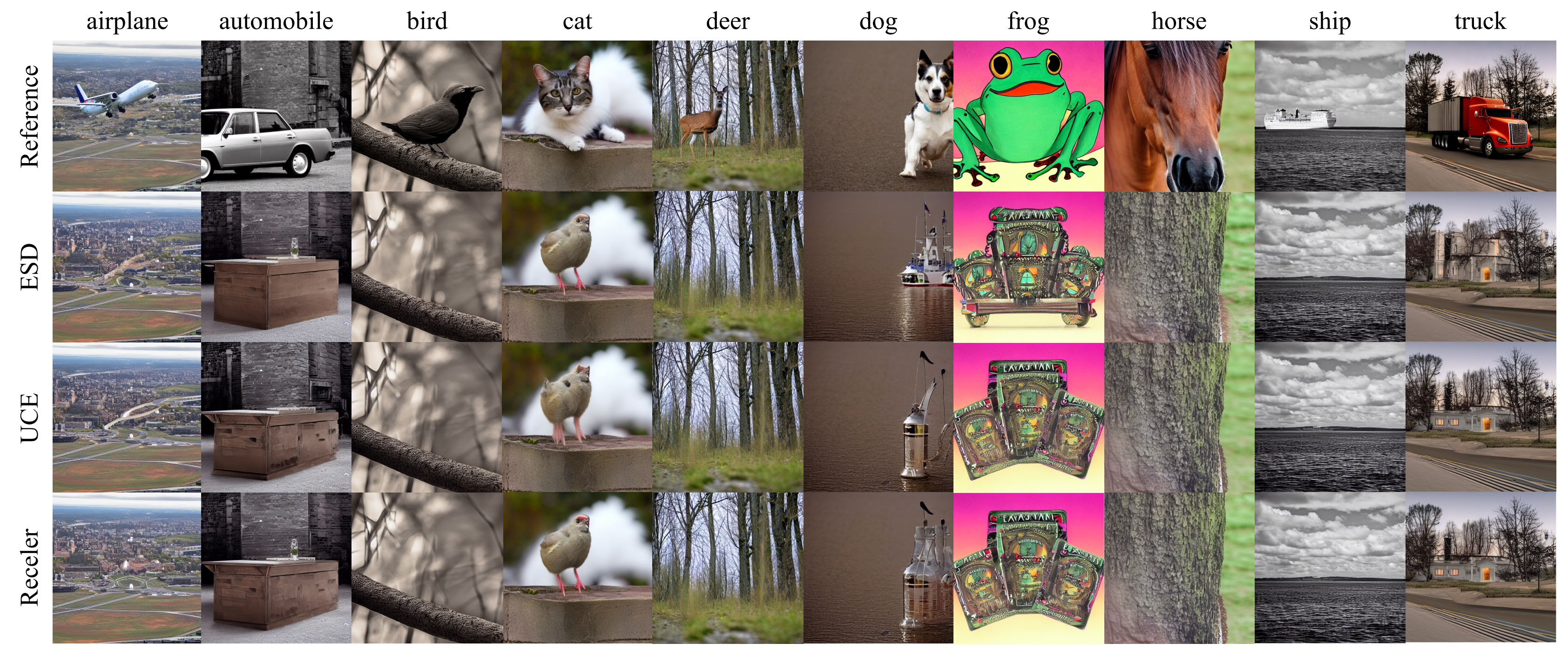


Figure. Comparison of erased models with IRECE across 10 concepts under **white-box** latent inversion. IRECE effectively removes the target concept while preserving the rest of the image.