GenErase: Generalizable and Semantically-Aware Concept Erasure in Diffusion Models

Supplementary Material

1. Paraphrases

In this section, we provide the list of paraphrases used for each target concept considered in our experiments. These target concepts correspond exclusively to those evaluated in the main erasure tasks and do not include the additional concepts from the generalization benchmark, GenBench-40. The complete set of paraphrases is presented in Tab. 1.

2. Object Eraser

In this section, we present extended results for the object erasure task. As shown in Tab. **??**, GenERASe consistently outperforms prior state-of-the-art methods across diverse object domains, achieving higher Erase Success Rate (ESR) and Preserve Success Rate (PSR) while incurring only a marginal increase in $\text{FID}_{\text{Non-Tar}}$. This slight rise in FID, confined to the low-value regime, reflects minor perceptual differences rather than semantic drift, whereas the substantial ESR gain indicates stronger generalization to varied object prompts. The stable PSR further confirms that non-target semantics remain intact, demonstrating that GenERASe achieves effective erasure without compromising fidelity or prior preservation.

3. Art Style Eraser

In this section, we evaluate artistic-style eraser and compare against prior baselines. Because the underlying text-to-image model does not reliably resolve paraphrases of artist names, we report results for the simple (non-synonym-expanded) eraser setting. In these comparisons, GenErase nearly matches the performance of the current state of the art (AdaVD) and clearly outperforms earlier guard-railing methods. We show the quantitative results in Tab. 3 and qualitative results in Fig. 1

Figure 1. **Qualitative Results for Artistic style eraser.** Col 1: Van Gogh Style erased Col 2-5: Andy Warhol, Caravaggio, Monet and Picasso styles preserved nearly perfectly.

4. Localization

Localization refers to the model's ability to selectively modify only the visual regions associated with the target concept while leaving unrelated content intact. This property is crucial for ensuring that concept erasure operates precisely, without distorting co-occurring or background elements in complex multi-entity scenes. We present qualitative examples illustrating that **GenErase** achieves spatially localized edits, only the target character or object is altered, while all non-target subjects remain visually consistent. These results highlight that our method effectively disentangles spatial semantics, ensuring fine-grained control even in crowded or context-rich images. Although minor visual variations may appear in non-target regions, their underlying semantic content remains unchanged. Qualitative results are shown in Fig. 2.

Figure 2. **Localization results.** *Target Concept:* Tom Cruise. In each example, the target identity (Tom Cruise) is replaced by a generic man, while non-target subjects and scene semantics remain unchanged.

5. Role of Gating Mechanism

We study the role of the gating mechanism in Tab. 4. While ESR remains nearly unchanged across settings, removing the gate leads to a notable degradation in PSR and a sharp rise in FID for non-target concepts (e.g., $4.30 \rightarrow 19.58$ for *Donald Trump*). This indicates that the gate stabilizes edits by preventing distributional drift in non-target regions and mitigating over-erasure. Overall, the gating mechanism acts as a crucial control for preserving semantic fidelity under strong erasure operations.

Table 1. **Target concepts and their corresponding paraphrases.** Each paraphrase represents an alternate textual expression used to test robustness of the erasure method under semantically equivalent but lexically varied prompts.

Target Concept	**Paraphrases / Alternate Descriptions**
Donald Trump	*President of the United States of America, Husband of Melania Trump*
Mark Zuckerberg	*Founder of Facebook, CEO of Facebook*
Dwayne Johnson	*WWE fighter known as The Rock, Actor who is also known as The Rock*
Brinjal	*Aubergine, Eggplant*
Spectacles	*Reading-glasses, Eye-glasses*
Dog	*Animal known as man's best friend, Pet known as man's best friend*
Melania Trump	*Wife of Donald Trump who is the first lady of United States America*
Robert Downy Jr.	*Tony Stark, Face of the actor who played the role of Iron Man*
Prince William	*Future King of England, Husband of Kate Middleton*
Elon Musk	*Founder of Tesla, SpaceX and PayPal, CEO of Tesla, SpaceX and PayPal*

Table 2. **Object Erasure:** Evaluation includes paraphrased targets. ESR ↑, PSR ↑, HM ↑, and non-target FID ↓. **Blue**: Best, **Orange**: Second-Best.

