One-dimensional Adapter to Rule Them All: Concepts, Diffusion Models and Erasing Applications

Supplementary Material

Contents

1 . Introduction 2

2 . Related Work 2

3 . Method 3
- 3.1. SPM as a 1-dim Lightweight Adapter 4
- 3.2. Latent Anchoring 4
- 3.3. Facilitated Transport 5

4 . Experiments 5
- 4.1. Single and Multiple Concept Removal 5
- 4.2. Training-Free Transfer Study 7
- 4.3. Versatile Erasing Applications 8

5 . Conclusion 8

6 . Acknowledgment 8

A. Analysis of SPM 1
- A.1. Dimension of SPM 1
- A.2 Component verification of SPM 1
- A.3. Sensitivity Analysis of η 2

B. Extended Experimental Results 2
- B.1. SPM for SD v2.1 and SDXL v1.0 2
- B.2. Cross-Application Multi-Concept Erasure . . 3
- B.3. Memorized Image Removal 3
- B.4. Generations of COCO-30k Caption 4
- B.5. SPM with Surrogate Concepts 4
- B.6. Numerical Results of Artistic Style Erasure . 6
- B.7. Samples of Nudity Removal 7
- B.8. Erasing Concepts with Shared Words 8
- B.9. Impact on Latent Representations 8

C. Detailed Experiment Settings 8
- C.1. Implementation Details 8
- C.2. Comparative methods 8

D. Additional Results 9
- D.1. Additional Samples of Single Concept Erasure 9
- D.2. Additional Samples of 20 Concepts Erasure . 9
- D.3. Additional Samples of Artistic Style Erasure 9
- D.4. Additional Samples of Training-Free Transfer 9
- D.5. Full Numerical Results of Object Erasure . . 9
- D.6. Failure Cases 10

E. Comparison with Concept-based Manipulation Methods 10

F. Societal Impact 11

Warning: The Appendix includes prompts with sexual suggestiveness and images with censored nude body parts.

A. Analysis of SPM

A.1. Dimension of SPM

To achieve precise erasure with minimum inference-time overhead, we have opted for a lightweight SPM with a one-dimensional intermediate layer. In addition to the effective and efficient results in the main text obtained with $d = 1$, here we explore the impact of dimensionality, i.e., the capacity, on the erasing performance. Tab. 3 shows the numerical results of SPMs with $d = 1, 2, 4, 8$ respectively, where the instance *Mickey* is eliminated as an example. As can be seen, with the increase in the intermediate dimension of the adapter, there is no discernible trend in the metrics for the target concept, other concepts, and general COCO captions. Fig. 8 also validates the robustness of SPM in the generated content. Thus, despite the low cost of increasing dimensionality, 1-dimensional SPM proves to be sufficient for our concept erasing task.

A.2. Component verification of SPM

Latent Anchoring (LA) and *Facilitated Transport* (FT) serve as a dual safeguard against the concept erosion phenomenon. In this section, we validate the effectiveness of each component during both fine-tuning and generation. Numerical results in Tab. 4 show that, without LA and FT, solely focusing on erasing can improve the metrics related to the targeted concept, but qualitative results in Fig. 9 demonstrate that

dim	Mickey CS↓	Mickey CER↑	Snoopy FID↓	Spongebob FID↓	Pikachu FID↓	Dog FID↓	Legislator FID↓	*General* FID_g↓
SD	71.94	2.50	-	-	-	-	-	13.24
1	63.04	13.50	31.28	36.02	25.62	7.40	10.67	13.25
2	61.96	15.25	32.08	37.01	26.60	8.43	11.94	13.26
4	62.70	14.88	31.21	36.09	26.06	7.53	10.69	13.23
8	62.01	16.62	32.04	36.58	26.27	7.96	10.99	13.25

Table 3. **Numerical analysis of the dimension of SPM**. In erasing *Mickey*, elevating the intermediate dimensionality of the SPM results in minimal fluctuations in performance concerning target erasure, concept preservation, and general generation capability. It sufficiently demonstrates that a one-dimensional setting is a judicious choice for both effectiveness and efficiency.

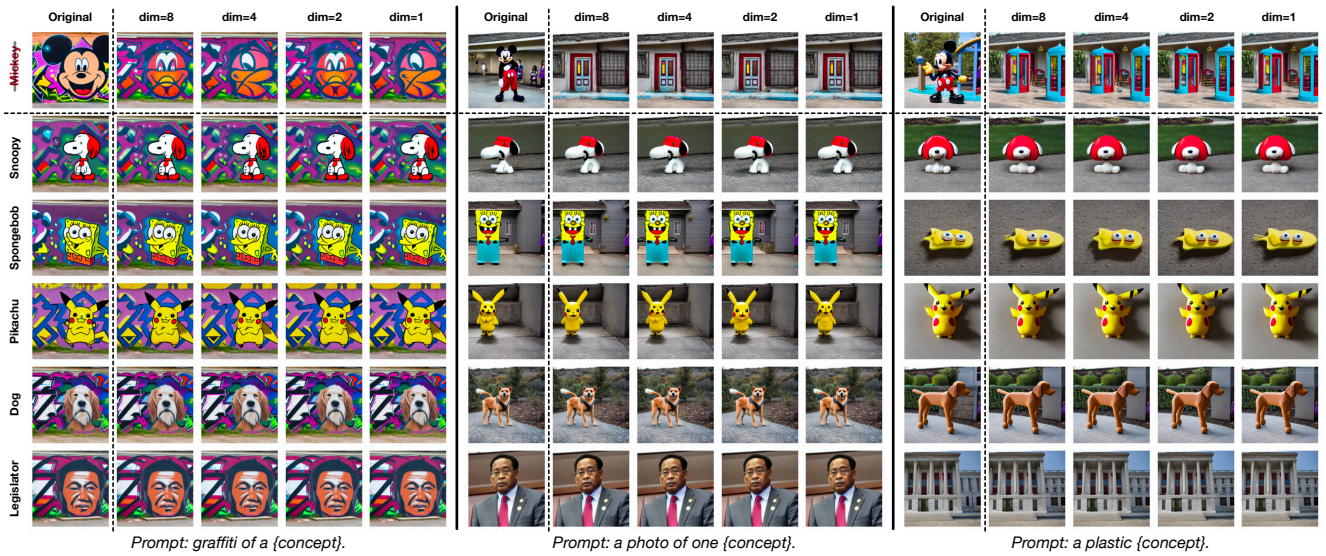

Figure 8. **Dimension Analysis of SPM**. The target concept *Mickey* is erased with $8, 4, 2$ and 1-dimensional SPM, and we showcase the results generated with three CLIP templates. It demonstrates that 1-dimensional SPM proves to be sufficient for both elimination and preservation.

