CRCE: Coreference-Retention Concept Erasure in Text-to-Image Diffusion Models

Yuyang Xue, Edward Moroshko, Feng Chen, Jingyu Sun, Steven McDonagh, Sotirios A. Tsaftaris

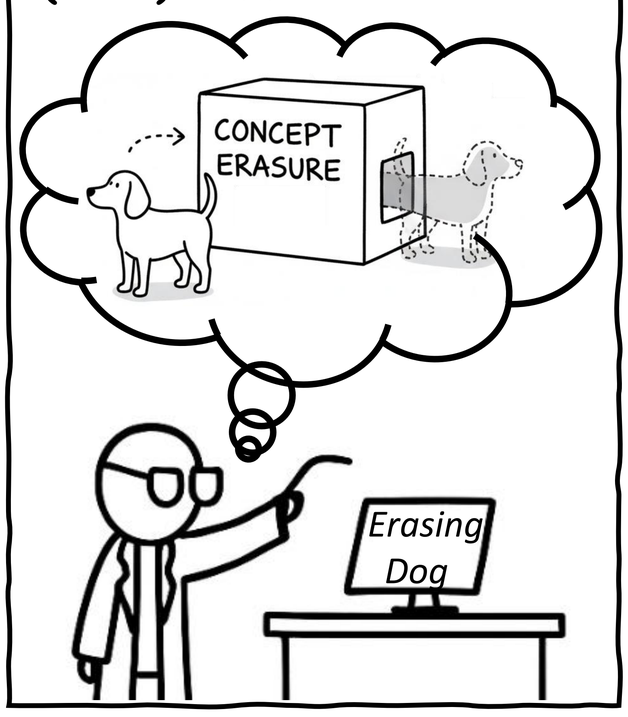
School of Engineering, University of Edinburgh, Edinburgh, UK

BMVC 2025

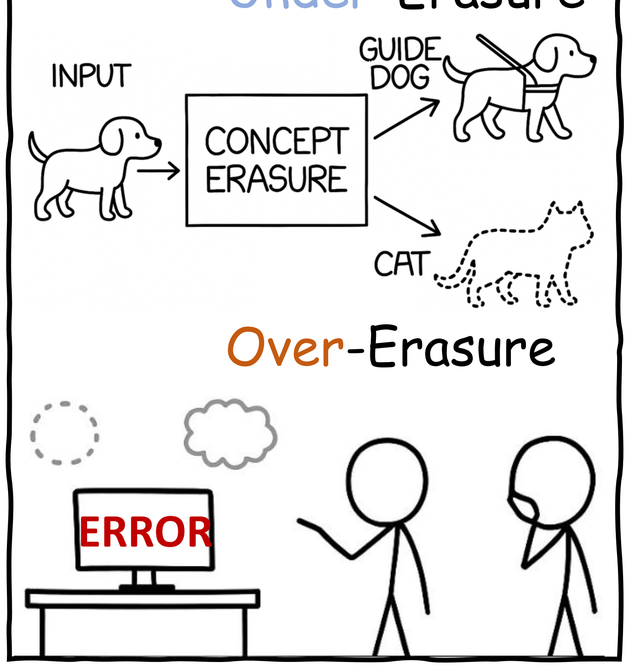
Our Code

Our Group

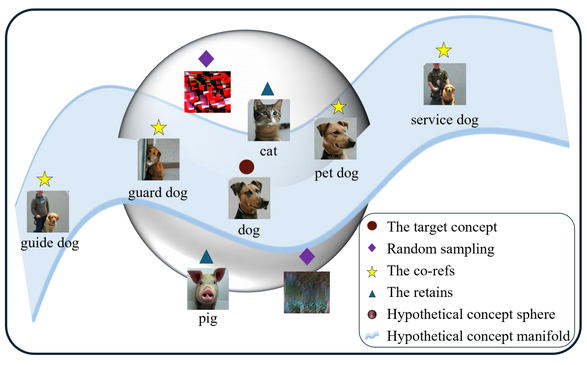
Concept Erasure can help erasing concept from Text-to-Image (T2I) diffusion models

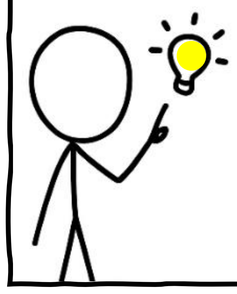


However, there are **under**-/**over**-erasure issues exist …

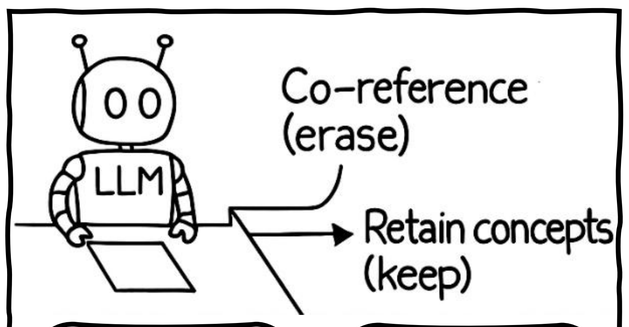


Euclidian Distance and *Clip Similarity* **do not work** in concept space…





…Wait, can we make use of **Large Language Models (LLMs)**?



