EraseAnything: Enabling Concept Erasure in Rectified Flow Transformers

EraseAnything: Making AI Models Forget Unwanted Stuff!

作者: 高岱恒

时间: 2025年5月26日

https://tomguluson92.github.io/projects/eraseanything/

https://github.com/tomguluson92/eraseanything

ICML

What's the Big Problem?

Amazing Creation

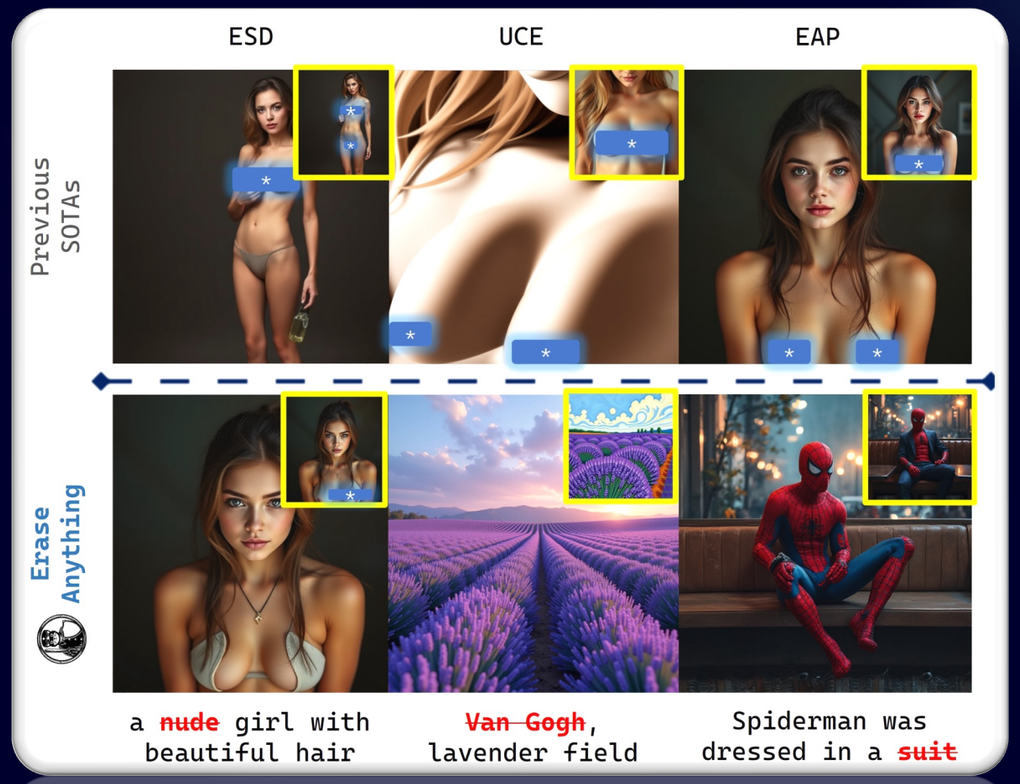
T2I AI models create pictures from words like "a cat riding a skateboard!"

