LOCALIZED CONCEPT ERASURE IN TEXT-TO-IMAGE DIFFUSION MODELS VIA HIGH-LEVEL REPRESENTATION MISDIRECTION

Uichan Lee*, Jeonghyeon Kim*, Sangheum Hwang†

Department of Data Science, Seoul National University of Science and Technology

*Equal contribution

†Corresponding Author

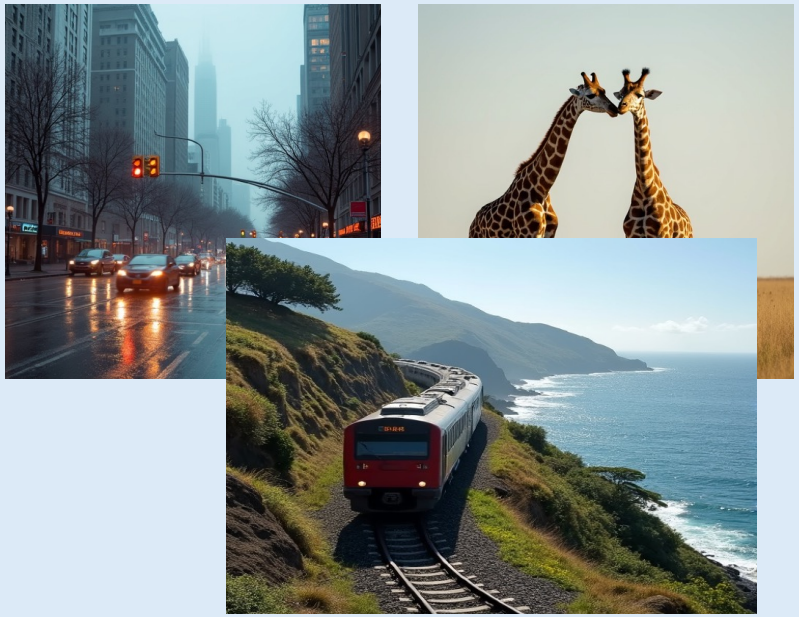
DAINTLAB

서울과학기술대학교 SEOUL NATIONAL UNIV OF SCIENCE & TECHNOLOGY 1910

ICLR

Problem Statement: The Safety Challenge

Generative Capabilities



Text-to-Image models (e.g., Stable Diffusion and Flux) have demonstrated remarkable performance in generating photorealistic images.

Critical Risks

⚠ **Copyright Infringement :** Mimicking specific artistic styles or IP.

⚠ **NSFW / Harmful Content :** Generating nudity, violence, or sexually explicit material

⚠ **Privacy Violations :** Reproducing private individuals' likenesses

Goal : Remove specific concepts without retraining the entire model.

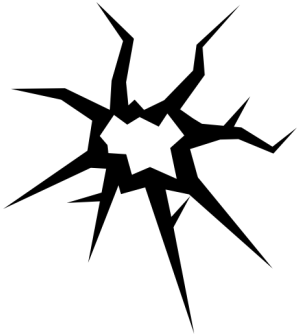
Current methods still have challenges

U-net Fine-Tuning (e.g., Mace, Salun)

Pros:
- **Effective erasure**

Cons:
- **Computationally expensive**
- Often degrades non-target images

Text Encoder Editing (e.g., Diff-QuickFix)

