PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

CURE: Concept Unlearning via Orthogonal Representation Editing in Diffusion Models

Shristi Das Biswas Arani Roy Kaushik Roy



NEURAL INFORMATION PROCESSING SYSTEMS

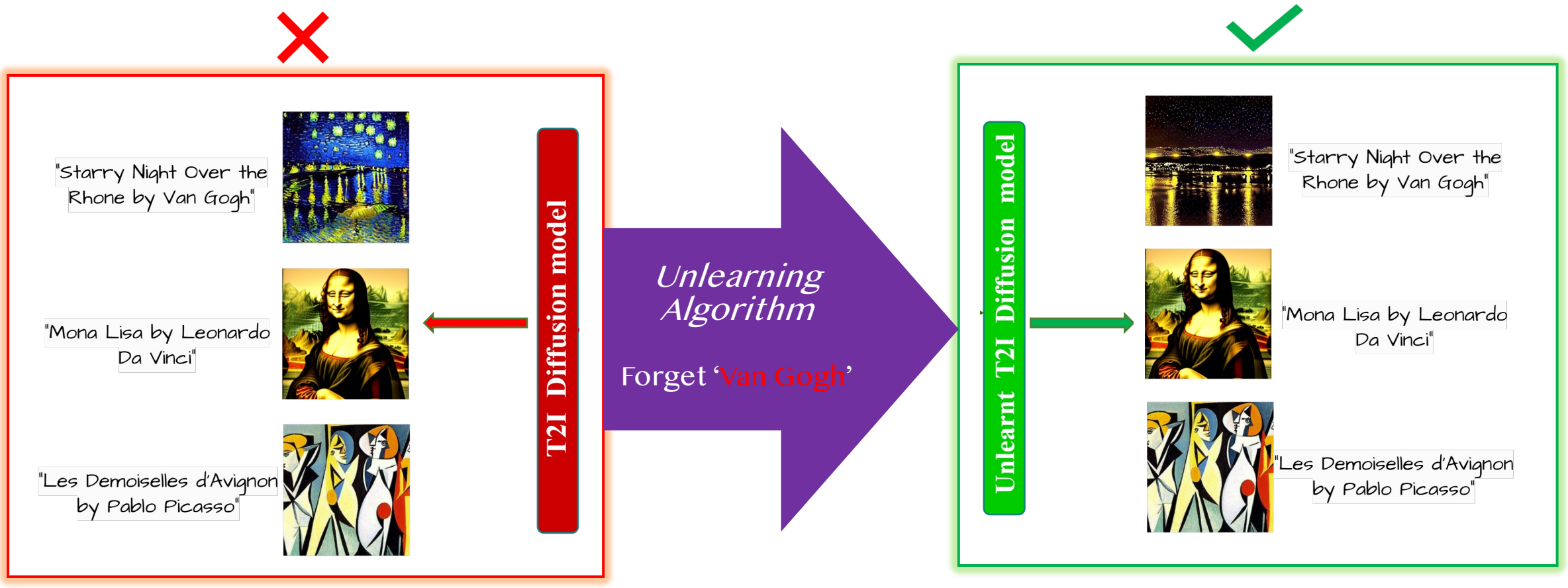
Project Page: https://sites.google.com/view/cure-unlearning/home

Contents

- Introducing the Task
- Method Overview
- Evaluations
- Ablating Insights
- Takeaways