Unwanted Content

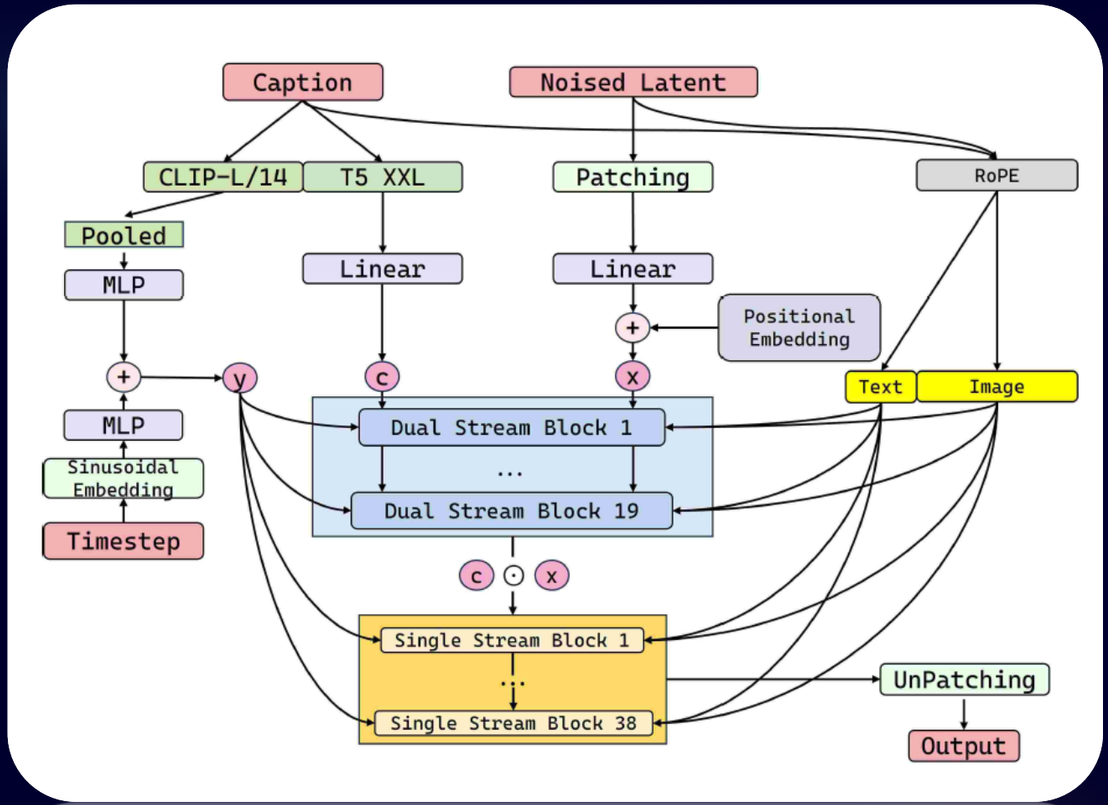
They sometimes generate harmful, inappropriate, or NSFW content.

New Architecture Challenge

Newer models like Stable Diffusion v3 and Flux use different "brains." Old forgetting tricks don't work.