Coreferences

Certainty

Normal, High, Very High, …

Retains

Certainty

Very High, High, Low,…

LLMs guide the erasure of **coreference** concept and keep **retain** with **certainty**!

🧩 Background & Motivation

- **Problem:**
  Text-to-Image (T2I) diffusion models can produce undesirable content.
  Therefore, we have **Concept Erasure**.
- **Challenges:**
  Under-erasure → residual traces remain (e.g., *Service dog* when erasing *Dog*).
  Over-erasure → unrelated but similar or close concepts removed (e.g., *Cat* erased too).
- **Observation:**
  Human semantic similarity ≠ CLIP embedding distance
  geometric proximity in CLIP space fails to capture true conceptual relations.

💡 Contributions Highlight

- **CRCE Framework:**
  First to address both *under-* and *over-erasure* by leveraging **Large Language Models (LLMs)** to infer **coreferential** (to erase) and **retained** (to preserve) concepts semantically.
- **Novel Loss Function:**
  Integrates **LLM-generated certainty-weighted** coref/retain sets to jointly balance erasure and preservation, superior *Erasure–Retention* balance.
- **CorefConcept Dataset:**
  Introduces a new dataset mapping **20 object**, **15 IP**, and **15 celebrity** concepts to corefs/retains with certainty scores.

🔬 Method and Results

➤ LLM-Guided Concept Generation

- LLM generates **coref** and **retain** lists for each target concept
- Each item assigned a certainty:
  Very High → *Very Low* (weights: 1.0–0.2)
- **Corefs**: synonyms / visually equivalent forms
- **Retains**: semantically adjacent but distinct (e.g., “cat” when erasing “dog”)

➤ Coref-Retain Loss

$$L = L_{\text{CE}} + \frac{1}{\text{M}} \sum_{c' \in \text{Coref}} C_{c'} \left\| \epsilon_\theta(c') - \epsilon_{\theta^*}^{\text{anchor}}(c) \right\|^2 + \frac{1}{N} \sum_{c' \in \text{Coref}} C_{c'} \| \epsilon_\theta(r) - \epsilon_{\theta^*}(r) \|^2$$

- L_{CE} is the original concept erasure loss function
  M and N are the number of *corefs* and *retains*
  c is the target concept, c' is its *coref*, and r is its *retain*
  C is LLM-weighted certainty
- Combined Loss + LLM-weighted *coref*/*retain* regularizers
- Balances erasure precision and knowledge preservation

➤ Evaluations

- Compared against **ESD-x, FMN, UCE, SPM, Receler, MACE, RealEra**
- Evaluated on **Object**, **IP**, and **Celebrity** concepts
- Metrics:
  ↓ **AccU** – residual of target concept
  ↓ **AccC** – residual of corefs (train/test)
  ↑ **AccR** – retention of unrelated concepts (train/test)
- Using **Qwen2-VL** Vision-Language model for robust visual semantic detection evaluation

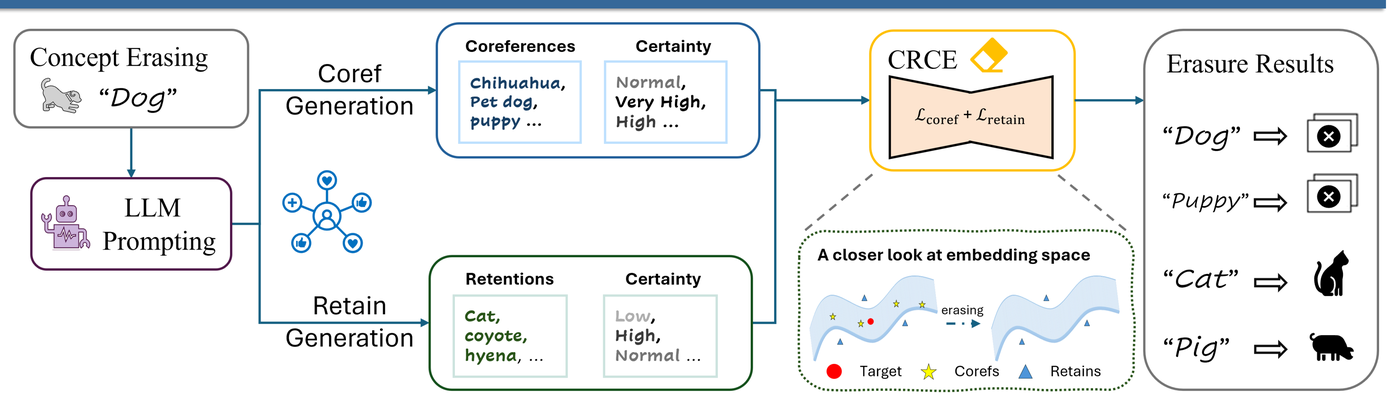


Figure 1. The overview of our pipeline.