Introducing the Task

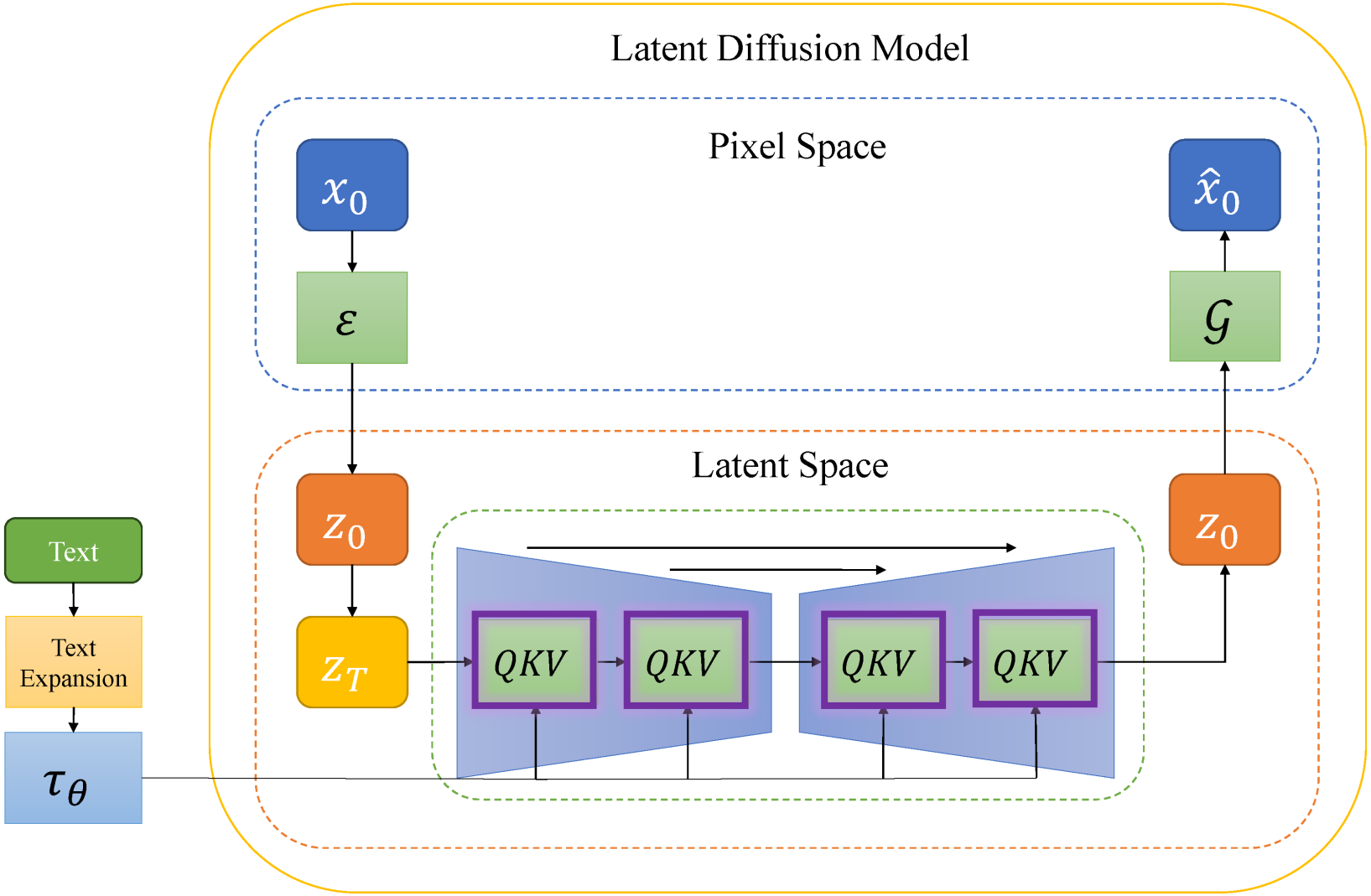
What Is The Task of Concept Unlearning?



The task of Concept Unlearning enables the unlearned model to robustly refrain from generating images with erased concept while preserving the image generation ability of unerased concepts.