Architecture Challenge



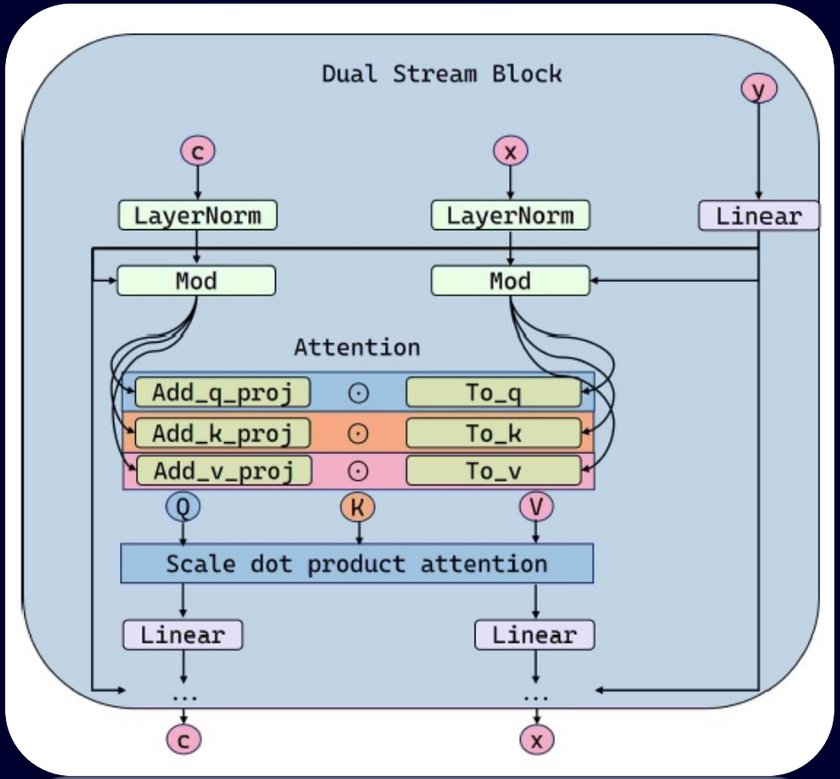


Figure 8. **Model architecture of Flux [dev].** Flux [dev] use frozen CLIP-L 14 and T5-XXL as text encoders for conditioned caption feature extraction. The coarsed CLIP embedding concatenated with timestep embedding y are used to modulation mechanism. The fine-grained T5 c concatenated with image latents x are input to a stacked of double stream blocks and single stream blocks to predict output in the VAE encoded latent space. Concatenation is indicated by $\odot$.

Introducing EraseAnything!

First of Its Kind

Specifically designed for newest T2I models like Flux.

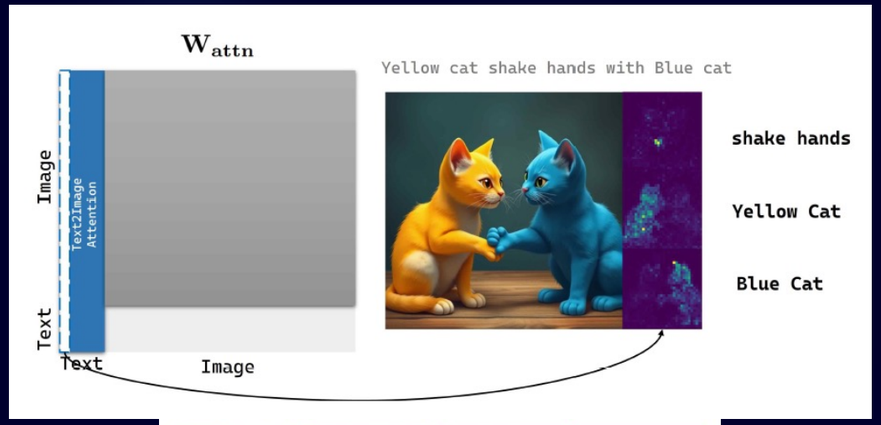
Selective Memory

Teaches AI to forget bad concepts while keeping good ones.

Versatile Solution

Works on various unwanted concepts, not just inappropriate content.

METHOD	$ACC_e \downarrow$	$ACC_{ir} \uparrow$	$ACC_g \downarrow$
CA (ENTITY)	14.8	89.2	27.3
CA (ABSTRACTION)	25.2	88.3	29.6
CA (RELATIONSHIP)	22.7	88.6	23.1
OURS (ENTITY)	12.5	91.7	18.6
OURS (ABSTRACTION)	21.1	90.5	24.7
OURS (RELATIONSHIP)	18.4	90.2	19.3



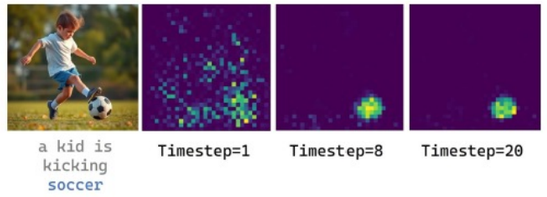


Figure 2. Correlations between text and attention maps.

How Does it Work?

Level 1: Getting Rid of Bad Stuff

We tell the model to reduce "thinking" about unwanted things.

Uses attention map regularizer to pinpoint and suppress bad "thoughts".

Level 2: Keeping the Good Stuff

Ensures model still creates great images of everything else.

Uses self-contrastive learning to maintain artistic skills.

Why is EraseAnything Special?

Smart Attention Localization

Precisely identifies where concepts exist in AI's "brain".

Reverse Self-Contrastive Loss

Makes unwanted concept less similar to everything else.

Pioneer for New Models

First effective solution for Flux models.

Universal Application

Works for objects, styles, and relationships!