LA	FT	Mickey CS↓	Mickey CER↑	Snoopy FID↓	Spongebob FID↓	Pikachu FID↓	Dog FID↓	Legislator FID↓	*General* FID_g↓
SD		71.94	2.50	-	-	-	-	-	13.24
		45.68	78.88	103.50	120.97	98.70	37.80	60.61	13.66
✓		53.67	35.12	50.33	57.35	42.69	16.52	27.29	13.12
	✓	54.06	41.12	42.25	44.54	35.61	17.69	28.34	13.28
✓	✓	63.04	13.50	31.28	36.02	25.62	7.40	10.67	13.25

Table 4. **Numerical component verification** with the *Mickey* erasure as an example. Despite the influence of Latent Anchoring (LA) and Facilitated Transport (FT) on the metrics of the target concept, as visually demonstrated in Fig. 9, the main entity does not exhibit the targeted semantics. Instead, it is attributed to changes in other parts, such as the background. With the prerequisite of sufficient target erasure, the metrics of other concepts and general COCO captions is greatly improved by LA and FT.

our method persistently pursuing a lower CS metric yields diminishing returns. More importantly, it comes at the cost of severe alteration and erosion of non-target concepts: The FID of *Snoopy* surges dramatically from 31.28 to 103.50, and the metric of *legislator*, which is semantically distant, also increases by 5.68 times. The FID increase is evident in the visual outcomes presented in Fig. 9. All concepts, regardless of their semantic proximity, show alterations in their generation. And close concepts such as the *Spongebob* and *Pikachu* are severely eroded. With LA for regularization, the FID of the concepts near the target is reduced by ∼50%, which demonstrates that the capability of Diffusion Model (DM) is efficiently retained. Generations of Fig. 9 also demonstrate that the semantic information of other concepts is well preserved, with only minimum alterations present. After deployment, the *Facilitated Transport* of SPMs further ensures the erasing of their corresponding targets, while minimizing

η	Snoopy CS↓	Snoopy CER↑	Mickey FID↓	Spongebob FID↓	Pikachu FID↓	Dog FID↓	Legislator FID↓	*General* FID_g ↓
SD	71.94	2.50	-	-	-	-	-	13.24
0.5	57.49	18.13	27.06	30.44	18.90	9.76	6.50	13.24
1	55.48	20.12	28.39	30.75	18.61	10.11	7.40	13.24
3	52.86	31.75	28.90	32.41	21.40	11.65	8.66	13.24
10	50.59	41.38	29.80	33.75	22.29	12.57	10.08	13.26

Table 5. **Sensitivity** of η on erasing *Snoopy* while preserving related concepts and other general concepts.

the impact on the generation of other prompts. As can be seen from Tab. 4 and Fig. 9, we can obtain generation results with minimal distortion on non-target concepts.

A.3. Sensitivity Analysis of η

Results in Tab. 5 present the sensitivity of the SPM to η in Eq. 4. As we have discussed in Sec. 3.2, increasing η leads to better removal on targeted concept; however, the alteration phenomenon could also manifest in the inspected non-targets. It is worth noting that even with significant adjustments of η, the FID_g metric indicates that SPMs preserve a strong generative consistency on general concepts, demonstrating its robustness.

B. Extended Experimental Results

B.1. SPM for SD v2.1 and SDXL v1.0

To validate the generality of our proposed SPM, we conduct erasure experiments on the more advanced SD v2.1 [37][5] and SDXL v1.0 [32][6], using instance removal as an example.

[5] https://huggingface.co/stabilityai/stable-diffusion-2-1
[6] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

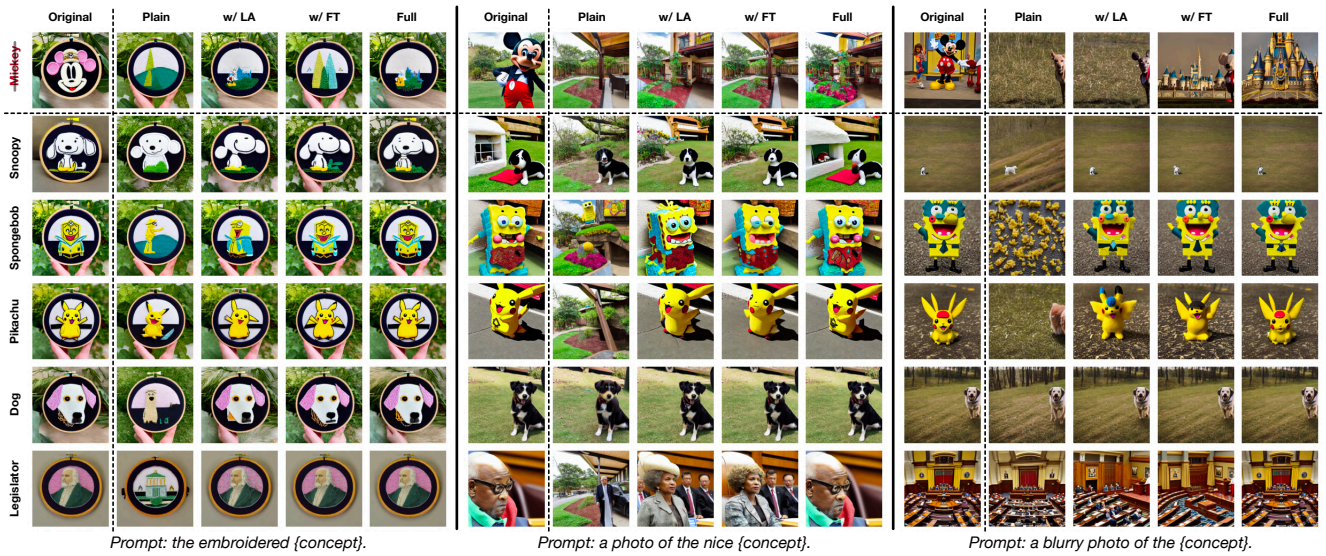

Figure 9. **Component verification of SPM** with the *Mickey* erasure as an example. Columns from left to right are generations obtained from: original SD v1.4, the erasing baseline, erasing with Latent Anchoring (LA), erasing with Facilitated Transport (FT) and erasing with both LA and FT as our full method. Qualitative results demonstrate that both proposed components effectively suppress the concept erosion without compromising erasure efficacy. Simultaneously, utilizing both of them helps minimize the generation alterations.

Fig. 10 and Fig. 12 show that, without intricate parameter search, the proposed SPM with one intrinsic dimension generalizes to different generative structures and achieves precise erasure. The erasing and preservation performance demonstrate that our conclusions drawn on SD v1.4 equally apply to newer models, notwithstanding variations in distribution for the targeted concept (e.g., the ubiquitous degeneration of *Snoopy* and *Spongebob* in SD v2.1). This allows for efficient adaptation to updates in open-source models, ensuring the safety of generated content.

Besides the alteration mitigation of concepts from other prompts, examples from SDXL v1.0 also show that the proposed SPM effectively preserves non-target descriptions within a rich prompt, such as *outdoor*, *sailing*, *cyberpunk* and *wheat*.

B.2. Cross-Application Multi-Concept Erasure

Besides the multi-instance SPM overlays, here we explore cross-application multi-concept erasure. In Fig. 13, we present samples for the combinations of applications involving artistic styles and instances: *Van Gogh* + *Cat* and *Comic* + *Snoopy*. We observe that the concept composition and negation of SPMs not only manifest within the same application but also in collaborative interactions across different applications. For example, in the original SD v1.4, the term *comic* refers to a multi-panel comic style sometimes with speech bubbles. Therefore, when we prompt "comic of Snoopy", it generates a multi-panel comic of Snoopy. Thus, when the *comic* element is erased, the output becomes a single panel of Snoopy without affecting the cartoon style of the instance itself. Furthermore, when the SPMs of both elements are plugged in, both the comic style and Snoopy disappear, leaving only "Sailing out of the sea" as the generation condition.