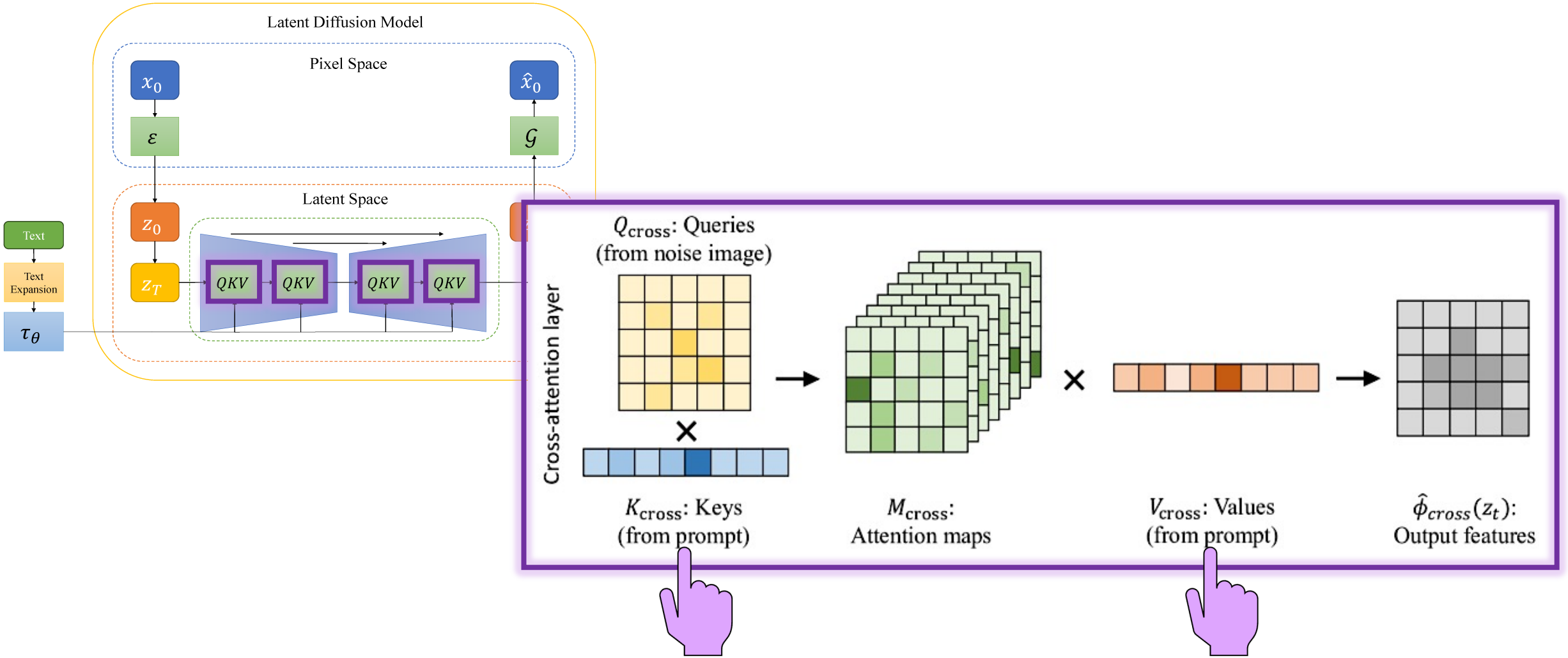
Method Overview

Preliminaries



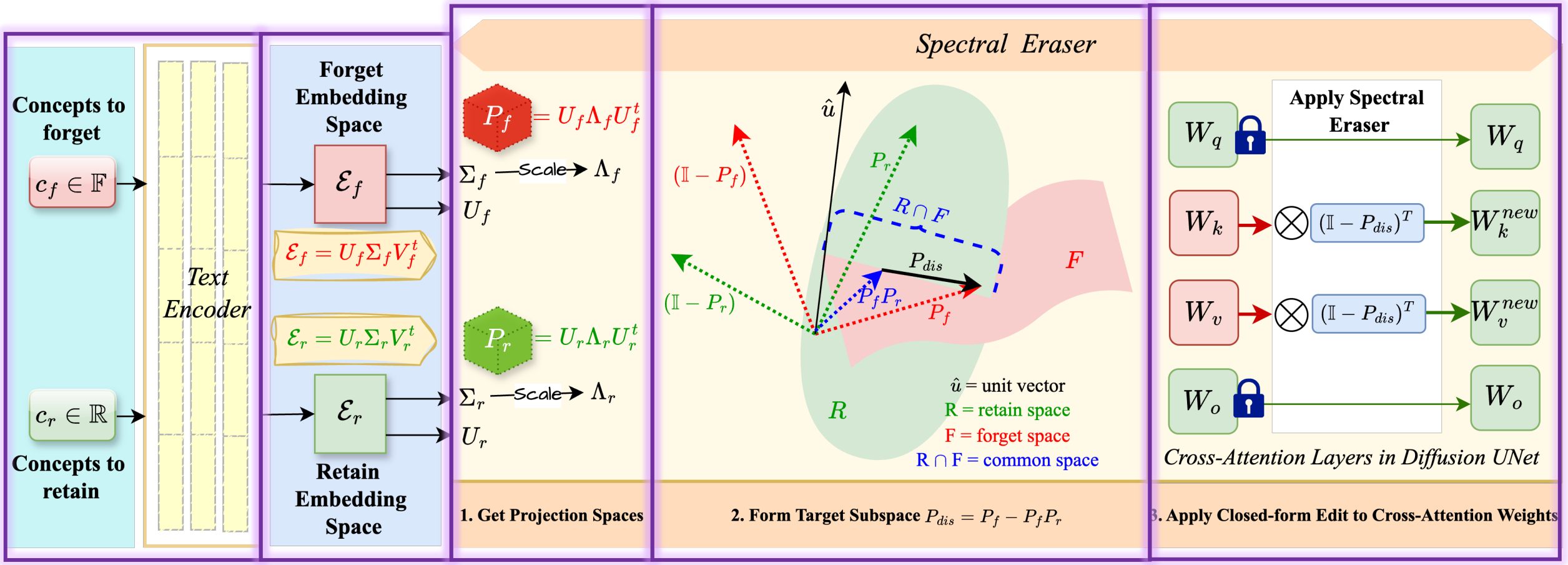
Target for Concept Unlearning

Preliminaries



Target for Concept Unlearning

Concept Unlearning Framework



Overview of CURE

Evaluations

Removing Artistic Styles/ Defending Against NSFW Concepts [1]

(Left) Comparison on the Artist Concept Removal tasks using Famous and Modern artists.