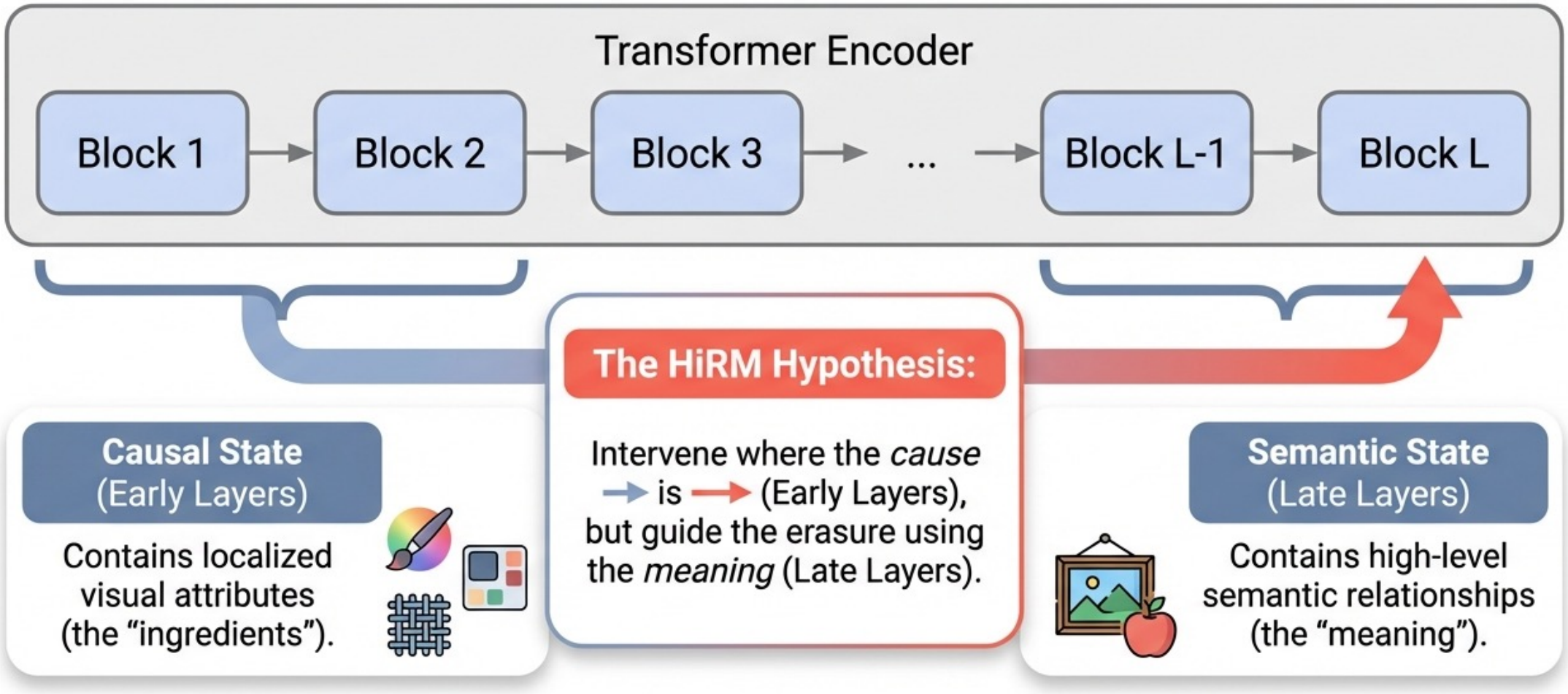


Pros: **Fast (modifies early layer)**

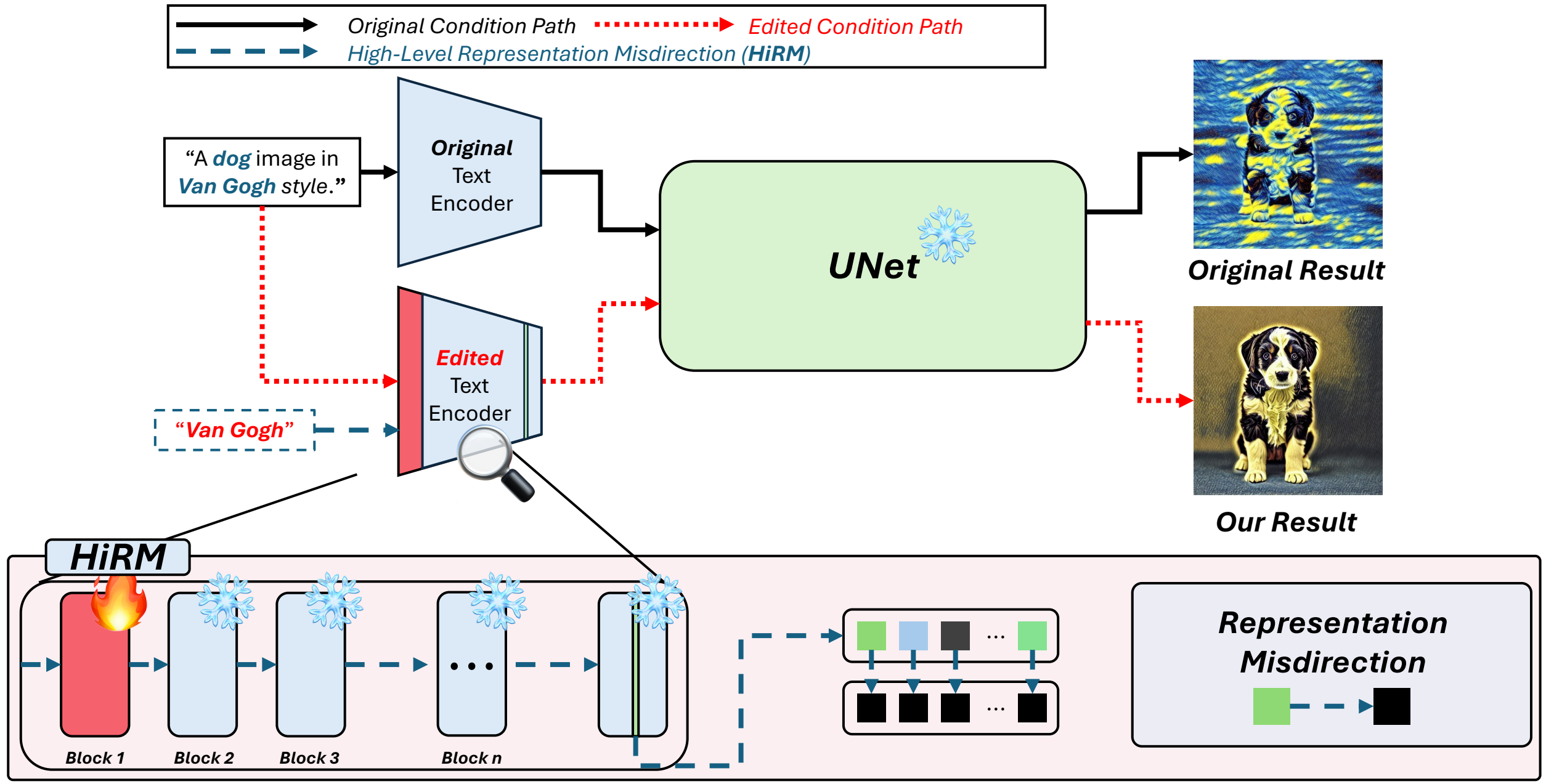
Cons:
- **Challenging to erase concepts more abstract** than artistic styles or objects (e.g., nudity)

Achieving both precision and efficiency in concept erasure
without degrading non-target outputs remains a key challenge.

Key Insight : Concept Localization & Semantic Integration



Method: High-Level Representation Misdirection (HiRM)



Method: High-Level Representation Misdirection (HiRM)

Erasure strategies: HiRM-R and HiRM-S

Misdirecting Vectors ■

Representation Misdirection ■ - - ▸ ■

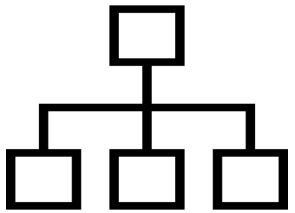
Loss Function

$$\mathcal{L} = \| \blacksquare - c \cdot \blacksquare \|$$

HiRM-R (Random)

- Steers target concept **toward a random vector**
- **Concept-agnostic vectors**