Reverse Self-Contrastive Loss (BO)

Algorithm 1 BO formulation in EraseAnything

Input: unlearned concept dataset and irrelevant dataset D_{un} and D_{ir}, learning rates $\alpha_{low}, \alpha_{up}$, total iteration steps M.

for $iteration = 1$ **to** M **do**

  for c_{un} sampled from D_{un} **do**

    PREPARATION

    ❶ Construct a meaningful sentence c involve c_{un}.

    ❷ Shuffle c to avoid overfitting.

    ❸ Find tokenized index idx_{start} : idx_{end} of c_{un} from c.

    LOWER LEVEL: c_{un} ERASURE

    ❹ Update LoRA $\Delta\theta$ with Eq. (2)+Eq. (3) under α_{low}.

    UPPER LEVEL: c_{ir} PRESERVING

    ❺ Retrieve c_{ir}, c_{syn} *w.r.t* to c_{un} and replace them into c separately to have $F^{ir,syn}$.

    ❻ Update LoRA $\Delta\theta$ with Eq. (4)+Eq. (5) under α_{up}.

  end for

end for

Reverse **S**elf **C**ontrastive loss (**RSC**): our training goal is to align the central feature F^{un} with the dynamically shifting F^{ir}, while simultaneously pushing them apart from the synonym feature F^{syn}. The strategy here is to deviate from the conventional self-contrastive learning approach, which would typically aim to make F^{un} more akin to F^{syn}, thereby enhancing the model's sensitivity to the term slated for removal. By inverting this approach, we aim to steer the network towards gradually discarding the concept of "nude" during learning, effectively obfuscating it within an array of irrelevant concepts.

$$\mathcal{L}_{rsc} = \log\left(\frac{\sum_{i=0}^{K}\exp\left(\frac{F^{un}\cdot F^{k_i}}{\tau}\right)}{\exp\left(\frac{F^{un}\cdot F^{syn}}{\tau}\right)}\right). \tag{5}$$

$$\begin{aligned} &\min \mathcal{L}_{lora+rsc}(\Delta^*\theta; D_{ir}) \\ s.t. \quad &\Delta^*\theta = \min \mathcal{L}_{esd+attn}(\Delta\theta; D_{un}) \end{aligned} \tag{6}$$

F_{un}, F_{syn}, F_...

Table 6. AI Agent template in generating c_{ir} (c_{un} = "nude").

Role	Content
System	*'You are a helpful assistant and a well-established language expert'*
User	Hello, please return K (K=3) English words that you think with Human intuition are **no_relation/far/mid** in the semantic space from the English word: c_{un}, and only reply the result with JSON format is as follows: {"**no_relation**": [(word1, similarity_score1), ...], "**far**": [(word1, similarity_score1), ...], "**mid**": [(word1, similarity_score1), ...]}
Response	{"**no_relation**": [("cloud", 0.1), ("tree", 0.2), ("carpet", 0.1)], "**far**": [("hot", 0.3), ("color", 0.4), ("wet", 0.3)], "**mid**": [("image", 0.5), ("figure", 0.6), ("portrait", 0.5)]}

Table 1. Find the closest synonyms of **nude**.

METHOD	TOP-3 CLOSEST SYNONYMS
CLAUDE 3.5	"NAKED", "UNDRESSED", "UNCLOTHED"
GPT-4O	"BARE", "NAKED", "UNCLOTHED",
KIMI	"NAKED", "UNCLOTHED", "BARE"
T5 FEATURE	"LEAN", "DEER", "GIRL"

Heuristic >> Cosine Similarity !!!