B.3. Memorized Image Removal

In preventing DMs from memorizing training images, thereby causing copyright infringement or privacy leakage [2, 9, 48], we follow ESD [8] and ConAbl [18] to erase classical masterpieces and investigate the impact on the artistic style and other paintings.

Compared to concrete objects with variations and abstract concepts with diversity, the erasure of a memorized image necessitates more precision. Take *The Great Wave off Kanagawa* by Hokusai and *The Starry Night* by Vincent van Gogh for example, as shown in Fig. 11, SPM can be precisely applied to erase a range within a memorized image, without perceptible changes for closely related artists or other paintings.

We then quantitatively analyze the erasure of specifically targeted images and the preservation of related artworks in comparison with the original SD v1.4 generated outputs. The similarity between an image pair is estimated via the SSCD [31] score, which is widely adopted for image copy detection [18, 31, 47], where a higher score indicates greater similarity and vice versa. As we can see from Tab. 6, the SSCD scores of all the targeted artworks are brought down to levels below 0.1, which demonstrates successful erasure. Meantime, the erasure scope is effectively confined to a single image, as evident from the robust SSCD scores maintained in the generated content of the same and other artists.

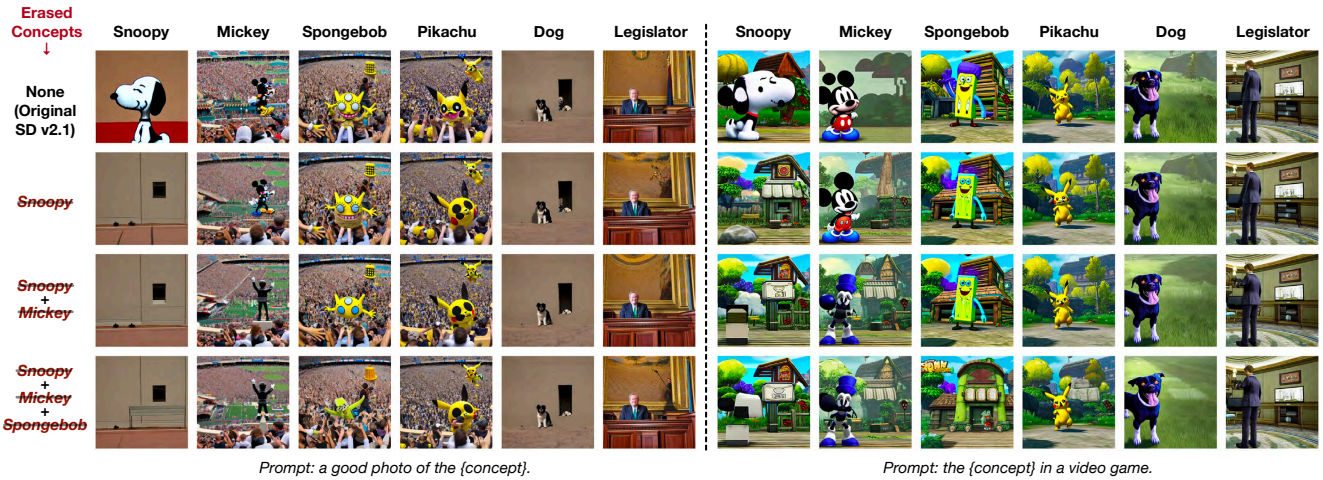

Figure 10. **Samples from SD v2.1 with one and multiple instance removed.** Our method can easily generalize to generative models with different architectures, and the erasing and preservation performance demonstrates that our conclusions remain applicable.

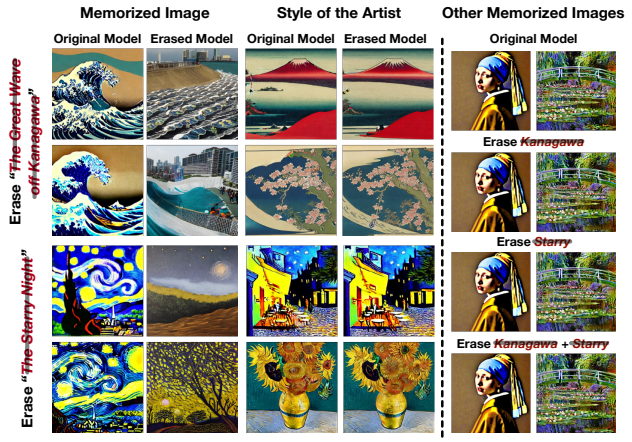

Figure 11. **Erasing specific images** memorized by the original DM (1-2 columns) with SPMs does not affect its ability to generate its artistic style (3-4 columns) or other images (5-6 columns).

Prompt	SSCD
Erasing ***The Great Wave off Kanagawa***	
The Great Wave off Kanagawa	0.04
Red Fuji by Hokusai	0.77
Plum Blossom and the Mooni by Hokusai	0.75
Girl with a Pearl Earring by Johannes Vermeer	0.99
Water Lilies by Claude Monet	0.91
Erasing ***The Starry Night***	
The Starry Night	0.09
Café Terrace at Night by Vincent van Gogh	0.87
Sunflowers by Vincent van Gogh	0.84
Girl with a Pearl Earring by Johannes Vermeer	0.98
Water Lilies by Claude Monet	0.94
Erasing ***The Great Wave off Kanagawa*** *and* ***The Starry Night***	
The Great Wave off Kanagawa	0.07
The Starry Night	0.08
Girl with a Pearl Earring by Johannes Vermeer	0.98
Water Lilies by Claude Monet	0.89

Table 6. **Quantitative results of specific image erasure** evaluated with the SSCD [31] model between the generated images of the original and erased DMs. A higher SSCD score indicates greater similarity. It shows that the a targeted image can be successfully eliminated without eroding artworks of the same or other artists.

B.4. Generations of COCO-30k Caption

The capacity of DMs with the proposed erasing SPM has been demonstrated numerically in Tab. 1, where the general FID remains stable throughout the continuous erasure of multiple concepts. Take the cases in Fig. 14 for example, the original generation results from SD v1.4 can align with the general objects, backgrounds, and actions indicated in the prompts. However, during the erasure of specific cartoon characters, previous methods exhibit the random disappearance of the original concepts, indicating a decline in the capability of concept perception or text-to-image alignment. In contrast, the non-invasive SPMs can preserve the original capacity of DMs to show stable performance for non-target concepts and general prompts.

