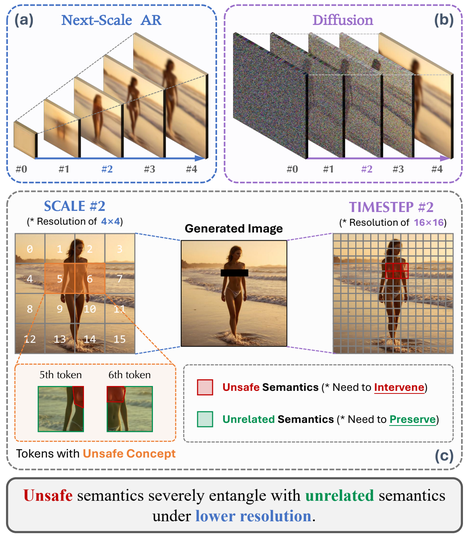
ScaleErasure: Inference-Time Minimal Intervention for Precise Concept Erasure in Next-Scale Autoregressive Image Generation

Cong Wang ; Haiyu Wu ; Zhiwei Jiang ; Zifeng Cheng ; Fei Shen ; Yafeng Yin ; Qing Gu

Introduction

Fig. 1(a,b) Multi-Stage Generation:
Both **next-scale AR** and **diffusion** follow a multi-stage generation philosophy.

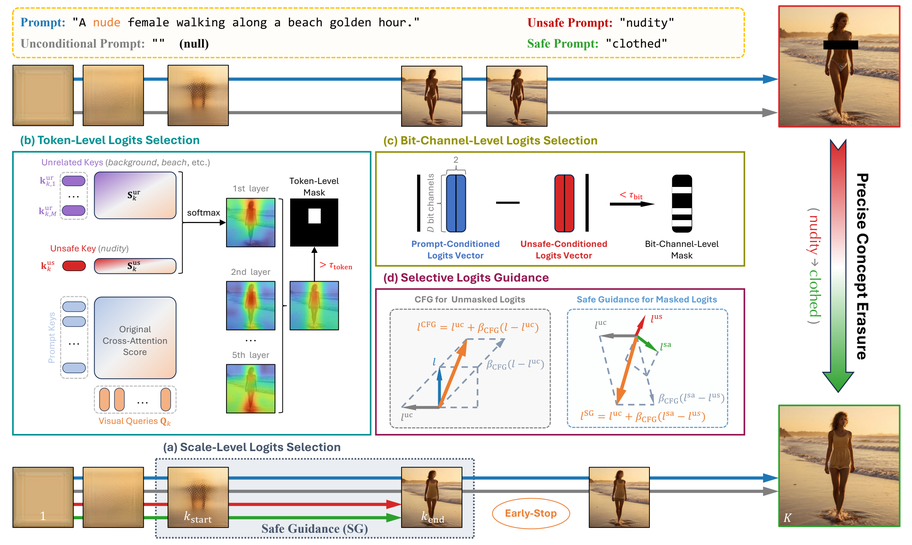
Fig. 1(c) Semantic Entanglement:
Unlike **diffusion**, **next-scale AR** operates at low resolutions in early scales, where **unsafe and unrelated semantics are compressed into the same token**.



How to ***precisely*** select and erase unsafe concepts ***under severe semantic entanglement***?

ScaleErasure: inference-time minimal intervention via selective logits guidance across ***scales***, ***tokens***, and ***bit channels***.

Method



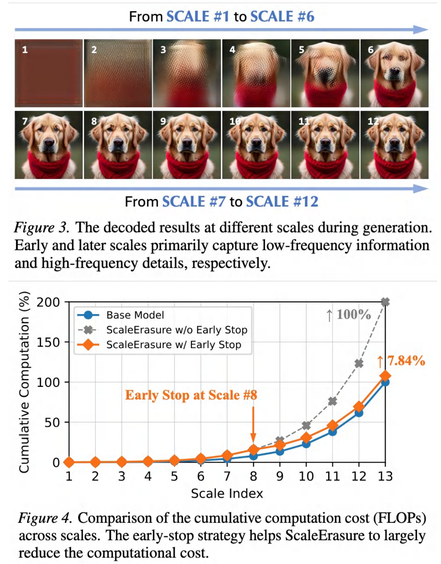


Figure 3. The decoded results at different scales during generation. Early and later scales primarily capture low-frequency information and high-frequency details, respectively.