Erasing & Keeping

$$\mathcal{L}_{esd} = \mathbb{E}\Big[v_{\theta_o+\Delta\theta}(x_t, c_{un}, t) - \eta \left\| v_{\theta_o}(x_t, c_{un}, t) - v_{\theta_o}(x_t, \emptyset, t) \right\|_2^2 \Big], \tag{2}$$

$$\mathcal{L}_{attn} = \sum_{idx=start}^{end} F_{idx}^{un}. \tag{3}$$

$$\mathcal{L}_{lora} = \mathbb{E}\left[\left\| v - v_{\theta+\Delta\theta}(u_t, c, t) \right\|_2^2\right], \tag{4}$$

$$\mathcal{L}_{rsc} = \log\left(\frac{\sum_{i=0}^{K} \exp\left(\frac{F^{un}.F^{k_i}}{\tau}\right)}{\exp\left(\frac{F^{un}.F^{syn}}{\tau}\right)}\right). \tag{5}$$

Traditional Methods vs. EraseAnything

Old Models (SD)

- Used "U-Net" with "cross-attention"
- Easy to manipulate for erasure

New Models (Flux)

- Uses transformer-based components
- No explicit "cross-attention"
- Old methods leave "concept residue"

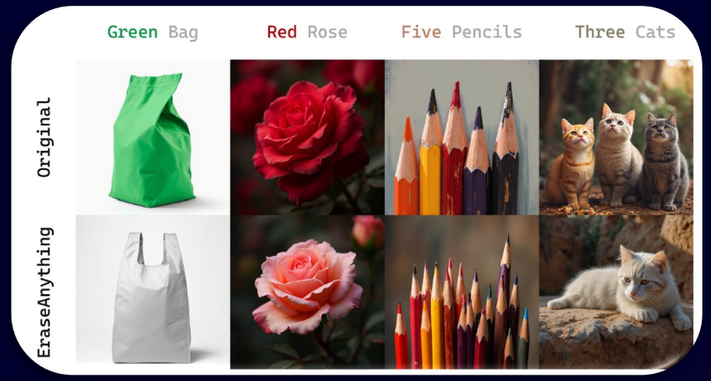
Our Solution

- Remove unwanted but Keep irrelevant
- Handles similar words like "nude" and "naked"
- Clean removal without residue

Seeing is Believing!

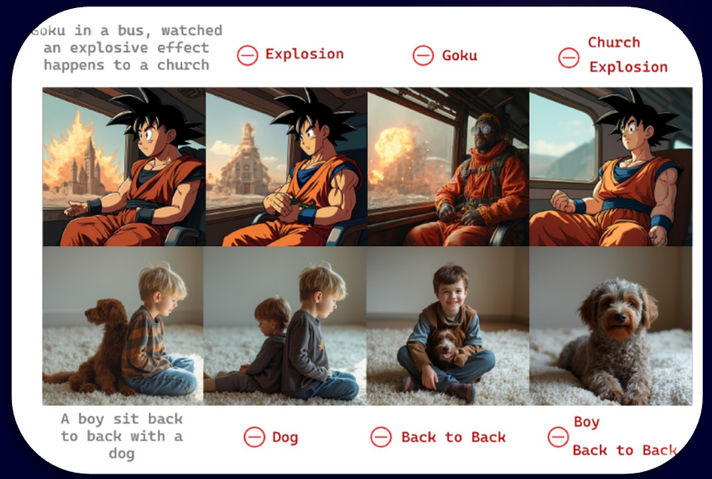
Complex Erasure

EraseAnything could erase quantity and color, which is superb than previous methods.



Multi-Concept Erasure

Successfully erases multi-concept in a same inference.



LoRA Disentanglement

Involve EraseAnything would not do harm to irrelevant concept generation!

How Good is it, Really?

Table 2. Assessment of Nudity Removal: (Left) Quantity of explicit content detected using the NudeNet detector on the I2P benchmark. (Right) Comparison of FID and CLIP on MS-COCO. The performance of the original Flux [dev] is presented for reference.

