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

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

1. IRECE Algorithm

We present the detailed algorithm of IRECE in Algorithm 1.

Algorithm 1 IRECE

Input: Sample prompt $\boldsymbol{c}_{\text{sam}}$, Target prompt $\boldsymbol{c}_{\text{tgt}}$, Initial latent $\boldsymbol{x}_T$, Noise schedule $\{\bar{\alpha}_t\}_{t=1}^{T}$, Intervention step t^*, Concept localization threshold τ, Standard model θ_{std}, Erased model θ_{era}.

for $t = T$ to 1 **do**

  if $t = t^*$ **then**

    Extract cross-attention maps from each layer ℓ of model θ (white-box: θ_{era}, black-box: θ_{std}):

$$\boldsymbol{A}^{\ell} \leftarrow A^{\ell}_{\text{cross}}(\boldsymbol{x}_t, \boldsymbol{c}_{\text{tgt}}; \theta)$$

    Aggregate maps after upsampling:

$$\boldsymbol{A} \leftarrow \sum_{\ell=1}^{L} \text{Upsample}(\boldsymbol{A}^{\ell})$$

    Construct binary mask:

$$\boldsymbol{M}(i,j) \leftarrow \begin{cases} 1, & \boldsymbol{A}(i,j) \geq \tau \\ 0, & \text{otherwise} \end{cases}$$

    Perturb target regions with Gaussian noise $\boldsymbol{\xi}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$:

$$\boldsymbol{x}_t \leftarrow (\boldsymbol{1} - \boldsymbol{M}) \odot \boldsymbol{x}_t + \boldsymbol{M} \odot \boldsymbol{\xi}_t$$

  end if

  Perform the denoising step with the erased model:

$$\boldsymbol{x}_{t-1} \leftarrow \text{DDIMStep}(\boldsymbol{x}_t, \boldsymbol{c}_{\text{sam}}, t, \theta_{\text{era}})$$

end for

Output: Generated output $\boldsymbol{x}_0$.

2. Implementation Details

Sampling. All sampling procedures are conducted using the `DDIMScheduler` in diffuser [5], with the guidance scale set to 7.5 and the number of inference fixed at 50.

Prompt Configurations. Table 1 presents the descriptions of the four prompt configurations used in the latent inversion evaluation.

Prompt	Description
""	Null text (unconditional generation with no textual guidance).
"image"	Generic placeholder describing the input image without specifying any object.
"object"	Coarse reference to the foreground object without explicitly naming it.
`TARGET`	Explicitly naming the target concept intended for erasure.

Table 1. Prompt configurations for latent inversion evaluation.

Parameters of IRECE. For robustness enhancement, the concept localization threshold τ is set to 0.4 and the intervention timestep t^* at 781.

3. More Experimental Results

3.1. Text Prompt Evaluation

We report per-class results for both text prompt and adversarial prompt evaluations in Table 2. Across the ten categories, *automobile* consistently exhibits the highest CRR under all methods, indicating that erasure is less effective for this class. A plausible reason is the large intra-class diversity of automobiles, which makes the concept harder to suppress. Despite this challenge, all methods still reduce CRR by at least 39% for *automobile*, confirming that suppression remains non-trivial but effective to some extent.

Methods	*airplane*	*automobile*	*bird*	*cat*	*deer*	*dog*	*frog*	*horse*	*ship*	*truck*	**Avg.**
					Text Prompt						
SD v1.4 [4]	94.67	98.67	92.67	96.00	99.33	98.67	92.00	98.67	94.67	96.00	**96.14**
ESD [1]	27.33	59.33	10.00	31.33	16.67	22.67	19.33	18.67	32.00	27.33	**26.47**
UCE [2]	36.00	44.00	3.33	9.33	7.33	6.00	14.67	5.33	30.67	28.00	**18.47**
Receler [3]	6.67	50.67	2.67	6.67	5.33	2.00	25.33	8.67	26.67	15.33	**15.00**
					Adversarial Prompt						
SD v1.4 [4]	85.33	95.56	94.67	95.56	100.00	97.33	99.33	96.00	92.67	93.33	**94.98**
ESD [1]	74.67	95.56	69.33	52.14	50.00	79.33	32.00	47.33	82.67	84.17	**66.72**
UCE [2]	70.67	72.67	30.00	14.67	9.33	22.67	35.33	18.67	43.33	52.00	**36.93**
Receler [3]	8.67	78.89	4.00	0.00	0.00	2.00	11.33	0.00	34.67	8.33	**14.79**