B.5. SPM with Surrogate Concepts

Without loss of generality, we present our main results with the empty prompt as the surrogate, thus freeing it from manually selecting one for each targeted concept to align their distributions, which could be challenging and ambiguous, especially for non-instance concepts [10, 18]. Simultaneously,

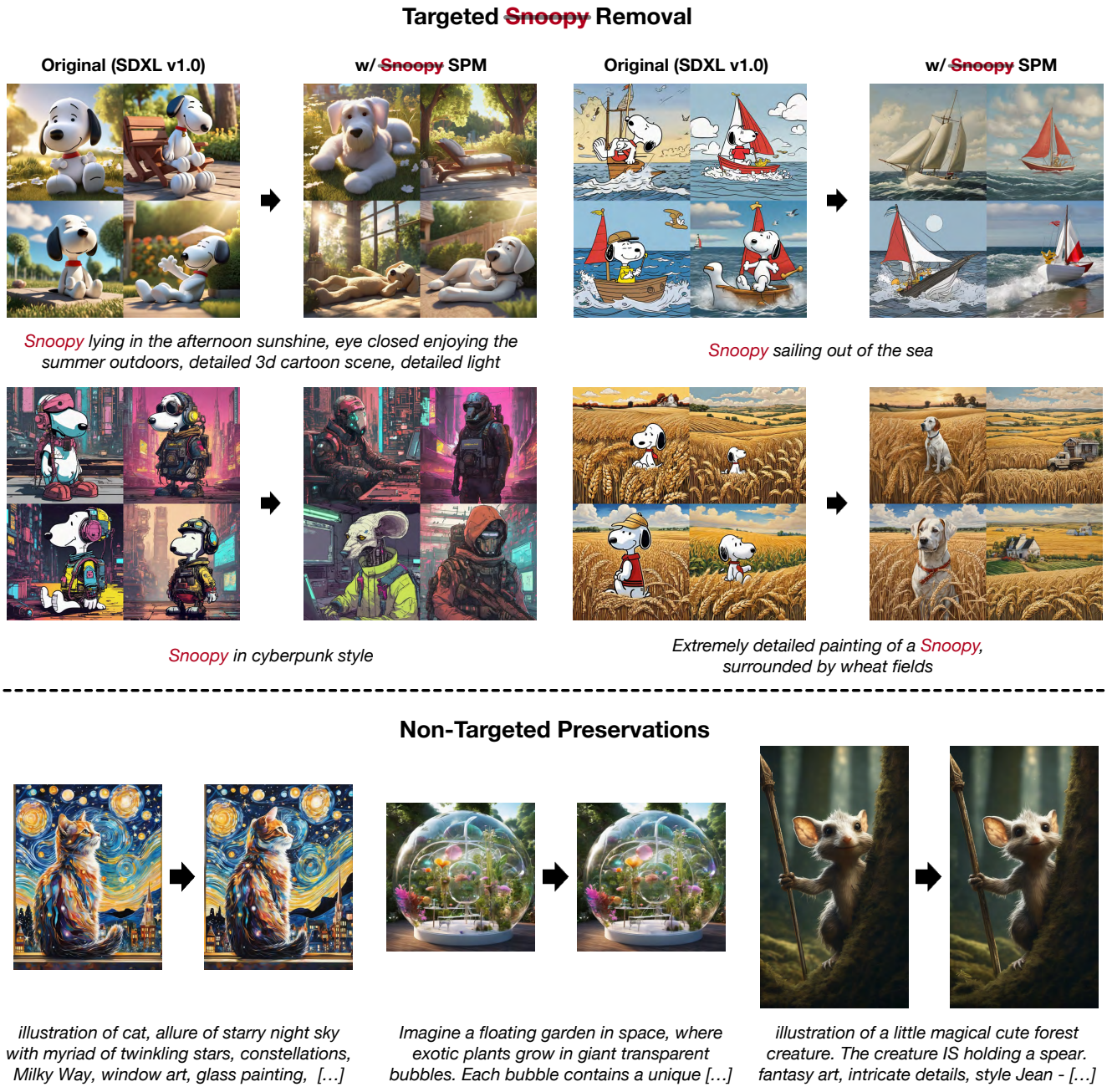

Figure 12. **Samples from SDXL v1.0 with the *Snoopy* SPM erasure.** In addition to the aforementioned effectiveness of erasure and preservation, as well as the generality across structures, we also observe that the proposed SPM effectively preserves non-target descriptions within a rich prompt, such as *outdoor*, *sailing*, *cyberpunk* and *wheat*.

our method also supports erasing a target towards a surrogate concept, which we informally term *concept reconsolidation*, to meet certain application requirements. Fig. 15 demonstrates the flexible application of SPM in reconsolidation through *Wonder Woman* $\rightarrow$ *Gal Gadot*, *Luke Skywalker* $\rightarrow$ *Darth Vader*, and *Joker* $\rightarrow$ *Heath Ledger* and $\rightarrow$ *Batman*. Both SD v1.4 generations and transfer results show that SPM not only precisely erases but also successfully rewrites the comprehension of specific concepts in generative models. It can thereby provide a result that aligns more closely with the user prompt while addressing potential issues such as copyright concerns.

Figure 13. **Samples from DMs with cross-application erasing.** In the combinations of *Van Gogh* + *Cat* and *Comic* + *Snoopy* erasure, we observe the concept composition and negation of SPMs across different applications.

B.6. Numerical Results of Artistic Style Erasure

In this section, we supplement the qualitative results in Fig. 7 and Fig. 21 with numerical analysis in Tab. 7. As can be seen, our method significantly surpasses the comparative approaches in the targeted style erasure, the preservation of other styles, and the general generation capability.

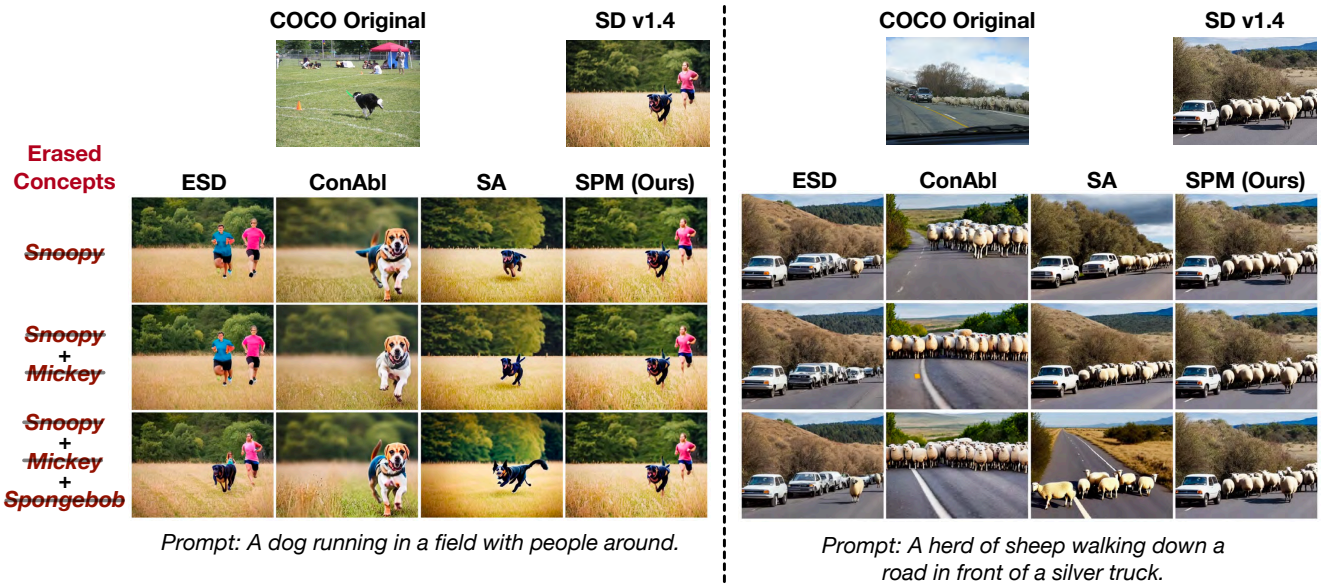

Figure 14. **Samples derived from prompts of COCO-30k Caption after one and multiple instances are erased from SD v1.4.** We observe that the content of the generated images aligns to the prompts with SPMs applied. No elements undergo erosion during the process of overlaying multiple concepts, and alterations are also well minimized.