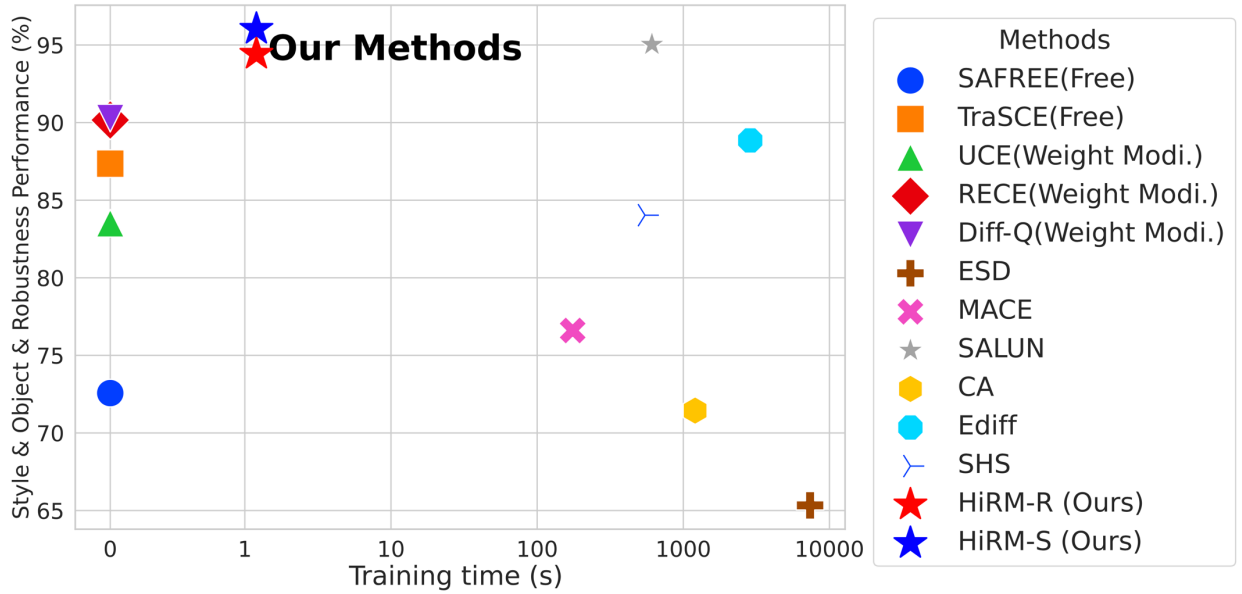
HiRM-S (Semantic)



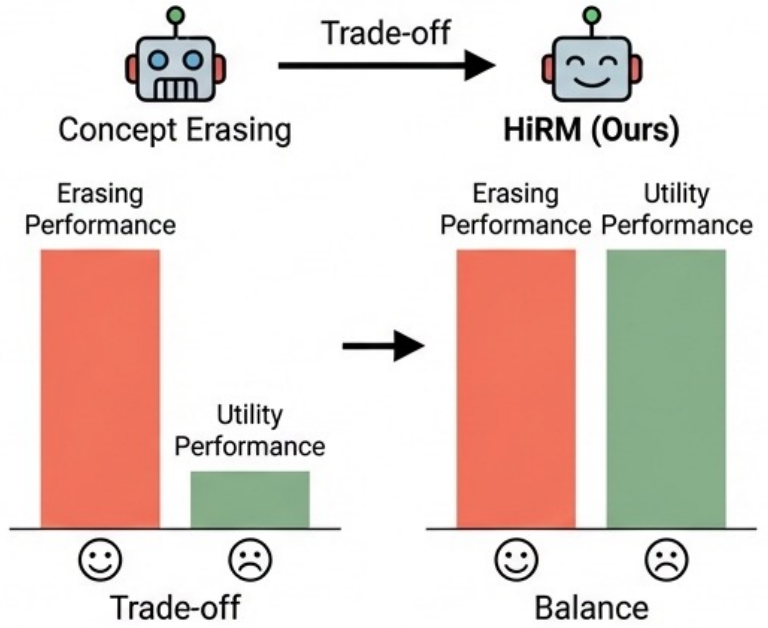
- Steer target concept **toward a "Super-category"** (e.g., "Van Gogh" to "**Painting**")