Table 2. Concept Reproduction Rate (CRR) of concept erasure methods on **text** and **adversarial prompts**, reported per class. Orange marks classes with CRR $>$ 50%, and red marks methods with average CRR $>$ 50%.

Methods	Settings	*airplane*	*automobile*	*bird*	*cat*	*deer*	*dog*	*frog*	*horse*	*ship*	*truck*	**Avg.**
ESD [1]	Text prompt	27.33	59.33	10.00	31.33	16.67	22.67	19.33	18.67	32.00	27.33	**26.47**
	White-box	78.00	97.33	93.33	98.00	94.67	96.67	82.67	84.67	96.67	88.67	**91.07**
	Black-box	44.67	82.67	15.33	37.33	30.00	31.33	25.33	18.00	86.67	40.67	**41.20**
	Black-box w/ perturb.	74.67	97.33	88.00	81.33	39.33	80.67	52.67	66.00	93.33	66.67	**74.00**
UCE [2]	Text prompt	36.00	44.00	3.33	9.33	7.33	6.00	14.67	5.33	30.67	28.00	**18.47**
	White-box	70.67	98.67	92.67	96.00	93.33	84.00	90.67	99.33	95.33	83.33	**90.40**
	Black-box	7.33	92.00	8.00	62.00	23.33	22.67	16.67	8.00	70.00	46.67	**35.67**
	Black-box w/ perturb.	34.00	86.67	54.67	93.33	65.33	48.00	29.33	14.00	59.33	57.33	**54.20**
Receler [3]	Text prompt	6.67	50.67	2.67	6.67	5.33	2.00	25.33	8.67	26.67	15.33	**15.00**
	White-box	50.00	89.33	18.00	39.33	41.33	92.00	34.67	22.67	82.67	90.00	**56.00**
	Black-box	8.67	64.00	0.67	0.00	1.33	1.33	16.67	0.00	10.67	20.67	**12.40**
	Black-box w/ perturb.	12.00	78.00	2.00	2.67	1.33	3.33	12.67	2.67	24.67	11.33	**15.07**

Table 3. Concept Reproduction Rate (CRR) of concept-erasure methods in **learned embedding evaluation**, reported per class. Orange marks classes with CRR $>$ 50%, and red marks methods with average CRR $>$ 50%.

3.2. Learned Embedding Evaluation

We present per-class results for learned embedding evaluation in Table 3. Compared to the text prompt baseline, CRR under the white-box setting rises above 50% for many categories across all methods, indicating that learned embeddings substantially reduces the effectiveness of erasure. In the black-box setting, CRR remains below 50% for most categories, but introducing perturbations substantially increases CRR. For example, ESD and UCE exceed 50% CRR in classes such as *bird*, *cat*, *deer*, and *truck*, highlighting that even black-box settings can become highly effective when enhanced with perturbations.

3.3. Latent Inversion Evaluation

We report per-class results for latent inversion evaluation in Table 4. In the white-box setting, concept erasure methods show consistently high CRR: with prompts such as “” and “image”, nearly all categories exceed 50%, while only `TARGET` achieves CRR below 50% in a few cases (e.g., *deer*, *horse*). In the black-box setting, the unconditional prompt “” still drives CRR above 50% for most categories, underscoring the vulnerability of erased models under latent inversion evaluation.