	Van Gogh		Picasso		Rembrandt		Andy Warhol		Caravaggio		*General*
	CS	FID	CS	FID	CS	FID	CS	FID	CS	FID	FID_g
SD v1.4	74.01	-	70.16	-	71.57	-	71.56	-	74.05	-	13.24
*Erasing **Van Gogh***											
SLD	54.60	166.40	67.85	70.49	63.44	123.82	68.79	89.03	61.02	120.59	17.55
ESD	**50.64**	195.76	63.48	94.88	65.10	93.35	61.63	124.43	65.18	90.54	13.96
ConAbl	54.60	180.47	62.83	95.93	65.96	87.54	65.46	101.18	64.54	91.22	13.91
SA	60.84	138.78	67.50	104.11	64.56	161.85	69.96	119.27	65.70	141.19	30.53
Ours	51.80	198.65	68.96	**35.39**	70.53	**56.12**	70.45	**60.71**	72.06	**62.20**	**13.22**
*Erasing **Picasso***											
SLD	69.89	110.79	58.11	139.59	70.70	93.31	68.60	86.32	65.38	107.92	15.93
ESD	67.65	94.43	57.45	170.59	69.00	81.24	60.88	126.48	68.64	85.80	14.62
ConAbl	66.70	119.26	55.45	210.29	69.85	82.06	62.30	133.67	65.32	96.24	14.49
SA	67.02	124.06	64.58	126.64	65.04	171.33	68.95	128.30	64.89	156.11	29.50
Ours	73.55	**43.70**	**49.22**	269.58	71.22	**53.89**	70.52	**62.73**	71.98	**61.70**	**13.24**
*Erasing **Rembrandt***											
SLD	66.20	104.31	68.33	71.98	42.41	175.45	69.58	81.66	57.14	138.69	18.56
ESD	64.83	95.26	66.14	66.74	34.48	220.91	64.46	98.32	57.60	118.70	14.21
ConAbl	65.02	101.18	65.81	62.75	53.53	133.64	66.66	89.04	57.88	118.35	14.26
SA	65.55	128.12	67.15	99.20	57.54	167.43	70.91	128.51	62.76	152.15	30.14
Ours	73.13	**46.89**	69.26	**34.26**	**32.69**	275.29	70.66	**58.68**	70.31	**68.65**	**13.26**

Table 7. **Quantitative Evaluation of artistic style erasure.** The best results are highlighted in bold, the second-best is underlined, and the grey columns are indirect indicators for measuring erasure on targets or alteration on non-targets. We observe superior performance of our SPMs in target erasure (CS, Clip Score), non-target preservation (FID) and general generation capacity (FID_g).

B.7. Samples of Nudity Removal

In Fig. 16, we present examples where implicit prompts of the I2P dataset [43] elicit the original model the SD v1.4 to generate inappropriate content. In addition to showcasing the effectiveness of our SPM on the original model, we also directly transfer the SPM to the ChillOutMix derivative

for validation. Results show that the proposed method can effectively suppress the exposure of different body parts merely through the erasure of a single word *nudity*. The training-free transfer results also demonstrate its efficacy on models optimized towards the generation of such content.

B.8. Erasing Concepts with Shared Words

Fig. 17 showcases the generative results of concepts that share common words with the erasing target. We find that synonyms are effectively erased (~~*Mickey Mouse*~~ vs *Mickey*), while different concepts (~~*Mickey Mouse*~~ vs *Mouse*, ~~*Batman*~~ vs {*}*man*) with shared terms, despite close semantic and visual proximity, are largely preserved. This verifies that the latent distance metric we designed for LA and FT in concept preservation is a more accurate representation of similarity than token-level overlap.

B.9. Impact on Latent Representations

Here we further investigate the impact of erasing a specific target on its surroundings in the continuous latent space, depicting the representations that are more similar, i.e., closer to the target but may lack natural language interpretability.

As shown in Fig. 18, the granularity of erasure extends beyond the object level, encompassing high-level patterns emerging in the generations associated with the target. It guarantees the thorough elimination of target, but also initiates discussions on the erasure granularity for interconnected concepts (e.g. *Minnie* & *Mickey* in Sec. D.6), which may lack universally agreed standards.

C. Detailed Experiment Settings

C.1. Implementation Details

Following previous arts [8, 10, 18], we primarily conduct our comparative experiments on SD v1.4. We also validate the the generality of the proposed SPM on SD v2.0 in Sec. 4.3 of the main text, and on the latest SD v2.1 and SDXL v1.0 in Sec. B.1. In the experiments, SPMs are injected into the linear and convolution layers of the U-Net. The pre-trained parameters are fixed, and only the parameters of the SPM are adjusted, introducing a negligible parameter overhead of approximately 0.05% to the model. In initialization, $\boldsymbol{v}_{sig}$ is zero-initialized and $\boldsymbol{v}_{reg}$ employs Kaiming initialization with $a = \sqrt{5}$, ensuring continuity with the original model at the beginning of the training process. Unless otherwise specified, we employ a training schedule consisting of 3,000 iterations with a batch size of 1 for training and 4 samples for latent anchoring. The parameters of SPM are optimized on an NVIDIA A100 GPU using the AdamW8bit optimizer, with a learning rate of 1e-4 and a cosine restart schedule incorporating a 500 iteration warmup period and 3 restart cycles. Except for the concept reconsolidation experiment in Sec. B.5, without loss of generality, surrogate concepts in all experiments are set to the empty prompt. The loss balancing factor λ in Eq. 8 is chosen as 10^3, and the sampling factor α and erasing guidance η is set to 1.0 without delicate hyper-parameter search.

All numerical and visual results of SD models presented in this study are obtained with a fixed seed of 2024, which is fed into the random generator and passed through the Diffusers[7] pipeline. We sample with 30 inference steps under the guidance scale of 7.5. Except for the nudity removal experiment, "bad anatomy, watermark, extra digit, signature, worst quality, jpeg artifacts, normal quality, low quality, long neck, lowres, error, blurry, missing fingers, fewer digits, missing arms, text, cropped, Humpbacked, bad hands, username" is employed as the default negative prompt.