Experiments : Concept Erasing Performance across Diverse Categories

Training time vs. Performance

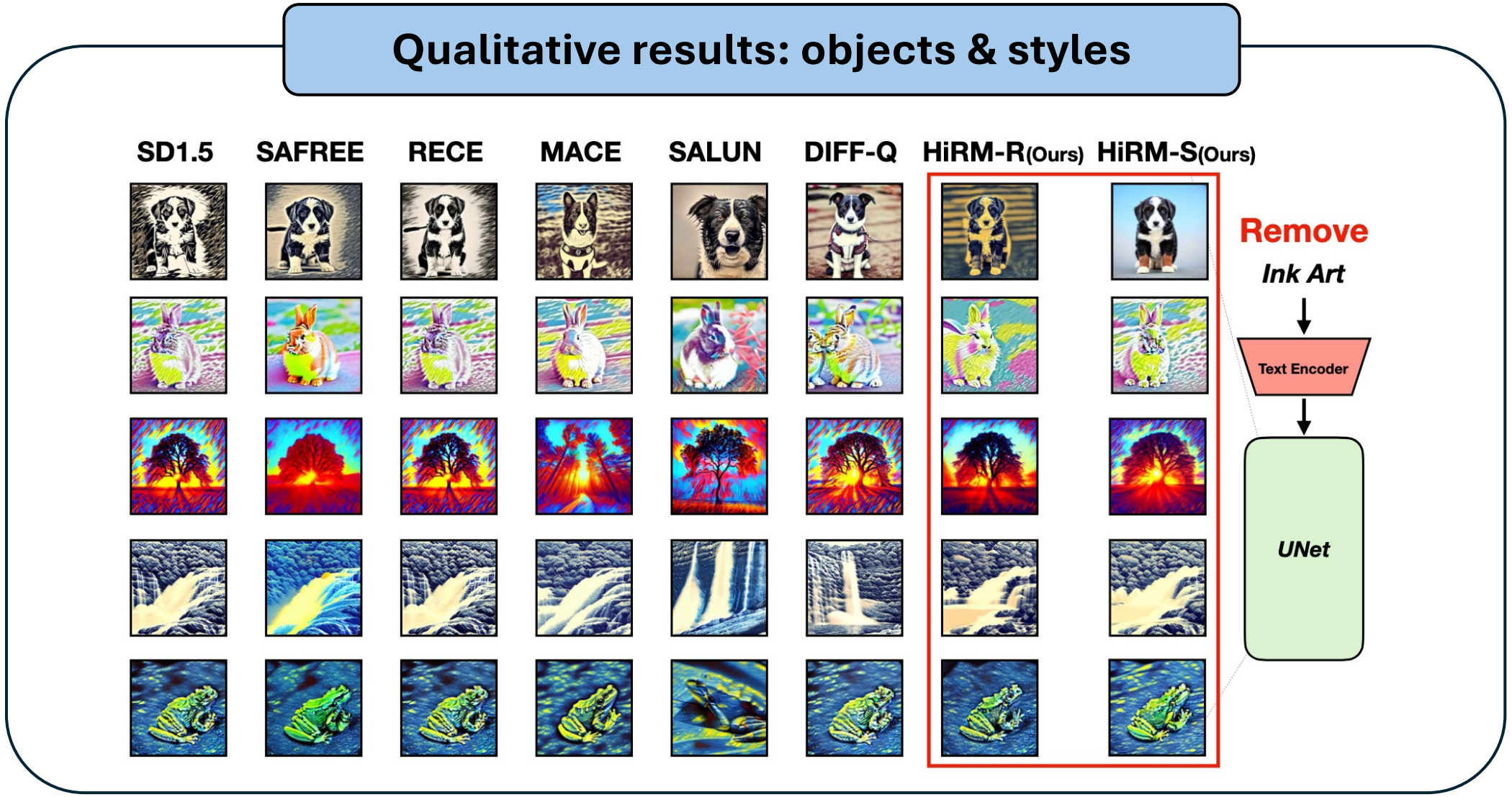


Erasing vs. Utility balance



- HiRM demonstrates the "Pareto Optimal" trade-off: High Erasing Performance, High Utility Performance, and Strong robustness against adversarial attacks.

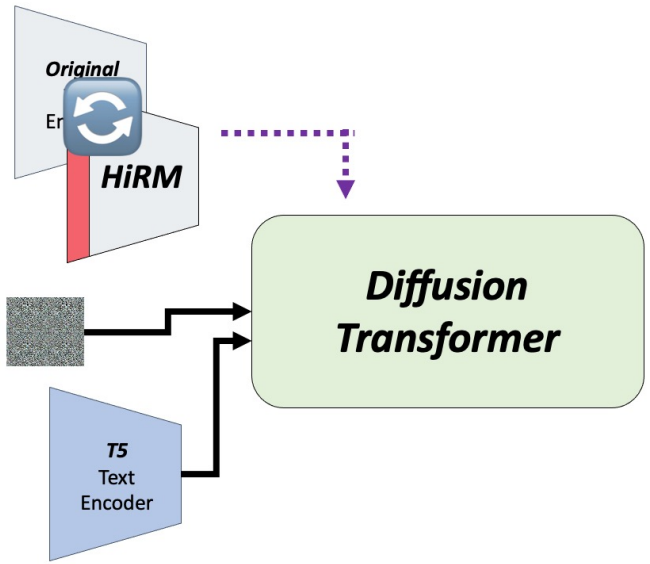
Experiments



- Qualitative comparison of concept erasure methods on the removal of the Ink Art style from the UnlearnCanvas benchmark.