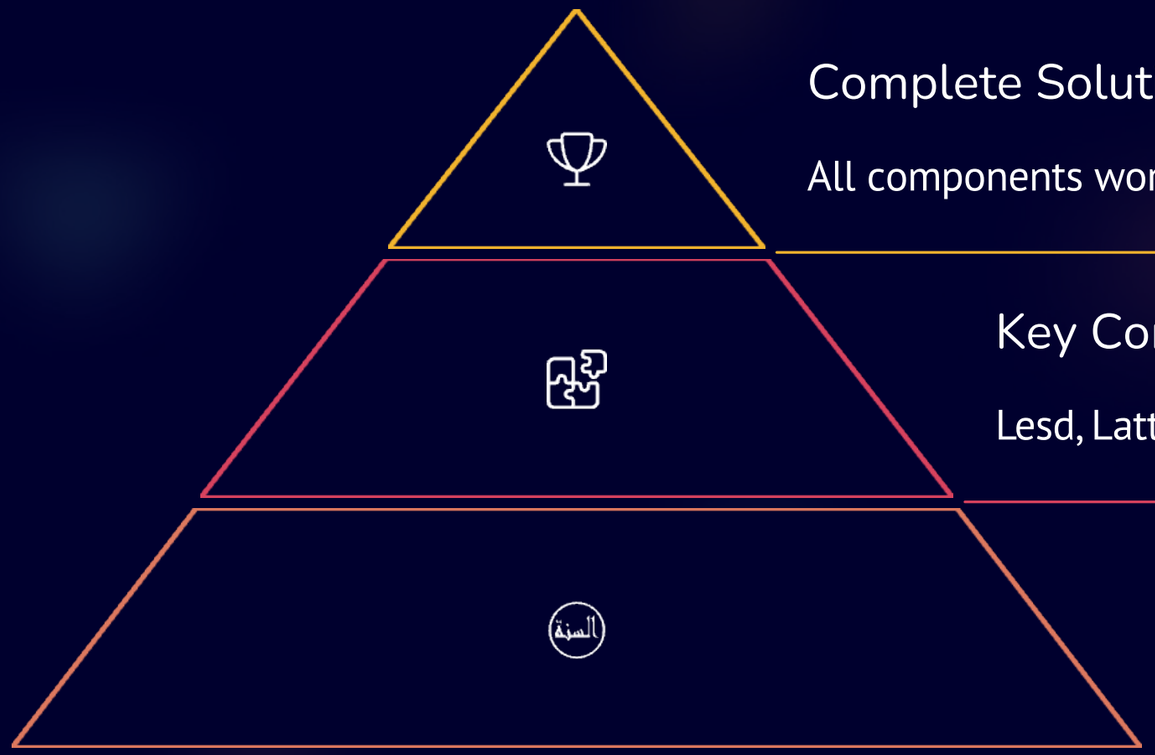
Method	Detected Nudity (Quantity)				MS-COCO 10K	
	Common	Female	Male	Total↓	FID↓	CLIP↑
CA (Model-based) (Kumari et al., 2023)	253	65	26	344	22.66	29.05
CA (Noise-based) (Kumari et al., 2023)	290	72	28	390	23.07	28.73
ESD (Gandikota et al., 2023)	329	145	32	506	23.08	28.44
UCE (Gandikota et al., 2024)	122	39	12	173	30.71	24.56
MACE (Lu et al., 2024a)	173	55	28	256	24.15	29.52
EAP (Bui et al., 2024)	287	86	13	386	22.30	29.86
Meta-Unlearning (Gao et al., 2024)	355	140	26	521	22.69	29.91
Ours	129	48	22	199	21.75	30.24
Flux.1 [dev]	406	161	38	605	21.32	30.87

Table 4. Performance Metrics of Nudity Detection Methods.

Concept	**Methods**	**Flux[dev]**	**ESD**	**CA**	**EraseAnything**
Nudity	Original (Org)	59.65%	7.36%	3.16%	2.46%
	MU-Attack (step 0)	64.56%	11.57%	15.44%	8.77%
	MU-Attack (steps 0,1,2)	65.96%	14.74%	16.49%	11.93%

EraseAnything achieves the best balance between removing unwanted content and maintaining image quality and resist attacking.

The "Why" Behind the Success



Complete Solution

All components working together for best results

Key Components

Lesd, Lattn, Llora, Lrsc loss terms

Ablation Study

Tested each component's contribution

Our study showed combining all loss terms gives the best performance in erasing concepts (lowest ACCe) while keeping irrelevant concepts intact (highest ACCir).

