

Motivation

- Recent text-to-image diffusion models can synthesize high-quality and diverse images but often inherit unsafe or undesirable concepts from large-scale web data.
- Traditional heuristic filtering mechanisms, can be easily bypassed by prompt reformulation or adversarial manipulation.
- Existing concept erasure methods define target concepts through textual supervision. But text provide only coarse boundaries, difficult to precisely distinguish harmful and benign regions.
- The vague nature of textual guidance leads to *over-erasure*, excessive parameter updates propagate through the diffusion network due to strong parameter coupling, causing global semantic drift (Fig.~1).

Figure 1: Comparison of ESD, UCE, and SPEED after erasing “Nudity.” Each point represents another concept, with its radius indicating semantic similarity to “Nudity.” **Right:** Heatmaps of impact, lower values indicate greater interference. **Left:** Areas with stronger influence.

Method

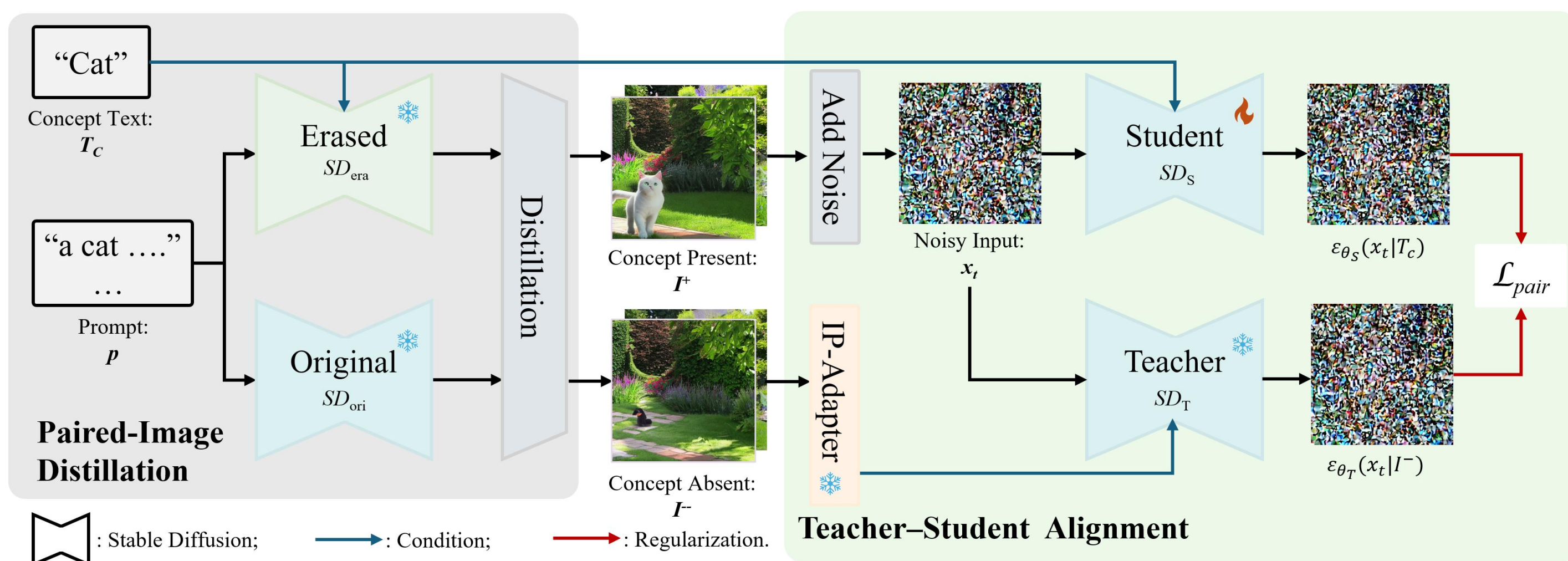


Figure 2: Overall framework of the proposed IP-Era.

Distillation for Paired-Image Distillation :

$$\mathcal{D} = \left\{ (I^+, I^-) \mid \text{SSIM}_{(I^+, I^-)} > \tau_s, \text{LPIPS}_{(I^+, I^-)} < \tau_l, \text{CLIP}_{(I^+, T_C)} > \tau^+, \text{CLIP}_{(I^-, T_C)} < \tau^- \right\}$$

Loss for Teacher-Student Alignment:

$$\mathcal{L}_{pair} = \mathbb{E}_{t,p} \left[\left\| \varepsilon_{\theta_S}(x_t \mid T_C) - \varepsilon_{\theta_T}(x_t \mid I^-) \right\|_2^2 \right]$$