	Dog				**Brinjal**				**Spectacles**			
Method	ESR	PSR	HM	FID	ESR	PSR	HM	FID	ESR	PSR	HM	FID
NP [**?**]	**80.84**	23.41	36.31	77.64	**81.03**	23.17	36.04	61.74	**77.38**	23.34	35.86	66.63
SLD [**?**]	78.80	23.53	36.24	64.04	75.78	23.29	35.63	42.98	75.86	23.35	35.71	42.64
AdaVD [**?**]	80.25	**23.60**	**36.47**	**21.24**	78.64	**23.61**	**36.32**	**11.29**	76.91	**23.61**	**36.13**	**5.34**
SAFREE [**?**]	80.78	23.54	36.46	79.56	79.65	23.59	36.40	75.32	77.54	23.65	36.25	65.71
GenErase (Ours)	**81.14**	**23.59**	**36.55**	**27.66**	**81.25**	**23.66**	**36.65**	**12.30**	**77.91**	**23.64**	**36.27**	**6.32**

Table 3. **Single-Style Erasure for *Van Gogh*.** We report CLIP Similarity (CS, ↓) for the target concept (T), representing erasure efficacy. For non-target styles (NT1–NT4), we report FID (↓) reflecting prior preservation. Baseline results are taken directly from AdaVD [**?**]. T : Van Gogh, NT1 : Picasso, NT2 : Monet, NT3: Andy Warhol NT4: Caravaggio. **Blue**: Best **Orange**: Second-Best

Method	**Target (CS)**	**Non-Target (FID)**			
	T (Van Gogh)	**NT1**	**NT2**	**NT3**	**NT4**
NP [**?**]	**24.90**	141.56	124.52	127.85	136.32
SLD [**?**]	27.48	103.96	109.11	103.89	119.32
AdaVD [**?**]	**24.87**	**6.82**	**2.66**	**8.36**	**6.84**
SAFREE [**?**]	25.82	130.35	128.71	127.72	134.46
GenErase (ours)	24.96	**15.25**	**13.52**	**27.58**	**22.52**

Table 4. **Ablation on gating mechanism.** The gate stabilizes edits and prevents over-erasure by suppressing unintended drift in non-target regions.

Target	**ESR**	**PSR**	**HM**	**FID$_{\text{Non-Tar}}$**
With Gate				
Donald Trump	81.61	26.71	40.25	4.30
Mark Zuckerberg	85.20	26.61	40.55	13.96
Spectacles	77.91	23.64	36.27	6.32
Brinjal	81.25	23.66	36.65	12.30
Without Gate				
Donald Trump	81.61	26.59	40.11	19.58
Mark Zuckerberg	85.21	26.47	40.39	26.38
Spectacles	77.91	23.57	36.19	19.14
Brinjal	81.25	23.58	36.65	21.36

6. Anchor Selection Strategy

To ensure neutral and stable replacements during erasure, GenErase employs a *training-free anchor selection* procedure from a small candidate pool (e.g., "a man", "a person", "someone"). For each anchor $a^{(m)}$, we evaluate its usable energy in the editable subspace, orthogonal to both the target and all preserve directions, defined as

$$E_m = \frac{1}{J}\sum_{j=1}^{J} \left\|(I - P_j)(I - u_j u_j^\top)\tilde{v}_{a^{(m)},j}\right\|_2^2, \qquad (1)$$

where $\tilde{v}_{a^{(m)},j}$ is the CA-V vector of the m-th anchor at token j, P_j is the preserve projector, and u_j is the normalized erasure direction. The anchor with the highest mean energy E_m is selected as it retains the most semantic structure within the free subspace while remaining strictly orthogonal to both preserved and target components.

This deterministic, geometry-grounded selection ensures numerically stable replacements without introducing stochastic variability or domain bias. Empirically, we observe only marginal differences in ESR, PSR, and FID across anchors, with this energy-based selection yielding a modest yet consistent improvement in overall visual stability and non-target

fidelity. For direct comparison with prior works [**?** **?**], we follow the evaluation setup of AdaVD [**?**] on *Melania Trump*, reporting CLIP Similarity (CS, ↓) for targets and FID (↓) for non-targets. As shown in Tab. 5, GenErase matches SOTA performance on direct prompts and outperforms all methods under paraphrasing, highlighting its superior generalization.