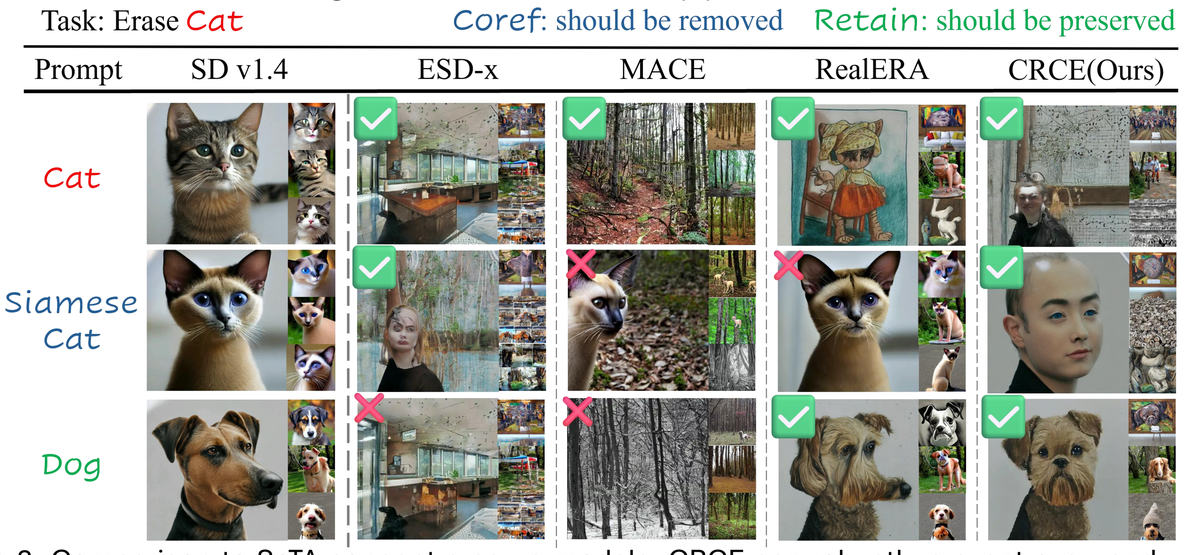


Figure 2. Comparison to SoTA concept erasure models, CRCE can robustly prevent over- and under-erasure.

	Object					Intellectual Property (IP)					Celebrity				
	$Acc_U \downarrow$	$Acc_C^{train} \downarrow$	$Acc_C^{test} \downarrow$	$Acc_R^{train} \uparrow$	$Acc_R^{test} \uparrow$	$Acc_U \downarrow$	$Acc_C^{train} \downarrow$	$Acc_C^{test} \downarrow$	$Acc_R^{train} \uparrow$	$Acc_R^{test} \uparrow$	$Acc_U \downarrow$	$Acc_C^{train} \downarrow$	$Acc_C^{test} \downarrow$	$Acc_R^{train} \uparrow$	$Acc_R^{test} \uparrow$
SD [31]	83.33	87.22	91.60	81.17	80.63	88.33	37.46	30.13	86.80	86.75	90.90	17.12	17.94	92.64	93.24
ESD-x [7]	3.33	37.72	39.92	73.82	69.89	3.33	10.67	7.73	62.80	71.16	*1.81*	11.09	10.76	72.9	68.37
FMN [41]	4.00	50.78	50.45	30.06	25.92	4.93	4.45	*3.89*	2.88	2.51	9.2	4.38	4.00	7.92	5.23
UCE [8]	1.27	8.52	8.40	17.91	14.58	*1.17*	1.38	6.57	1.01	1.02	1.47	3.14	3.03	5.83	5.00
SPM [25]	36.00	31.04	33.29	14.57	15.45	16.86	9.52	10.13	5.87	8.05	16.85	15.78	19.41	33.94	34.35
Receler [13]	2.40	**6.53**	**7.09**	11.68	8.71	0.27	**0.77**	**1.74**	0.87	0.91	0.53	*2.87*	1.76	4.11	4.96
MACE [23]	*2.10*	*12.10*	*9.05*	62.73	68.84	1.33	*3.33*	3.73	44.53	3.73	**0.00**	3.33	2.40	29.73	30.40
RealEra [22]	48.88	70.49	72.60	**88.92**	*77.15*	3.33	16.67	14.66	77.73	**81.81**	12.72	11.78	11.53	82.19	85.40
CRCE-Sphere	26.67	58.11	62.40	86.57	**84.84**	20.00	13.73	12.00	*75.20*	*79.47*	7.27	11.78	11.53	82.19	*82.70*
CRCE-Fixed	8.75	19.27	32.21	78.23	72.39	7.66	4.80	4.26	67.94	69.47	5.57	**1.36**	*2.59*	*81.54*	79.45
CRCE	**1.25**	12.81	26.31	*85.20*	79.56	**0.00**	6.27	7.46	**78.80**	80.25	*1.81*	2.04	**1.29**	**87.35**	**87.02**

Table 1. Quantitative comparison of concept erasure methods across Object, IP, and Celebrity categories.

This work was supported by the School of Engineering, the University of Edinburgh and CHAI - Causality in Healthcare AI Hub.