Details of experiments on artistic erasure. In contrast to erasing concrete concepts, we exclusively utilize CS and FID as metrics in artistic experiments because establishing a surrogate concept for calculating the CER of abstract artistic styles may be ambiguous. In the application of SPM, we recommend doubling the semi-permeability γ, yielding better erasure performance on abstract artistic styles without compromising the generation of other concepts.

Details of experiments on explicit content erasure. To fully achieve the potential of SPMs in mitigating implicit undesired content, we adopt an extended training schedule of 15K iterations, together with $\eta = 3.0$, $\lambda = 10^2$ and $\gamma = 2.0$.

C.2. Comparative methods

All experiments involving comparative methods are conducted using their respective official public codebases.

- **SLD (Safe Latent Diffusion) [43]**[8] . SLD is proposed for the mitigation of inappropriate content such as hate and sexual material. ESD later extends its application to the elimination of artistic styles [8]. Both the results reported by SA [10] and our preliminary experiments substantiate its suboptimal quality in generation outputs after instance removal. Thus we compare with SLD in the contexts of artistic style removal and nudity removal. The default hyper-parameter configuration of SLD-Medium is adopted to balance between the erasing and preservation. Note that we adopt the term *nudity* as the targeted concept in the nudity removal experiment, which demonstrates better performance in the I2P dataset [43] compared to the 26 keywords and phrases suggested by the authors.
- **ESD (Erased Stable Diffusion) [8]**[9] . Following the original implementation, we choose to finetune cross-attention parameters (ESD-x) for artistic styles, and finetune the unconditional weights of the U-Net module (ESD-u) for instances and nudity. All ESD models are trained for 1,000

[7] https://github.com/huggingface/diffusers
[8] https://github.com/ml-research/safe-latent-diffusion
[9] https://github.com/rohitgandikota/erasing

steps on a batch size of 1 with a 1e-5 learning rate using Adam optimizer.

- **ConAbl (Concept Ablation) [18]**[10] . Following the original implementation, instances and styles are removed by fine-tuning cross-attention parameters. We add the regularization loss for instance removal to ensure the quality of generation outputs. For both ConAbl and SA [10], which necessitate the specification of surrogate concepts, we choose *Snoopy* → *Dog*, *Mickey* → *Mouse*, *Spongebob* → *Sponge*, *All artistic styles* → *paintings* (adopted by ConAbl), and *nudity* → *clothed* (adopted by SA [10]). Each surrogate concept name is augmented by 200 descriptive prompts generated via the ChatGPT 4 [30] API, followed by the subsequent generation of 1,000 images.
- **SA (Selective Amnesia) [10]**[11] . Considering the time consumption, we reuse the pre-computed FIM released by the authors. Apart from the targeted prompts *sexual, nudity, naked, erotic* for the nudity removal experiment, all other configurations are consistent with the aforementioned ConAbl experiments. Each surrogate concept is grounded in the generation of 1,000 images. We primarily adhere to the configuration of celebrities for the erasure of instances.

D. Additional Results

D.1. Additional Samples of Single Concept Erasure

In addition to the samples of "graffiti of the concept" with *Snoopy*-SPM presented in Fig. 3 of the main text, we show the results with more CLIP prompt templates with *Snoopy*, *Mickey* and *Spongebob* erasure in Fig. 19. It can be observed that ESD sacrifices the generation of other concepts for thorough erasure. The offline memory replay of ConAbl and SA strikes a trade-off between the two: the former achieves better erasure but still erodes other concepts, while the latter, although capable of preserving the semantic content of other concepts, often fails in target erasure. Additionally, all comparative methods alter the alignment of the multi-modal space, which results in evident generation alterations.

D.2. Additional Samples of 20 Concepts Erasure

In addition to the generations presented in Fig. 1 of the main text, we give more randomly chosen examples and their generations as 20 Disney characters are progressively erased: *Mickey, Minnie, Goofy, Donald Duck, Pluto, Cinderella, Snow White, Belle, Winnie the Pooh, Elsa, Olaf, Simba, Mufasa, Scar, Pocahontas, Mulan, Peter Pan, Aladdin, Woody* and *Stitch*.

The comparison in Fig. 20 shows that simply suppressing the generation of the targeted concept as ESD does may lead to degenerate solutions, where the DM tends to generate images classified as the surrogate concept, i.e., empty background in this case. Thus, we observe not only the erosion of other concepts, but also their convergence to similar 'empty' images, indicating a pronounced impact on the capacity of the model. In contrast, with the deployment of our LA and FT mechanisms, our method successfully mitigates the concept erosion phenomenon, and the object-centric images generated with 20 SPMs are quite robust.

D.3. Additional Samples of Artistic Style Erasure

In addition to Fig. 7 of the main text, we present results obtained from the erasure of *Van Gogh*, *Picasso* and *Rembrandt* in Fig. 21. Consistent with the conclusion in Sec. 4.3, we find that previous methods tend to trade off between erasing and preservation, whereas our proposed SPMs can successfully erase the targeted style while maximally preserving the styles of other artists.

D.4. Additional Samples of Training-Free Transfer

In addition to Fig. 6 of the main text, we further investigate whether complex variations of the targeted concept generated by SD-derived models can be effectively identified and erased using the SPMs fine-tuned solely based on the official SD and a single class name. As Fig. 22 shows, with our prompt-dependent FT mechanism, the erasure signal applied on the feed-forward flow proves effective in eliminating cat patterns: causing them to either vanish from the scene or transform into human-like forms (given that popular community models often optimize towards this goal). We observe that when anthropomorphic cats transform into humans, cat ear elements are frequently preserved. This phenomenon might be attributed to a stronger association between cat ears and humans rather than cats, as both the official SD model and community models conditioned on the prompt of "cat ears" generate human figures with cat ears.

D.5. Full Numerical Results of Object Erasure

Tab. 8 presents the comprehensive numerical outcomes of the general object erasure experiments. In addition to the CS and CER metrics displayed for the target concept, and the FID for the non-target in Tab. 1 of the main text, the remaining metrics are depicted in gray. Here we explain the reason we exclude these metrics as indicators for measuring the concept erasing task.

A higher FID for the targeted concept suggests a more pronounced generative difference for the targeted concepts. However, it cannot conclusively demonstrate the accurate removal of the content related to the target. Conversely, CS and CER assess the correlation between the generated image and the target concept, providing reliable evidence of the efficacy of the erasure.