Experiments: Transferability in State-of-Art Diffusion Transformer Model

Transferability in Flux1.dev

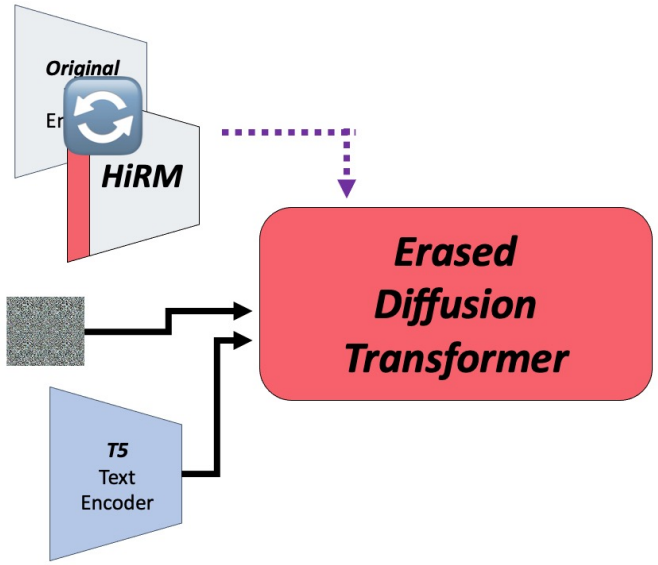


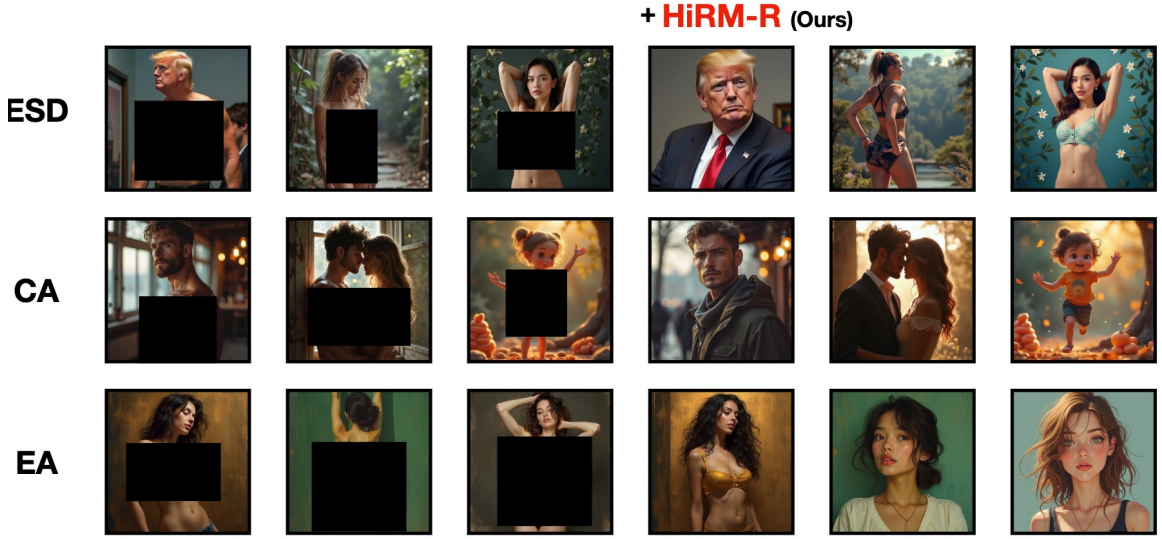
Method	Ring-16 ↓	Ring-38 ↓	Ring-77 ↓	MMA ↓	I2P ↓	COCO-1k (CLIP) ↑
Flux1.dev (Labs, 2024)	88.42	87.37	87.37	25.40	4.49	0.308
ESD (Gandikota et al., 2023)	41.05	33.68	32.63	7.10	2.93	0.307
CA (Kumari et al., 2023)	**11.58**	**11.58**	**6.32**	**3.20**	**1.83**	0.302
UCE (Gandikota et al., 2024)	65.26	66.32	62.11	9.30	3.27	**0.309**
EraseAnything (Gao et al., 2025)	29.47	24.21	26.32	6.60	2.64	0.305
Diff-Q (Basu et al., 2023)	57.89	51.58	64.21	12.20	4.44	0.306
HiRM-R (Ours)	37.89	38.95	44.21	9.70	3.23	0.308