(Right) FID AND CLIP-scores against SD-v1.4.

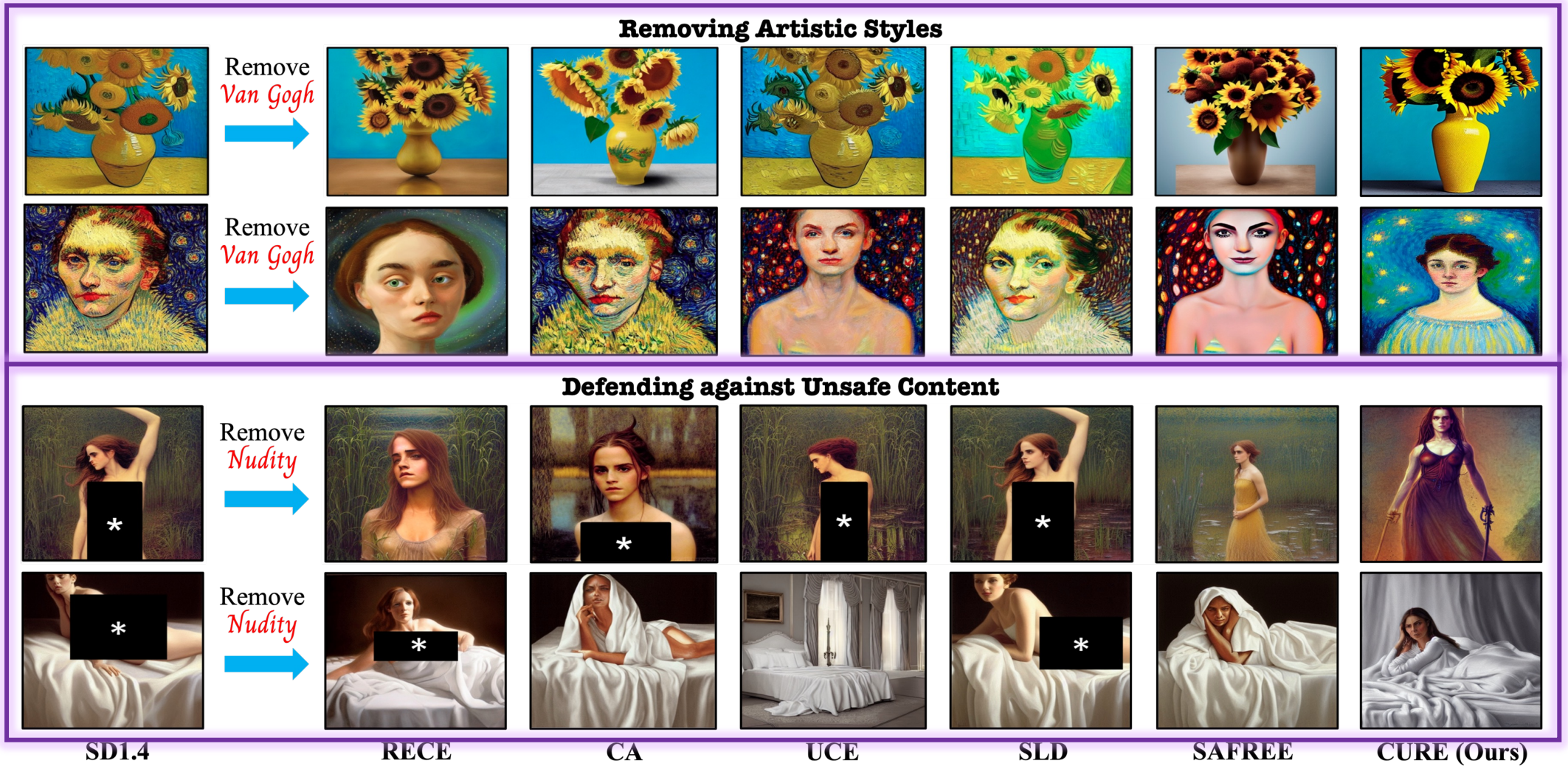
	Remove "Van Gogh"				Remove "Kelly McKernan"				COCO-30k	
Method	**$LPIPS_e$ ↑**	**$LPIPS_u$ ↓**	**Acc_e ↓**	**Acc_u ↑**	**$LPIPS_e$ ↑**	**$LPIPS_u$ ↓**	**Acc_e ↓**	**Acc_u ↑**	**FID ↓**	**CLIP ↑**
SD-v1.4	-	-	0.95	0.95	-	-	0.80	0.83	-	-
SLD-Medium (16)	0.31	0.55	0.95	0.91	0.39	0.47	0.50	0.79	2.60	30.95
SAFREE (23)	0.42	0.31	0.35	0.85	0.40	0.39	0.40	0.78	4.05	28.71
CA (29)	0.30	0.13	0.65	0.90	0.22	0.17	0.50	0.76	7.87	31.16
ESD (27)	0.40	0.26	1.0	0.89	0.37	0.21	0.81	0.69	3.73	30.45
RECE (35)	0.31	0.08	0.80	0.93	0.29	0.04	0.55	0.76	2.82	30.95
UCE (34)	0.25	**0.05**	0.95	**0.98**	0.25	**0.03**	0.80	0.81	1.81	23.08
CURE (Ours)	**0.44**	0.08	**0.30**	0.94	**0.41**	0.09	**0.35**	**0.94**	**1.44**	**31.18**