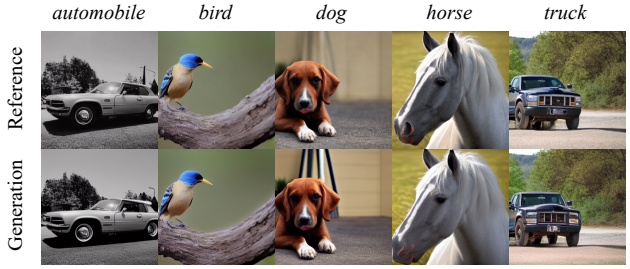

Figure 1. Qualitative results for generated images with “image” prompt under the black-box latent inversion evaluation.

Figure 1 shows that the “image” prompt strategy also performs strongly under white-box access, successfully

capturing the overall semantics in most cases. In contrast, the TARGET strategy exhibits markedly different behaviors across access settings. As shown in Figure 2, under black-box access it often achieves effective concept removal, producing outputs that deviate substantially from the original semantics. In the white-box case, however, the generated images continue to depict the target concept, albeit with noticeable disruptions in the corresponding regions.

Figure 2. Qualitative results for TARGET prompt generations under white-box and black-box latent inversion evaluation.

3.4. (Surrogate-based) Black-box with Different Backbones

We further investigate how backbone discrepancies between the surrogate and erased models affect the effectiveness of concept erasure. Since the null text prompt in the white-box setting yields the most prominent performance (Table 4), we mainly adopt it as our case study for **latent inversion evaluation**. Specifically, we generate inverted latents via DDIM inversion using SD v1.5, while all concept-erasure methods are evaluated on an erased model based on SD v1.4. As shown in Table 5, and in reference to the null-text prompt results in Table 4, when the surrogate and erased models use different backbones, the CRRs are lower than those in the white-box setting; meanwhile, they remains consistently higher than those in the corresponding black-box setting where both models share the same backbone. This is because the backbone mismatch causes the surrogate's representation space to deviate from that of the erased model, producing latents falling outside the regions where the erased model was trained to suppress the target concept.

3.5. Inference-time Robustness Enhancement for Concept Erasure (IRECE)

We report the detailed per-class CRR results before and after applying IRECE for three concept erasure methods: ESD [1], UCE [2], and Receler [3]. White-box results are shown in Table 6, and those under the black-box setting are provided in Table 7. Across most classes and prompt strategies, IRECE achieves a substantial reduction in CRR, indicating improved robustness against concept re-emergence.

IRECE introduces two tunable hyperparameters: the intervention timestep t^* and the concept localization threshold τ. Their effects are shown in Figure 3.

Intervention Timestep. This parameter specifies when the erasure is applied during the diffusion process. Applying it too early disrupts the overall image structure, as the latent representation is not yet sufficiently developed. Applying it too late leaves too few denoising steps, often leading to blending artifacts and incomplete suppression.

Concept Localization Threshold. This parameter controls the aggressiveness of erasure. A lower τ produces broader masks that may unintentionally remove nearby non-target regions, while a higher τ results in tighter masks that risk leaving traces of the target concept. Its role is analogous to the scale coefficient in Receler [3], which adjusts the trade-off between erasure strength and image fidelity.

References

[1] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 4, 5, 6

[2] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 2, 3, 4, 5, 6

[3] Chi-Pin Huang, Kai-Po Chang, Chung-Ting Tsai, Yung-Hsuan Lai, Fu-En Yang, and Yu-Chiang Frank Wang. Receler: Reliable concept erasing of text-to-image diffusion models via lightweight erasers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 5, 6

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[5] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, 2022. 1