- Qualitative & Quantitative comparison of concept erasure methods on Flux1.dev using the Adversarial benchmark

Experiments: Synergistic Effects

Synergistic effects with other methods



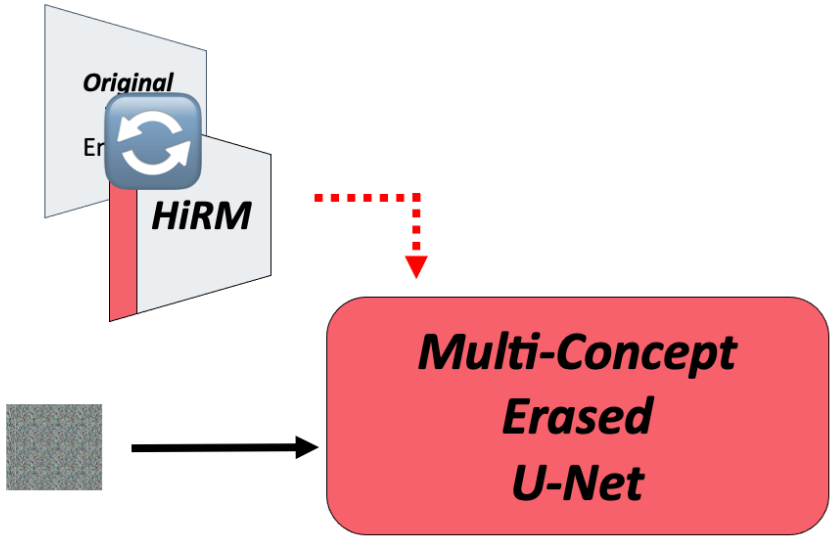


Method	Ring-16 ↓	Ring-38 ↓	Ring-77 ↓	MMA ↓	I2P ↓	COCO-1k (CLIP) ↑
ESD (Gandikota et al., 2023)	41.05	33.68	32.63	7.10	2.93	**0.307**
+ HiRM-R (Ours)	**12.63**	**7.37**	**4.21**	**3.30**	**2.51**	0.306
CA (Kumari et al., 2023)	11.58	11.58	6.32	**3.20**	1.83	**0.302**
+ HiRM-R (Ours)	**3.16**	**2.11**	**2.11**	4.40	**1.11**	**0.302**
EraseAnything (Gao et al., 2025)	29.47	24.21	26.32	6.60	2.64	**0.305**
+ HiRM-R (Ours)	**3.16**	**1.05**	**3.16**	**2.50**	**1.57**	**0.305**

- Synergistic effects of combining HiRM with denoiser-based concept erasing methods on the Flux1.dev architecture

Experiments: Synergistic Effects

Synergistic effects with other methods



Method	Adversarial Robustness					Multi Concept		COCO-10k	
	Ring-16↓	Ring-38↓	Ring-77↓	MMA↓	I2P↓	Acc_e↓	Acc_r↑	FID↓	CLIP↑
SPEED (Li et al., 2025)	-	-	-	-	-	**3.46**	**88.48**	-	-
HiRM-S (Ours)	**1.05**	**1.05**	**0.00**	3.30	0.66	-	-	**6.75**	**0.306**
S-HiRM-S (w/ Ours)	**1.05**	**1.05**	2.11	**1.70**	**0.43**	3.64	79.30	7.43	0.304

- Synergistic effects of combining HiRM (text encoder) and SPEED (U-Net) for adversarial robustness and multi-concept erasure

Summary & Contributions

HiRM offers a lightweight, effective ‘Safety Patch’ for Text-to-Image Diffusion models

Performance

- Demonstrates a strong trade-off between erasure effectiveness and generation quality on UnlearnCanvas, NSFW, and adversarial attack settings across diverse targets.

Synergistic Effect

- Exhibits synergistic effects as a complementary module.

Efficiency

- Performing concept erasure within the text encoder offers significant advantages in terms of efficiency. **(using only 1.6s per a concept)**