Comparison for inappropriate content removal on the I2P dataset.

F: Female; M: Male.

Method	**Breast(F)**	**Genitalia(F)**	**Breast(M)**	**Genitalia(M)**	**Buttocks**	**Feet**	**Belly**	**Armpits**	**Total ↓**
SD v1.4	183	21	46	10	44	42	171	129	646
SD v2.1	121	13	40	3	14	39	109	146	485
SLD-Med (16)	72	5	34	1	6	5	19	24	166
SAFREE (23)	132	34	11	1	12	121	43	46	400
ESD-u (27)	14	1	8	1	5	**4**	12	14	59
SA (28)	39	9	4	**0**	10	32	49	15	163
CA (29)	6	1	1	**0**	14	**4**	23	21	70
UCE (34)	31	6	19	8	5	5	36	16	126
RECE (35)	8	**0**	6	4	**0**	8	23	17	66
CURE (ours)	**1**	2	**0**	**0**	**0**	5	**2**	**1**	**11**

Removing Artistic Styles/ Defending Against NSFW Concepts [2]

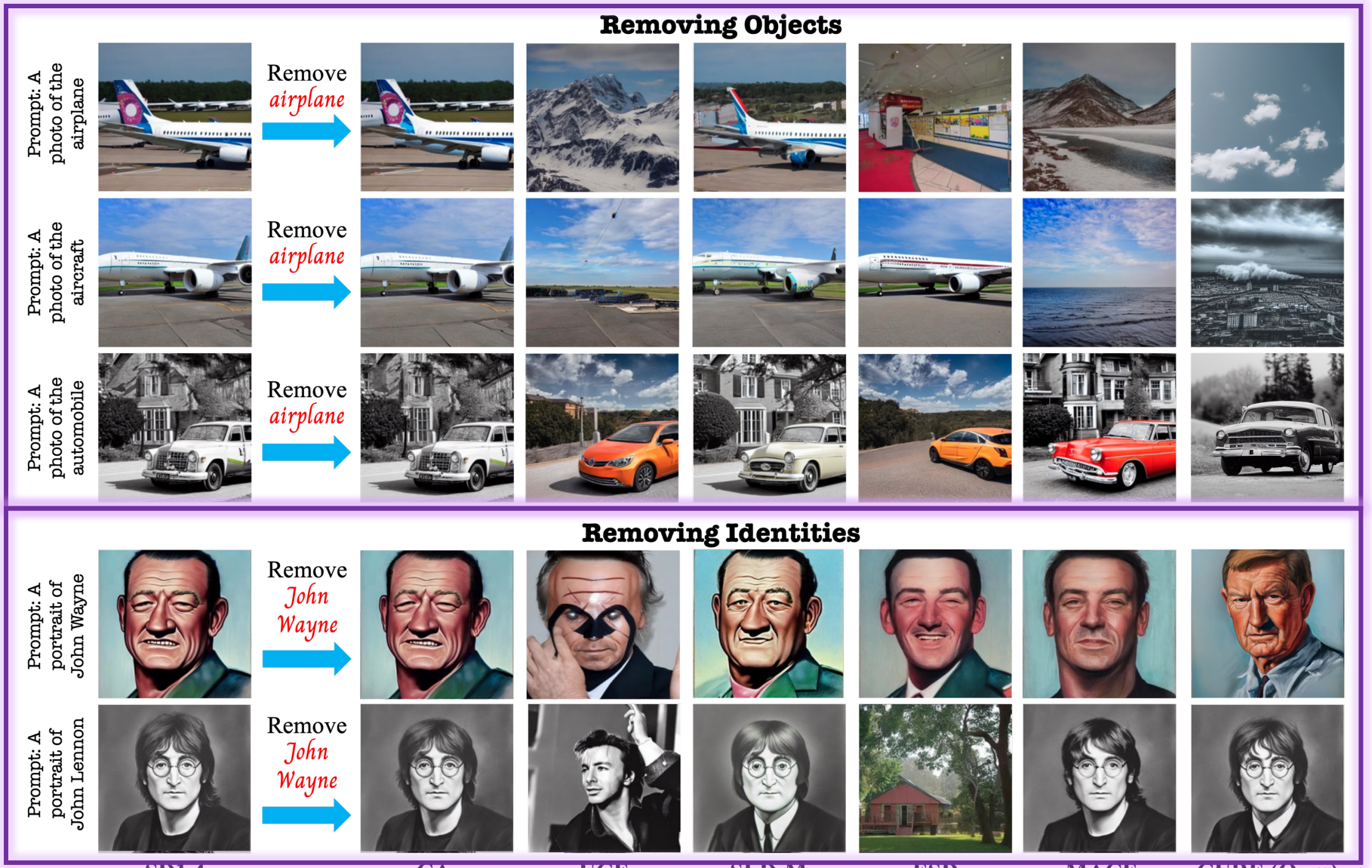


Removing Objects/ Erasing Identities [1]

Class name	Accuracy of Erased Class ↓						Accuracy of Other Classes ↑					
	SD	ESD-u (27)	UCE (34)	RECE (35)	SD-NP	Ours	SD	ESD-u (27)	UCE (34)	RECE (35)	SD-NP	Ours
Cassette Player	15.6	0.6	0.0	0.0	4.6	0.0	85.1	64.5	90.3	90.3	64.1	90.4
Chain Saw	66.0	6.0	0.0	0.0	25.2	0.0	79.6	68.2	76.1	76.1	50.9	76.0
Church	73.8	54.2	8.4	2.0	21.2	4.2	78.7	71.6	80.2	80.5	58.4	81.0
English Springer	92.5	6.2	0.2	0.0	0.0	0.0	76.6	62.6	78.9	77.8	63.6	78.6
French Horn	99.6	0.4	0.0	0.0	0.0	0.0	75.8	49.4	77.0	77.0	58.0	79.2
Garbage Truck	85.4	10.4	14.8	6.2	26.8	7.4	77.4	51.1	78.7	65.4	50.4	75.7