Methods	Prompt Strategy	*airplane*	*automobile*	*bird*	*cat*	*deer*	*dog*	*frog*	*horse*	*ship*	*truck*	**Avg.**
						White-box Setting						
ESD [1]	Text prompt	27.33	59.33	10.00	31.33	16.67	22.67	19.33	18.67	32.00	27.33	**26.47**
	“”	88.67	98.00	89.33	94.67	94.67	98.00	92.67	95.33	86.67	90.67	**92.87**
	“image”	71.33	86.00	76.00	68.67	57.33	73.33	78.67	65.33	70.67	59.33	**70.67**
	“object”	72.67	80.67	52.00	48.67	38.67	39.33	68.00	48.00	60.00	48.67	**55.67**
	TARGET	68.00	86.67	52.00	50.67	35.33	58.67	58.67	55.33	63.33	46.00	**57.47**
UCE [2]	Text prompt	36.00	44.00	3.33	9.33	7.33	6.00	14.67	5.33	30.67	28.00	**18.47**
	“”	92.00	98.00	91.33	96.67	98.00	98.67	92.67	97.33	89.33	95.33	**94.93**
	“image”	70.00	77.33	83.33	73.33	75.33	62.00	76.67	78.67	73.33	76.67	**74.67**
	“object”	60.00	75.33	46.67	35.33	30.67	29.33	74.00	37.33	49.33	41.33	**47.93**
	TARGET	84.67	96.67	90.67	94.67	86.67	94.67	91.33	92.67	88.00	92.00	**91.20**
Receler [3]	Text prompt	6.67	50.67	2.67	6.67	5.33	2.00	25.33	8.67	26.67	15.33	**15.00**
	“”	91.33	100.00	90.67	94.67	97.33	98.67	92.00	96.67	88.00	94.67	**94.40**
	“image”	72.67	84.67	82.00	74.67	68.67	82.00	80.67	78.00	68.67	80.67	**77.27**
	“object”	67.33	72.67	54.00	50.00	48.00	38.67	77.33	42.67	63.33	54.67	**56.87**
	TARGET	61.33	91.33	60.67	70.00	26.67	62.00	80.67	34.67	54.00	56.00	**59.73**
						Black-box Setting						
ESD [1]	Text prompt	27.33	59.33	10.00	31.33	16.67	22.67	19.33	18.67	32.00	27.33	**26.47**
	“”	71.33	81.33	56.00	48.67	34.00	58.00	81.33	58.00	54.67	30.00	**57.33**
	“image”	51.33	73.33	24.67	34.67	16.00	40.00	62.00	29.33	39.33	36.00	**40.67**
	“object”	60.67	83.33	37.33	33.33	32.00	46.00	80.00	43.33	58.67	56.67	**53.13**
	TARGET	46.67	74.00	16.67	32.00	35.33	34.00	57.33	27.33	32.67	38.67	**39.47**
UCE [2]	Text prompt	36.00	44.00	3.33	9.33	7.33	6.00	14.67	5.33	30.67	28.00	**18.47**
	“”	91.33	98.00	92.67	96.67	98.00	98.67	93.33	97.33	90.00	96.00	**95.20**
	“image”	54.00	70.00	62.67	60.67	54.67	41.33	75.33	43.33	34.67	92.00	**58.87**
	“object”	31.33	74.00	20.67	8.00	15.33	21.33	63.33	30.00	37.33	68.00	**36.93**
	TARGET	40.67	74.00	11.33	28.00	30.00	30.00	50.00	25.33	24.67	33.33	**34.73**
Receler [3]	Text prompt	6.67	50.67	2.67	6.67	5.33	2.00	25.33	8.67	26.67	15.33	**15.00**
	“”	84.00	87.33	82.00	68.00	66.67	82.67	84.00	84.67	77.33	73.33	**79.00**
	“image”	43.33	73.33	36.67	40.00	26.00	50.00	56.67	43.33	27.33	51.33	**44.80**
	“object”	70.00	66.67	33.33	33.33	46.00	48.00	71.33	40.67	56.67	78.00	**54.40**
	TARGET	28.67	69.33	12.00	17.33	23.33	12.00	50.67	23.33	17.33	26.00	**28.00**

Table 4. Concept Reproduction Rate (CRR) of concept-erasure methods in **latent inversion evaluation**, reported per class. Orange marks classes with CRR $> 50\%$, and red marks methods with average CRR $> 50\%$.