User Study & Conclusion

3.86/5

User Rating

Outstanding performance across all aspects

99%+

Safer AI

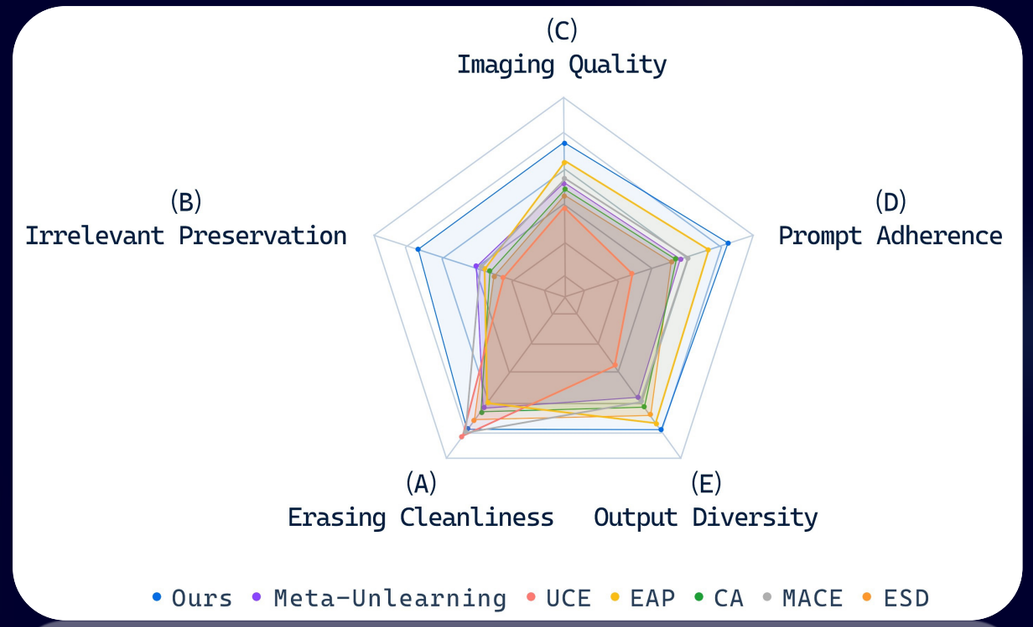
Makes powerful AI models more ethical

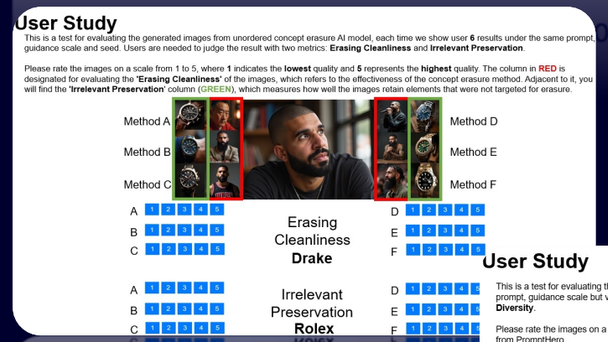
1st

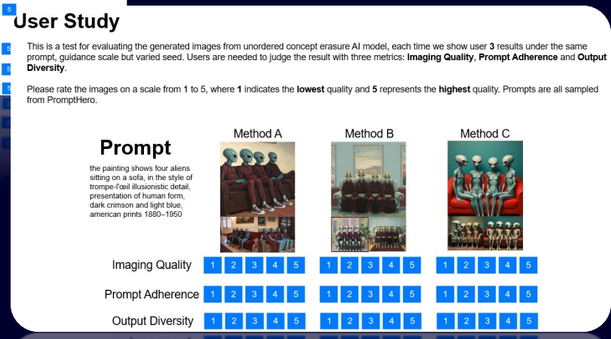
Innovation

First solution for newest T2I models

EraseAnything achieves perfect balance between removing unwanted concepts and preserving creative abilities. It's a game-changer for AI safety.