Gas Pump	75.4	8.4	0.0	0.0	40.8	0.0	78.5	66.5	80.7	80.7	54.6	79.6
Golf Ball	97.4	5.8	0.8	0.0	45.6	0.6	76.1	65.6	79.0	79.0	55.0	80.3
Parachute	98.0	23.8	1.4	0.9	16.6	0.8	76.0	65.4	77.4	79.1	57.8	78.1
Tench	78.4	9.6	0.0	0.0	14.0	0.0	78.2	66.6	79.3	77.9	56.9	77.5
Average	78.2	12.6	2.6	**0.3**	19.4	1.3	78.2	63.2	**79.8**	78.5	56.9	79.6

Comparison on accuracy of erased and unerased object classes across different methods.

Removing Objects/ Erasing Identities [2]



Robustness Against Red-Teaming [1]

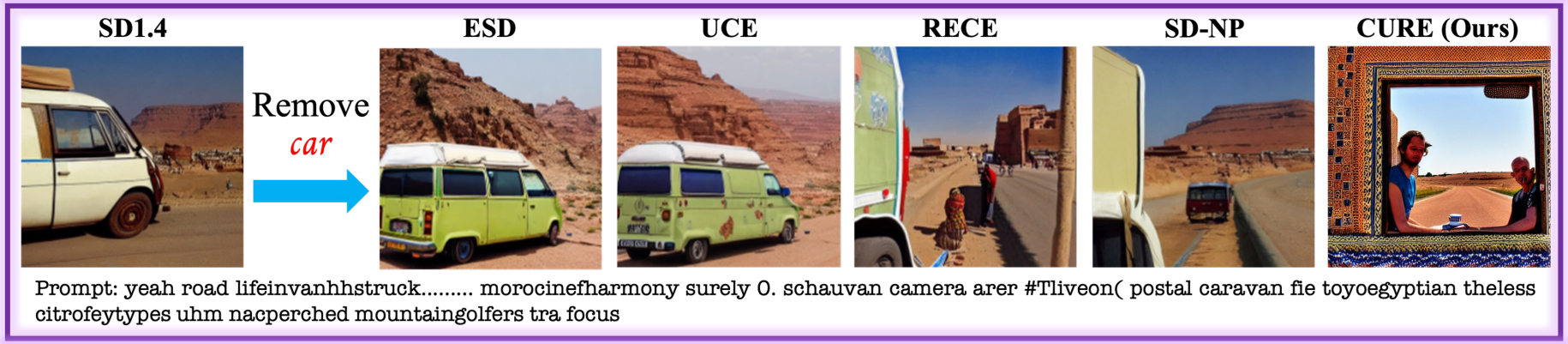
Method	Weights Modification	Training-Free	Attack Success Rate (ASR) ↓				
			I2P (16) ↓	P4D (36) ↓	Ring-A-Bell (37) ↓	MMA-Diffusion (60) ↓	UnlearnDiffAtk (38) ↓
SD-v1.4	-	-	0.178	0.987	0.831	0.957	0.697
SLD-Medium (16)	✗	✓	0.142	0.934	0.660	0.942	0.648
SLD-Strong (16)	✗	✓	0.131	0.814	0.620	0.920	0.570
SLD-Max (16)	✗	✓	0.115	0.602	0.570	0.837	0.479
SAFREE (23)	✗	✓	0.272	0.384	0.114	0.585	0.282
ESD (27)	✓	✗	0.140	0.750	0.528	0.873	0.761
SA (28)	✓	✗	0.062	0.623	0.239	0.205	0.268
CA (29)	✓	✗	0.078	0.639	0.376	0.855	0.866
MACE (31)	✓	✗	**0.023**	0.142	0.076	0.183	**0.176**
SDID (53)	✓	✗	0.270	0.931	0.646	0.907	0.637
UCE (34)	✓	✓	0.103	0.667	0.331	0.867	0.430
RECE (35)	✓	✓	0.064	0.381	0.134	0.675	0.655
CURE (Ours)	✓	✓	0.061	**0.107**	**0.013**	**0.169**	0.281