Methods	*airplane*	*automobile*	*bird*	*cat*	*deer*	*dog*	*frog*	*horse*	*ship*	*truck*	**Avg.**
ESD [1]	82.67	94.00	93.33	91.33	79.33	90.67	89.33	94.67	90.67	93.33	**89.93**
UCE [2]	86.00	58.67	77.33	70.00	58.00	83.33	82.67	88.00	94.67	82.00	**78.07**
Receler [3]	86.67	94.00	93.33	87.33	78.67	92.67	90.00	95.33	90.67	96.00	**90.47**

Table 5. Concept Reproduction Rate (CRR) of concept-erasure methods in **latent inversion evaluation with different backbone**, reported per class. Orange marks classes with CRR $> 50\%$, and red highlights methods with average CRR $> 50\%$. Overall, all categories exceed the 50% threshold, showing that the erased model fails to suppress concepts embedded in latents inverted from a different backbone.

Methods	Prompt	IRECE	*airplane*	*automobile*	*bird*	*cat*	*deer*	*dog*	*frog*	*horse*	*ship*	*truck*	**Avg.**
ESD [1]	""		88.67	98.00	89.33	94.67	94.67	98.00	92.67	95.33	86.67	90.67	**92.87**
		✓	24.00	62.67	28.00	42.00	20.67	40.67	50.67	32.67	40.67	9.33	**35.14**
		Δ	-64.67	-35.33	-61.33	-52.67	-74.00	-57.33	-42.00	-62.66	-46.00	-81.34	-57.73
	"image"		71.33	86.00	76.00	68.67	57.33	73.33	78.67	65.33	70.67	59.33	**70.67**
		✓	27.33	62.67	20.00	33.33	19.33	31.33	39.33	30.67	36.67	10.67	**31.13**
		Δ	-44.00	-23.33	-56.00	-35.34	-38.00	-42.00	-39.34	-34.66	-34.00	-48.66	-39.53
	"object"		72.67	80.67	52.00	48.67	38.67	39.33	68.00	48.00	60.00	48.67	**55.67**
		✓	24.67	64.00	19.22	32.67	19.33	26.00	56.67	29.33	32.00	19.33	**32.32**
		Δ	-48.00	-16.67	-32.78	-16.00	-19.34	-13.33	-11.33	-18.67	-28.00	-29.34	-23.35
	TARGET		68.00	86.67	52.00	50.67	35.33	58.67	58.67	44.00	63.33	46.00	**56.33**
		✓	39.33	76.67	19.33	43.33	37.33	47.33	52.00	44.00	60.00	22.67	**44.20**
		Δ	-28.67	-10.00	-32.67	-7.34	+2.00	-11.34	-6.67	0.00	-3.33	-23.33	-12.14
UCE [2]	""		92.00	98.00	91.33	96.67	98.00	98.67	92.67	97.33	89.33	95.33	**90.67**