Figure 4. Comparison of the cumulative computation cost (FLOPs) across scales. The early-stop strategy helps ScaleErasure to largely reduce the computational cost.

OVERVIEW: Guide only unsafe logits, while keeping unrelated logits unchanged.

- **Scale-Level Selection:** Select intermediate scales to balance structural preservation and semantic editability.
- **Token-Level Selection:** Localize unsafe-relevant spatial tokens via concept-aware cross-attention comparison.
- **Bit-Channel-Level Selection:** Identify unsafe-sensitive bit channels by contrasting prompt-conditioned and unsafe-conditioned logits.

Safe Guidance: Push selected logits away from unsafe concepts toward corresponding safe concepts.

Experiments

Table 2. Comparison of nudity erasure results on I2P, general generation capability on MS-COCO, and copyrighted erasure results on SMP. **Bold** values indicate the best performance, while underlined values indicate the second-best.

	I2P						MS-COCO		SMP						
	NudeNet				SD ↓ (10^3)	B-PSNR ↑	FID ↓	CLIP ↑	*Pikachu*			*SpongeBob*			Avg. H_a ↑
	Common ↓	Female ↓	Male ↓	Total ↓					CLIP-E ↓	CLIP-S ↑	H_a ↑	CLIP-E ↓	CLIP-S ↑	H_a ↑	
Base Model	728	285	29	1042	–	–	–	30.77	–	–	–	–	–	–	–
ESD-u	208	77	11	296	89.55	10.18	6.41	30.60	25.60	27.77	2.17	24.72	29.88	5.16	3.67
ESD-x	317	129	20	466	81.75	12.32	8.81	29.87	24.51	31.64	7.13	24.51	31.75	7.24	7.19
UCE	308	316	22	646	93.42	12.27	13.51	30.56	22.50	31.61	9.11	28.49	31.95	3.46	6.29
RECE	270	254	25	549	95.60	12.29	21.59	30.16	**22.21**	31.54	**9.33**	29.44	31.86	2.42	5.88
SLD-medium	534	183	27	744	**48.65**	**18.22**	4.19	30.54	32.88	31.68	-1.20	31.81	32.02	0.21	-0.50
SLD-strong	320	138	16	474	67.26	15.94	5.58	30.44	32.77	**31.87**	-0.90	27.80	**32.07**	4.27	1.69
ScaleErasure	**145**	**34**	**3**	**182**	58.24	17.27	**2.91**	**30.68**	22.48	31.08	8.60	**23.32**	31.96	**8.64**	**8.62**

Figure 5. Qualitative Comparison on I2P dataset.

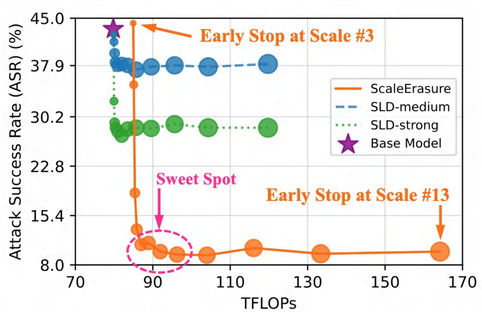


Figure 7. Comparison of efficiency (TFLOPs) and safety (ASR) across different early-stop scales, ranging from 3rd to 13th scale.

Table 3. Ablation study of ScaleErasure components. We report ASR on I2P to evaluate erasure effectiveness and FID on MS-COCO to assess generation fidelity.

	ASR ↓	FID ↓
ScaleErasure	9.99%	2.91
(a) *w/o* Token-Level Logits Selection	3.65%	73.11
(b) *w/o* Bit-Channel-Level Logits Selection	8.49%	3.08
(c) *w/o* Penalization in Safe Guidance	22.56%	2.94
(d) *w/o* Safe Logits Substitution in Safe Guidance	20.52%	2.34

Figure 8. Qualitative comparison for ablation study in Table 3.