In contrast, CS and CER solely measure the relevance of the content to the concept, potentially overlooking generation

[10] https://github.com/nupurkmr9/concept-ablation
[11] https://github.com/clear-nus/selective-amnesia

	Snoopy			Mickey			Spongebob			Pikachu			Dog			Legislator			*General*
	CS	CER	FID	CS	CER	FID	CS	CER	FID	CS	CER	FID	CS	CER	FID	CS	CER	FID	FID_g
SD v1.4	74.43	0.62	-	71.94	2.50	-	72.99	0.62	-	72.60	0.88	-	63.73	0.88	-	57.64	8.88	-	13.24
*Erasing **Snoopy***																			
ESD	**44.50**	**77.62**	163.93	54.01	45.13	129.07	59.81	18.12	113.90	64.92	12.62	72.18	62.74	4.38	45.94	56.44	11.25	55.18	13.68
ConAbl	59.81	5.50	199.44	64.51	20.00	110.85	67.96	2.25	79.49	69.92	3.75	71.22	64.55	0.25	96.36	57.50	7.75	55.74	15.42
SA	64.59	0.25	122.15	72.54	2.88	53.64	73.35	0.75	57.65	73.27	0.50	42.95	64.70	0.25	75.72	58.06	7.38	47.42	16.84
Ours	55.48	20.12	108.60	71.52	2.88	**28.39**	72.75	0.88	**30.75**	72.45	1.00	**18.61**	63.73	1.00	**10.11**	57.67	9.38	**7.40**	**13.24**
*Erasing **Snoopy** and **Mickey***																			
ESD	**45.49**	**67.00**	169.72	**44.23**	**83.12**	191.61	54.12	36.38	145.71	58.20	28.25	114.25	62.14	6.62	51.05	55.86	13.25	64.74	13.69
ConAbl	60.05	4.00	210.29	56.14	14.00	186.71	62.99	5.75	112.15	68.77	5.75	105.43	64.22	0.00	79.40	57.84	7.38	56.17	15.28
SA	63.33	10.75	167.87	60.93	51.12	180.91	66.02	14.38	148.33	74.55	3.00	129.52	67.55	0.38	137.91	58.79	35.38	151.94	17.67
Ours	55.11	20.62	110.93	52.04	39.50	142.36	72.27	0.75	**36.52**	72.14	1.00	**26.69**	63.69	0.62	**13.45**	57.62	8.25	**16.03**	**13.26**
*Erasing **Snoopy**, **Mickey** and **Spongebob***																			
ESD	**46.94**	**60.38**	160.21	**44.79**	**80.25**	186.85	**43.76**	**85.88**	211.59	53.53	43.62	137.23	62.23	4.50	50.77	54.96	18.00	73.96	13.46
ConAbl	60.88	1.12	191.86	55.10	23.12	194.34	58.46	15.38	224.36	69.36	3.88	102.79	64.43	0.00	67.43	57.16	8.12	55.72	15.50
SA	64.53	15.25	187.74	61.15	61.88	183.66	60.59	49.88	181.60	71.77	5.38	167.79	69.10	2.88	183.26	57.38	57.50	185.29	18.32
Ours	53.72	25.75	117.73	50.50	44.50	149.53	51.30	41.87	163.06	71.48	1.25	**33.19**	63.64	0.75	**14.69**	57.63	8.75	**20.66**	**13.26**

Table 8. **Extended quantitative Evaluation of instance erasure.** The best results are highlighted in bold, the second-best is underlined, and the grey columns are indirect indicators for measuring erasure on targets or alteration on non-targets.

alterations until they amount to substantial concept erosion. Inversely, a marked increase in FID indicates a significant alteration after the erasure process.

Nevertheless, we can derive valuable insights from these numerical results as well. Methods that exhibit a significant FID increase, while retaining similar CS and CER levels as the original model, such as ConAbl and SA, are subject to generation alterations. Regarding the target concept, despite a smaller increase in FID, the qualitative results depicted in Fig. 19 demonstrate that our method effectively preserves the non-targeted concept in the prompt, whereas other erasure techniques may erode these contents.

D.6. Failure Cases

SPM effectively removes the targeted concept and maintains consistency in non-target contexts. However, even when the original model fails to align with the prompt to generate the target, it may still function and alter the generation results. As illustrated in Fig. 23 (a), the original model does not generate the targeted instance when the prompt including the target "Mickey" is input. After the SPM is plugged in, the erased output also demonstrates a certain level of alteration. It occurs because the input prompt triggers the FT mechanism and activates SPM, thereby disrupting the generation process.

Another failure scenario, shown in Fig. 23 (b), examines the generation of a concept (*Minnie*) closely related to the target (*Mickey*). The outputs of non-target concept, given their semantic similarity to the target, exhibit generation alteration towards erosion. However, whether the erasure should prevent the generation of the target as a whole or fine-grain to iconic attributes of the target (such as the round black ears of Mickey), may still be a subject for more discussion.

In the application of nudity removal, variations in body part exposure and the implicit nature of prompts from the I2P dataset add complexity to the task. While Fig. 5 and Fig. 16 illustrate the superiority of SPM over dataset cleansing and previous approaches, it has not yet met our expectations for a safe generative service. As depicted in Fig. 23 (c), even though effective erasure has been achieved on the I2P dataset, when transferred to community DMs, the SPM fails to clothe the characters with $\gamma = 1.0$ for the same prompt unless we manually amplify the strength to 2.0.

These examples highlight that a safe system necessitates a clearer definition of erasure, more precise erasing methods, and the combination of multi-stage mitigations.

E. Comparison with Concept-based Manipulation Methods

In addition to **concept erasing**, for which SPM is designed, there is a diverse body of research on concept-based manipulations for DMs. Here we elaborate on the distinctions of SPM from these settings, and discuss its potential extensions within them.

Concept personalization methods are proposed mainly to introduce new concepts to DMs, such as specialized objects or styles that have not been learnt during the pretraining scheme. For example, Dreambooth [39] and Textual Inversion [7] take a few personalized images of a new concept to finetune the model parameters or representations. If given prepared data, SPM is also capable of fulfilling this function. Nonetheless, these approaches primarily focus on the establishment of a new concept and exploring its variations across various contexts, as opposed to modifying or erasing

an existing concept while ensuring the preservation of other relevant concepts in a self-supervised manner.

Image editing, along with **inpainting** methods, alter a given image based on the textual conditions. These works focus on aligning the input image with the specified prompt, while our erasing task emphasizes the detection of potential risks from any user prompts and the flexibility of generating safe content. The editing task aims to change the target content while preserving the rest within the input image, which partially parallels our motivation within the latent space. However, the absence of an original image for generation guidance increases the difficulty of preservation in our task. To tackle this challenge, our framework opts for editing via a 1-dim adapter instead of parameter fine-tuning of DM, and designs LA and FT mechanism to mitigate alternation.

Concept editing application bears a closer resemblance to our task, where the interpretation of a specific concept is altered for safety, diversity, and faireness. In fact, as illustrated in Sec. B.5, SPM can also **rewrite** one concept with another using surrogate concepts, promising further extensibility of SPM.

F. Societal Impact

The proposed SPM provides a non-invasive, precise, flexible and transferable solution for erasing undesirable content from DMs while preserving the overall quality and coherence of generated content. Aligning with emerging regulatory frameworks and ethical standards, it can effectively address concerns related to the use of artificial intelligence in content generation, including copyright infringement, privacy breaching, misleading or mature content dissemination, etc. However, the choice of targeted concept is neutral, which also exposes it to potential misuse, such as concealing or reconsolidate specific events [10]. We believe that strengthening oversight and review mechanisms should be considered to ensure transparency and explainability in the decision-making process of the erasing models. Users and stakeholders should have a clear understanding of how the model is governed to build trust in the evolving landscape of content generation.