		✓	37.33	82.00	48.67	71.33	44.00	70.67	64.67	46.00	53.33	26.67	**54.47**
		Δ	-54.67	-16.00	42.66	-25.34	-54.00	-28.00	-28.00	-51.33	-36.00	-68.66	-36.20
	"image"		70.00	77.33	83.33	73.33	75.33	62.00	76.67	78.67	73.33	76.67	**70.67**
		✓	28.00	49.33	26.00	51.33	29.33	31.33	58.00	38.67	38.67	22.67	**37.33**
		Δ	-42.00	-28.00	57.33	-22.00	-46.00	-30.67	-18.67	-40.00	-34.66	-54.00	-33.33
	"object"		60.00	75.33	46.67	35.33	30.67	29.33	74.00	37.33	49.33	41.33	**45.66**
		✓	28.67	71.33	18.00	22.00	17.33	17.33	56.00	32.00	38.00	19.33	**32.00**
		Δ	-31.33	-4.00	28.67	-13.33	-13.34	-12.00	-18.00	-5.33	-11.33	-22.00	-13.67
	TARGET		84.67	96.67	90.67	94.67	86.67	94.67	91.33	92.67	88.00	92.00	**86.87**
		✓	34.67	79.33	47.33	58.00	44.00	54.00	48.00	40.00	49.33	20.00	**47.47**
		Δ	-50.00	-17.34	43.34	-36.67	-42.67	-40.67	-43.33	-52.67	-38.67	-72.00	-39.40
Receler [3]	""		91.33	100.00	90.67	94.67	97.33	98.67	92.00	96.67	88.00	94.67	**94.40**
		✓	26.67	74.67	32.00	50.67	31.33	46.67	54.00	39.33	46.67	20.00	**42.20**
		Δ	-64.66	-25.33	-58.67	-44.00	-66.00	-52.00	-38.00	-57.34	-41.33	-74.67	-52.20
	"image"		72.67	84.67	82.00	74.67	68.67	82.00	80.67	78.00	68.67	80.67	**77.27**
		✓	28.00	56.00	34.00	42.00	22.00	40.67	48.00	39.33	32.00	18.00	**36.00**
		Δ	-44.67	-28.67	-48.00	-32.67	-46.67	-41.33	-32.67	-38.67	-36.67	-62.67	-41.27
	"object"		67.33	72.67	54.00	50.00	48.00	38.67	77.33	42.67	63.33	54.67	**56.87**
		✓	32.67	61.33	22.00	28.67	21.33	27.33	59.33	32.00	41.33	26.00	**35.20**
		Δ	-34.66	-11.34	-32.00	-21.33	-26.67	-11.34	-18.00	-10.67	-22.00	-28.67	-21.67
	TARGET		61.33	91.33	60.67	70.00	26.67	62.00	91.33	92.67	88.00	56.00	**70.00**
		✓	21.33	80.67	16.00	40.67	17.33	26.67	48.00	40.00	49.33	20.67	**36.07**
		Δ	-40.00	-10.66	-44.67	-29.33	-9.34	-35.33	-43.33	-52.67	-38.67	-35.33	-33.93