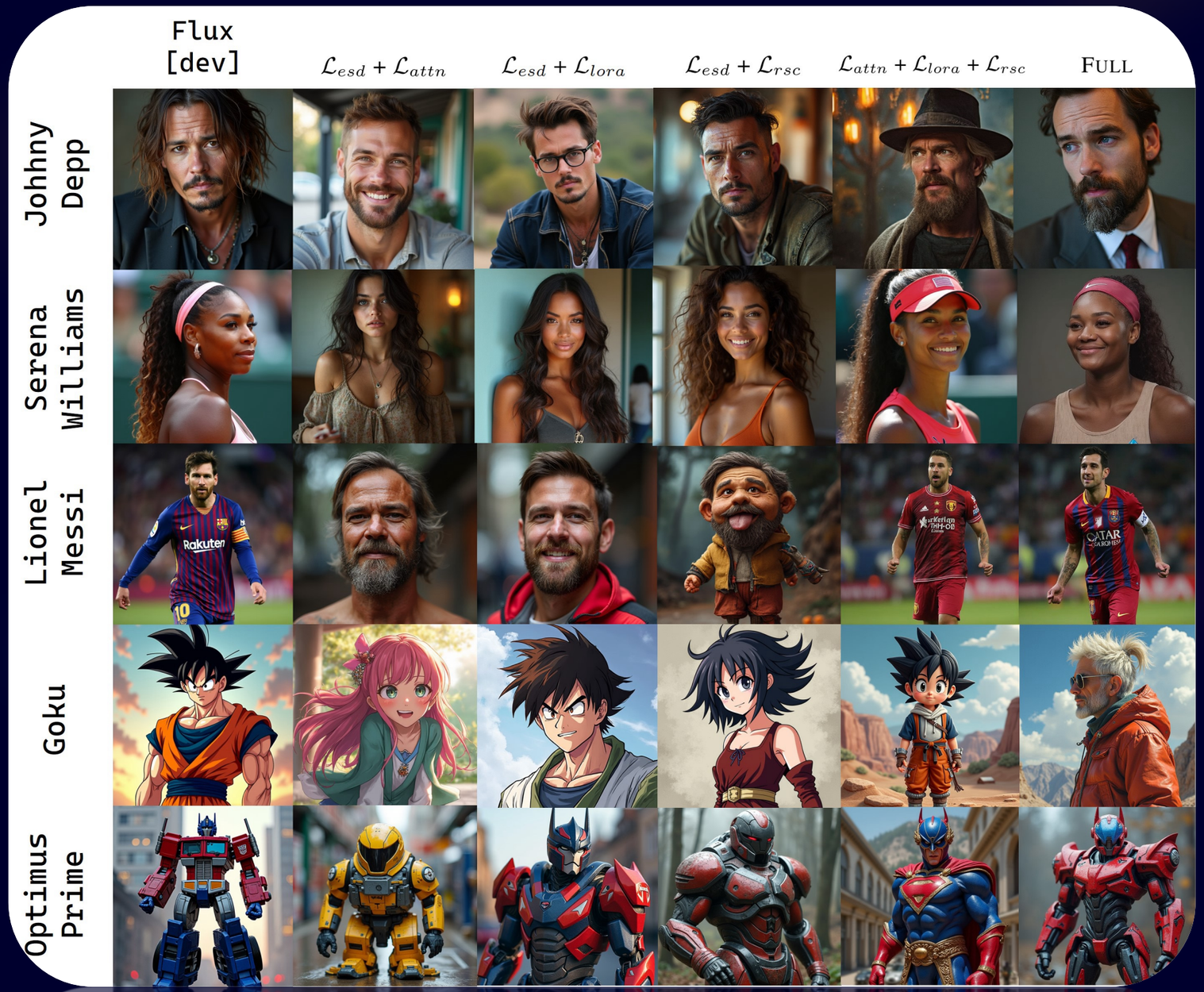






Visualization Results

Visualization Results



Thanks for reading!

samuel.gao023@gmail.com