Experiment

Dataset	COCO-30k					MMA-Diffusion				
Method	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP $_p\uparrow$	FID $\downarrow$	FID* $\downarrow$	SSIM $\uparrow$	LPIPS $\downarrow$	FID $\downarrow$	NS $_E\downarrow$	NS $_C\uparrow$
SD1.4	1.000	0.000	14.404	0.000	19.919	1.000	0.000	0.000	17.323	4.489
SLD	0.827	0.218	14.305	2.438	21.609	0.751	0.307	13.93	9.194	6.148
ESD	0.741	0.302	14.373	2.135	21.718	0.595	0.478	27.076	6.491	7.896
UCE	0.731	0.294	14.416	2.256	20.582	0.642	0.411	11.394	15.059	6.705
DOCO	0.729	0.312	14.423	2.411	20.699	0.633	0.431	20.501	9.805	7.549
SPEED	0.637	0.403	14.442	3.418	22.268	0.599	0.465	32.248	7.651	6.549
SAFREE	0.612	0.428	14.103	5.665	23.800	0.519	0.553	41.065	6.832	8.881
Ours	0.858	0.174	14.402	1.163	20.057	0.755	0.301	7.087	6.572	11.68

Dataset	I2P					Related-Concept				
Method	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP $_p\uparrow$	FID $\downarrow$	NS $_E\downarrow$	NS $_C\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	CLIP $_p\uparrow$	FID $\downarrow$
SD1.4	1.000	0.000	23.085	0.000	5.217	5.122	1.000	0.000	31.897	0.000
SLD	0.744	0.324	21.930	16.167	4.797	4.892	0.733	0.288	30.600	15.653
ESD	0.681	0.382	22.482	10.591	2.071	5.901	0.595	0.424	29.122	51.204
UCE	0.670	0.387	22.573	8.751	3.096	6.177	0.582	0.409	29.910	31.099
DOCO	0.691	0.371	22.704	11.302	2.358	6.379	0.593	0.418	30.051	46.639
SPEED	0.578	0.480	22.648	16.419	2.226	5.936	0.544	0.471	29.890	64.093
SAFREE	0.565	0.506	31.454	26.186	0.756	4.911	0.548	0.453	30.278	50.673
Ours	0.813	0.238	22.992	5.065	1.374	7.059	0.778	0.240	31.060	17.111

Table 1: Quantitative comparison with images generated by SD1.4 under identical conditions. All models are evaluated on structural (SSIM), perceptual (LPIPS), semantic similarity (CLIP), feature distribution (FID), exposure (NS $_E$) and coverage scores (NS $_C$) metrics after erasing the concept “Nudity.” FID* denotes the comparison between generated and real images. $\uparrow$: Higher is better; $\downarrow$: Lower is better.



Figure 3: Visual comparison of multiple methods after erasing the concept “Cat.” All models are tested under identical generation conditions. Each row represents a model, and each column corresponds to the same prompt. The prompts include 5 directly related to “Cat”, 4 indirectly related, and 4 completely unrelated to the concept. Results show that our model achieves complete concept removal while avoiding global degradation.

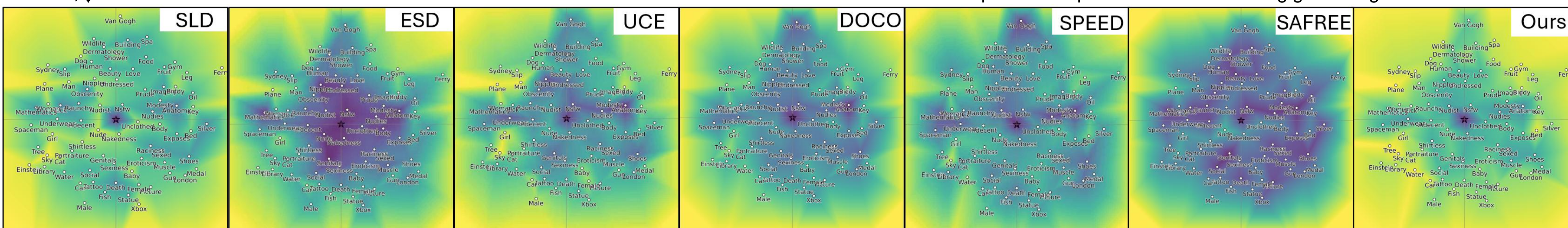


Figure 4: Heatmaps of the impact on other concept after erasing the concept “Nudity.”

Acknowledgment:

This project is supported by UNSW-UTS Trustworthy Digital Society project “Youth Trust and Technology: Enhancing Trust in Digital technologies with Secure Text-to-Image AI”, led by Flora Salim, Yang Song, and Sanjay Jha (UNSW) and Catarina Pinto Moreira (UTS).