Table 6. Effect of IRECE on per-class CRR (%) under the **white-box** latent inversion setting. Each block reports results for concept erasure methods without IRECE, with IRECE, and the corresponding change Δ (with IRECE minus without IRECE). A more negative Δ indicates stronger suppression of the target concept.

Methods	Prompt	IRECE	*airplane*	*automobile*	*bird*	*cat*	*deer*	*dog*	*frog*	*horse*	*ship*	*truck*	**Avg.**
ESD [1]	""		71.33	81.33	56.00	48.67	34.00	58.00	81.33	58.00	54.67	30.00	**57.33**
		✓	23.33	62.00	20.00	30.00	11.33	32.00	43.33	23.33	39.33	5.33	**29.00**
		Δ	-48.00	-19.33	-36.00	-18.67	-22.67	-26.00	-38.00	-34.67	-15.34	-24.67	-28.33
	"image"		51.33	73.33	24.67	34.67	16.00	40.00	62.00	29.33	39.33	36.00	**40.67**
		✓	20.67	51.33	14.00	20.00	14.00	18.00	36.67	22.00	30.67	6.67	**23.40**
		Δ	-30.66	-22.00	-10.67	-14.67	-2.00	-22.00	-25.33	-7.33	-8.66	-29.33	-17.27
	"object"		60.67	83.33	37.33	33.33	32.00	46.00	80.00	43.33	58.67	56.67	**53.13**
		✓	30.67	68.67	24.67	28.00	24.67	24.00	55.33	42.67	36.67	24.67	**35.74**
		Δ	-30.00	-14.66	-12.66	-5.33	-7.33	-22.00	-24.67	-0.66	-22.00	-32.00	-17.39
	TARGET		46.67	74.00	16.67	32.00	35.33	34.00	57.33	27.33	32.67	38.67	**39.47**
		✓	34.00	68.67	16.00	25.33	25.33	28.00	49.33	25.33	28.67	20.00	**32.07**
		Δ	-12.67	-5.33	-0.67	-6.67	-10.00	-6.00	-8.00	-2.00	-4.00	-18.67	-7.40
UCE [2]	""		91.33	98.00	92.67	96.67	98.00	98.67	93.33	97.33	90.00	96.00	**95.20**
		✓	37.33	84.00	46.00	68.67	47.33	70.67	65.33	44.00	56.00	28.00	**54.73**
		Δ	-54.00	-14.00	-46.67	-28.00	-50.67	-28.00	-28.00	-53.33	-34.00	-68.00	-40.47
	"image"		54.00	70.00	32.67	60.67	54.67	41.33	75.33	43.33	34.67	92.00	**58.87**
		✓	16.67	45.33	23.33	40.00	21.33	26.00	51.33	35.33	31.33	34.00	**33.40**
		Δ	-37.33	-24.67	-9.34	-20.67	-33.34	-15.33	-24.00	-8.00	-3.34	-58.00	-25.47
	"object"		31.33	74.00	20.67	8.00	15.33	21.33	63.33	30.00	37.33	68.00	**36.93**
		✓	13.33	65.33	12.67	7.33	8.00	13.33	50.00	28.67	32.00	29.33	**26.00**
		Δ	-18.00	-8.67	-8.00	-0.67	-7.33	-8.00	-13.33	-1.33	-5.33	-38.67	-10.93
	TARGET		40.67	74.00	11.33	28.00	30.00	30.00	50.00	25.33	24.67	33.33	**34.73**
		✓	26.00	63.33	11.33	14.67	12.67	23.33	44.00	18.67	24.00	19.33	**25.73**
		Δ	-14.67	-10.67	0.00	-13.33	-17.33	-6.67	-6.00	-6.66	-0.67	-14.00	-9.00
Receler [3]	""		84.00	87.33	82.00	68.00	66.67	82.67	84.00	84.67	77.33	73.33	**79.00**
		✓	24.00	62.67	39.33	43.33	27.33	43.33	44.00	40.00	42.00	13.33	**37.93**
		Δ	-60.00	-24.66	-42.67	-24.67	-39.34	-39.34	-40.00	-44.67	-35.33	-60.00	-41.07
	"image"		43.33	73.33	36.67	40.00	26.00	50.00	56.67	43.33	27.33	51.33	**44.80**
		✓	16.00	52.00	20.00	20.00	12.67	30.67	37.33	30.67	26.67	13.33	**25.93**
		Δ	-27.33	-21.33	-16.67	-20.00	-13.33	-19.33	-19.34	-12.66	-0.66	-38.00	-18.87
	"object"		70.00	66.67	33.33	33.33	46.00	48.00	71.33	40.67	56.67	78.00	**54.50**
		✓	36.67	60.00	23.33	26.67	30.00	32.00	57.33	30.00	32.00	32.67	**36.07**
		Δ	-33.33	-6.67	-10.00	-6.66	-16.00	-16.00	-14.00	-10.67	-24.67	-45.33	-18.43
	TARGET		28.67	69.33	12.00	17.33	23.33	12.00	50.67	23.33	24.67	26.00	**28.00**
		✓	14.67	60.00	7.33	12.67	11.33	5.33	53.33	21.33	24.00	12.00	**21.07**
		Δ	-14.00	-9.33	-4.67	-4.66	-12.00	-6.67	+2.66	-2.00	-0.67	-14.00	-6.93

Table 7. Effect of IRECE on per-class CRR (%) under the **black-box** latent inversion setting. Each block reports results for concept erasure methods without IRECE, with IRECE, and the corresponding change Δ (with IRECE minus without IRECE). A more negative Δ indicates stronger suppression of the target concept.