Table 5. **Single-Celebrity Erasure for *Melania Trump*.** We report CLIP Similarity (CS, ↓) for the target concept, T (direct), P (paraphrase), and HM, aligned with ESR. For non-targets, we report FID (↓), where NT1–NT4 are Bruce Lee, Marilyn Monroe, Anne Hathaway, and Tom Cruise. Baselines from [**?**]; paraphrased CS measured in our setup.**Blue**: Best **Orange**: Second-Best

Method	Target (CS)			Non-Target (FID)			
	T	P	HM	NT1	NT2	NT3	NT4
NP [**?**]	23.73	24.79	24.25	115.35	103.83	106.04	106.00
SLD [**?**]	25.45	24.67	25.05	90.69	93.93	104.48	88.31
AdaVD [**?**]	**23.28**	25.17	24.19	**7.32**	**6.86**	**6.52**	**5.74**
SAFREE [**?**]	**24.15**	**23.74**	69.63	87.85	85.82	101.23	
GenErase (Ours)	23.39	**23.89**	**23.64**	**7.58**	**7.15**	**6.97**	**6.54**

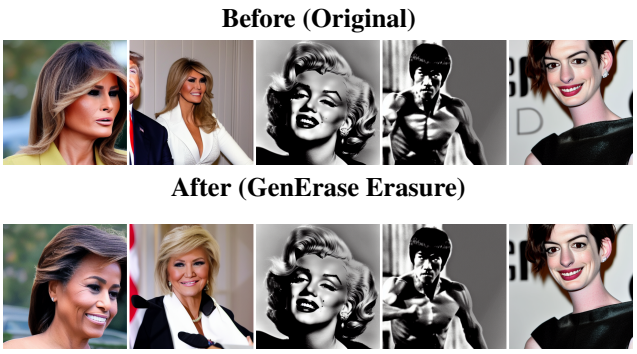

Figure 3. **Qualitative results for *Melania Trump* erasure.** Visual comparison *before* (top) and *after* (bottom) applying GenErase. Following the evaluation protocol of [**?**], results correspond to the quantitative metrics in Tab. 5, showing effective identity removal while preserving non-target content.

7. Computational Efficiency

We evaluate the runtime and memory footprint of GenErase against existing guard-railing methods. All operations, projection, gating, and replacement, occur directly in the attention value space, requiring no gradient updates or retraining. Empirically, GenErase incurs only a marginal increase in inference time while achieving higher erasure fidelity and preservation stability. As shown in Tab. 6, it also maintains a lower memory footprint than other projection-based approaches, owing to optimized, memory-efficient implementation. Overall, GenErase delivers strong concept erasure and generalization with negligible computational overhead.

Table 6. **Runtime and memory comparison** for erasing a target concept. Average pre-processing time, per-batch sampling time, total time for the batch, and peak VRAM usage are reported. Our computational burden is comparable to other projection-based methods, while the memory footprint remains smaller.

Method	Pre-Proc.	Sampling	Total	VRAM
NP [**?**]	0s	9.25s	9.25s	4.07 GB
SLD [**?**]	0s	13.96s	13.96s	16.52 GB
AdaVD [**?**]	1.13s	16.32s	17.45s	17.57 GB
SAFREE [**?**]	1.56s	9.85s	11.41s	19.45 GB
GenErase (Ours)	1.78s	16.89s	18.67s	14.32 GB

Table 7. Ring-A-Bell Ablation Study: Comparison of HGG and OER components.

Split	Full	-HGG	-OER	-Both
Clean	0.06	0.12	0.18	0.30
Attack	0.14	0.26	0.36	0.55

8. Adversial Robustness

We generate images using both clean and attack prompts and measure whether the erased concept can be recovered under attack using Attack Success Rate (ASR). To isolate contributions, we evaluate four variants (Tab. 7): Full, without HGG, without OER, and without both. We find that both OER and HGG are essential, and Full is most robust (lowest ASR).