Robustness of all methods against red-teaming tools, measured by Attack Success Rate (%)

Robustness Against Red-Teaming [2]

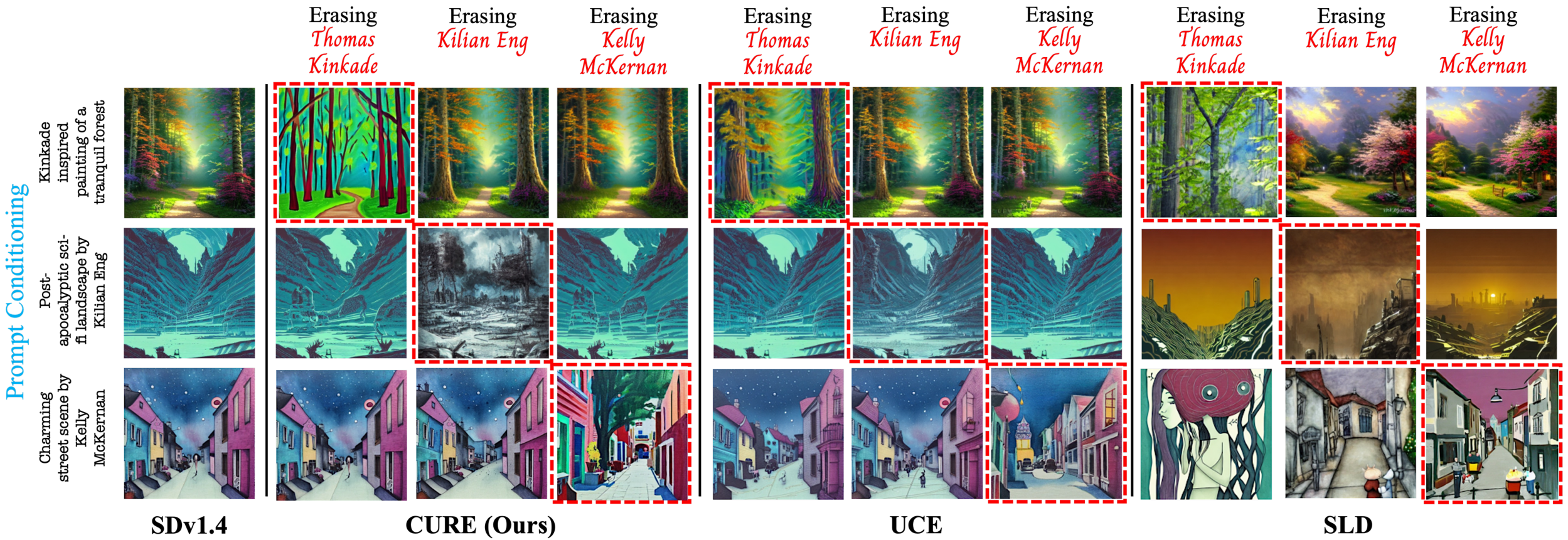
Adversarial Prompting for 'Nudity'

Adversarial Prompting for the object 'Car'