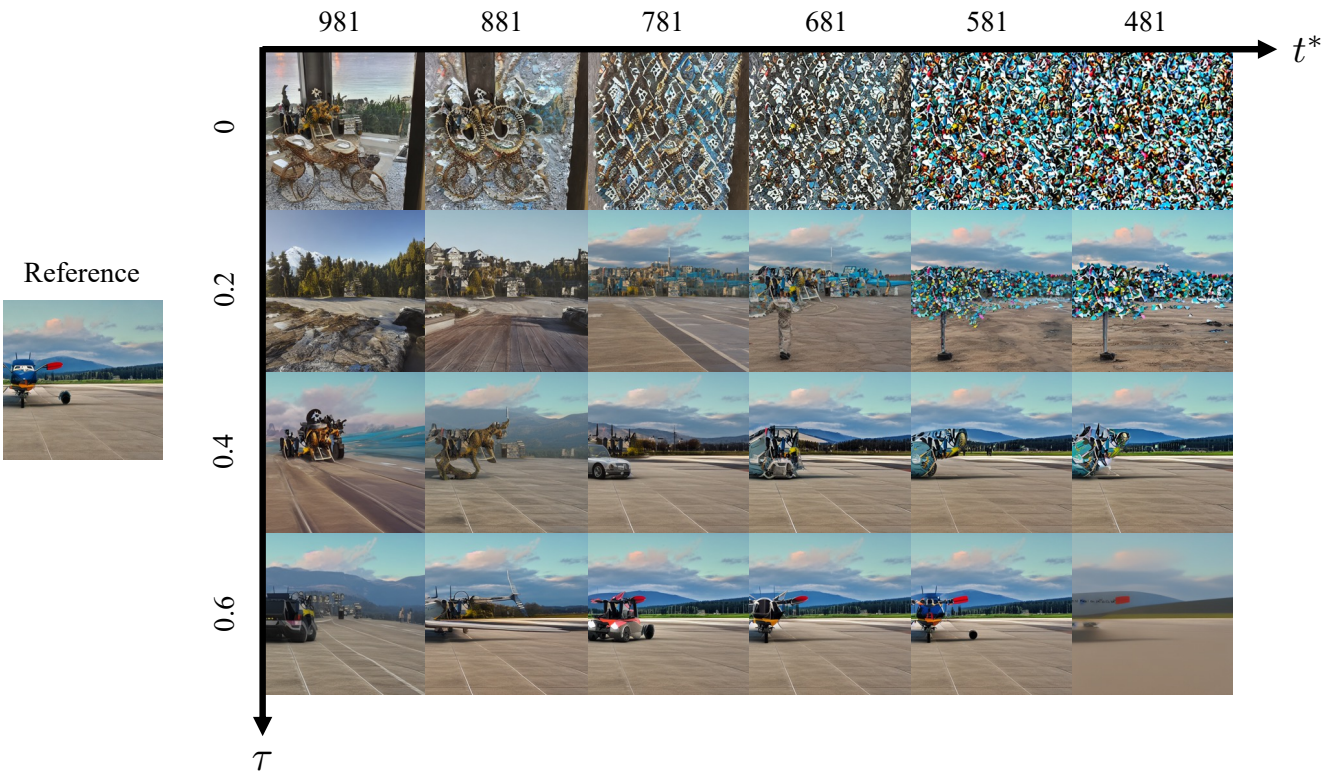

Figure 3. **Effect of intervention timestep and concept localization threshold on IRECE.** Columns correspond to intervention timesteps t^* (decreasing left to right), and rows to concept localization thresholds τ (increasing top to bottom).