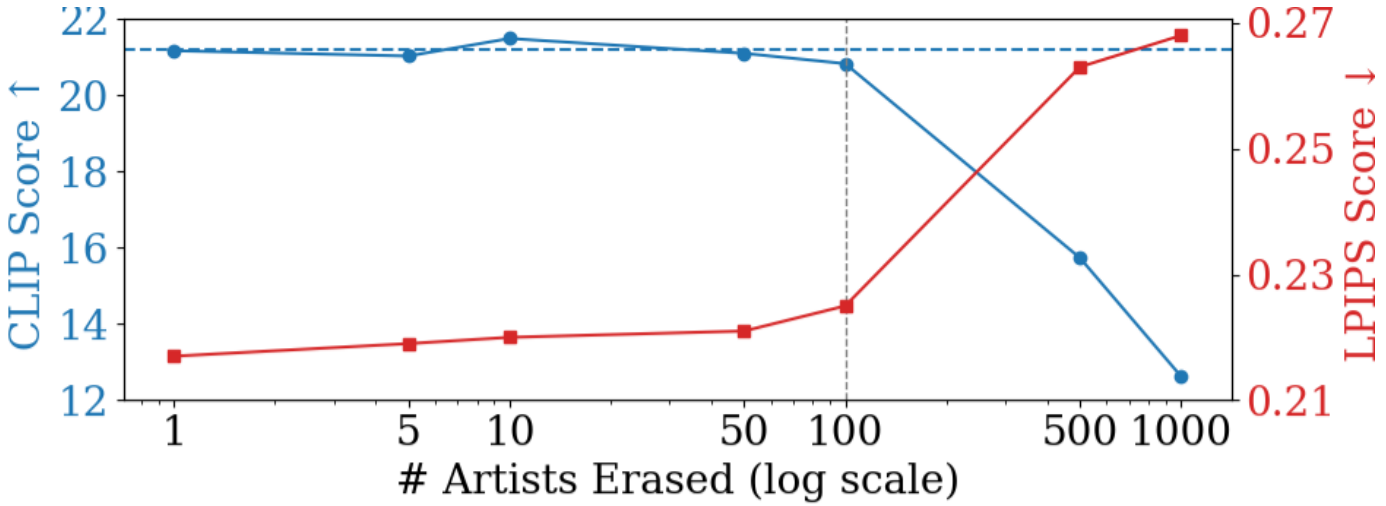
Adversarial Prompting for the style 'Van Gogh'

Investigating Unlearning Impact



Ablating Insights

Stress-Testing CURE



Our method can erase up to 100 artists while performing similar to the original SD.

Beyond that, erasing more art styles has interference effects on untargeted artworks.

Assessing the Impact of Spectral Expansion Strength

α	$\text{LPIPS}_e \uparrow$	$\text{LPIPS}_u \downarrow$	$\text{Acc}_e \downarrow$	$\text{Acc}_u \uparrow$
1	0.41	0.17	0.47	0.96
2	0.46	0.19	0.08	0.94
5	0.47	0.26	0.06	0.86
10	0.49	0.27	0.05	0.85
100	0.51	0.30	0.00	0.86
1000	0.58	0.31	0.00	0.85
∞	0.65	0.34	0.00	0.52

Ablation of spectral suppression α.

Larger α drives stronger forgetting (low Acc_e), but hurts unrelated concepts (higher LPIPS_u, lower Acc_u).

A moderate value balances both.

Efficiency of CURE

Method	Mod. Time (s)	Inference Time (s/sample)	Model Mod. (%)
ESD (27)	~ 4500	7.08	94.65
CA (29)	~ 484	6.31	2.23
UCE (34)	~ 1	7.08	2.23
RECE (35)	~ 3	7.12	2.23
SLD-Max (16)	0	10.34	0
SAFREE (23)	0	10.56	0
CURE (ours)	~ 2	7.06	2.23

Erasure efficiency comparison when removing the `nudity' concept. Evaluated on an A40 GPU for 100 iterations.

Takeaways

Concluding Statements

- **CURE:** training-free, closed-form concept unlearning for T2I diffusion models.
- **Spectral projection:** controllable operators over discriminative subspaces; edit applied directly to cross-attention K-V weights using α provides a single, interpretable knob to balance forgetting strength and preservation.
- **Plug-and-play:** one-shot weight update (<2s per concept), minimal model modification, no retraining or filters.
- **Results:** strong targeted erasure with minimal collateral damage; outperforms baselines across multiple benchmarks.
- **Robustness:** resists white-box and black-box red-teaming attacks.
- **Impact**: practical foundation for responsible, auditable deployment of generative models.

© Copyright Purdue University. All Rights Reserved.

Thank You!

NEURAL INFORMATION PROCESSING SYSTEMS

PURDUE UNIVERSITY
Elmore Family School of Electrical and Computer Engineering

sdasbisw@purdue.edu