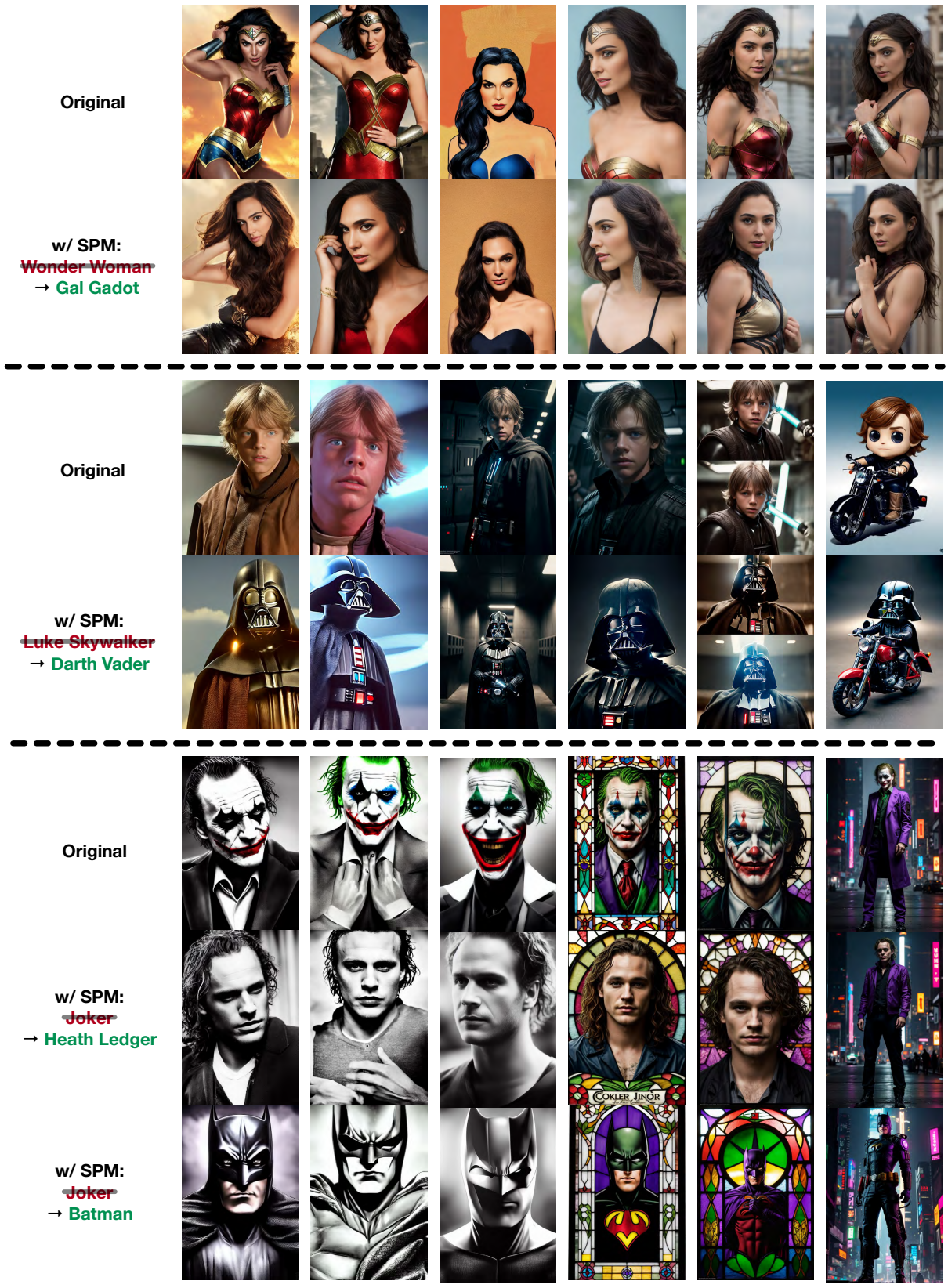

Figure 15. **Samples from SPM-applied DMs for concept reconsolidation.** By specifying a non-trivial surrogate for the targeted concept, SPM can adjust the distribution of undesired concepts to match the surrogate distribution in a context-aware manner.

Figure 16. **Samples conditioned on the I2P prompts with the concept *nudity* erased.** In each row, from left to right, we present the prompt from the I2P dataset, the generation outputs of SD v1.4, SD v1.4+SPM, ChillOutMix, and ChillOutMix+SPM.

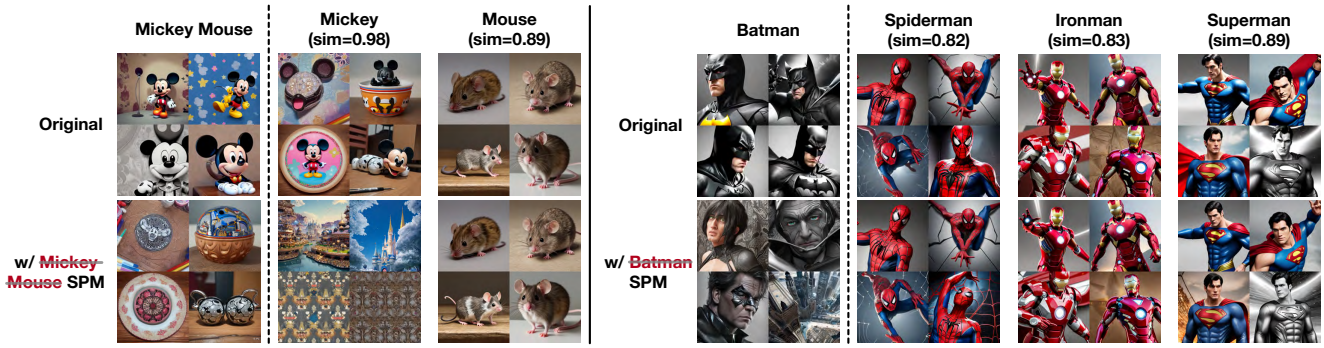

Figure 17. Impact of SPMs on concepts that share words with the targets (**sim** as cosine similarity).

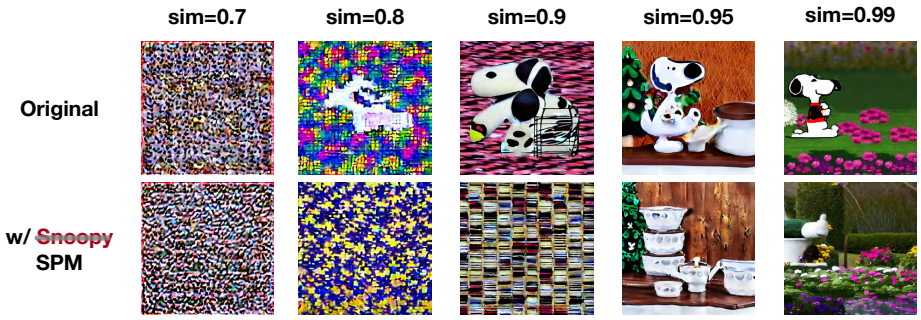

Figure 18. Impact of the *Snoopy*-SPM on semantic representations near the target (**sim** as cosine similarity) in the continuous latent space.

Figure 19. **Additional samples of single concept erasure** with *Snoopy* (top), *Mickey* (middle), and *Spongebob* (bottom), as the targets. While previous methods entail a trade-off between erasing and preservation, SPM allows us to reconcile both aspects.

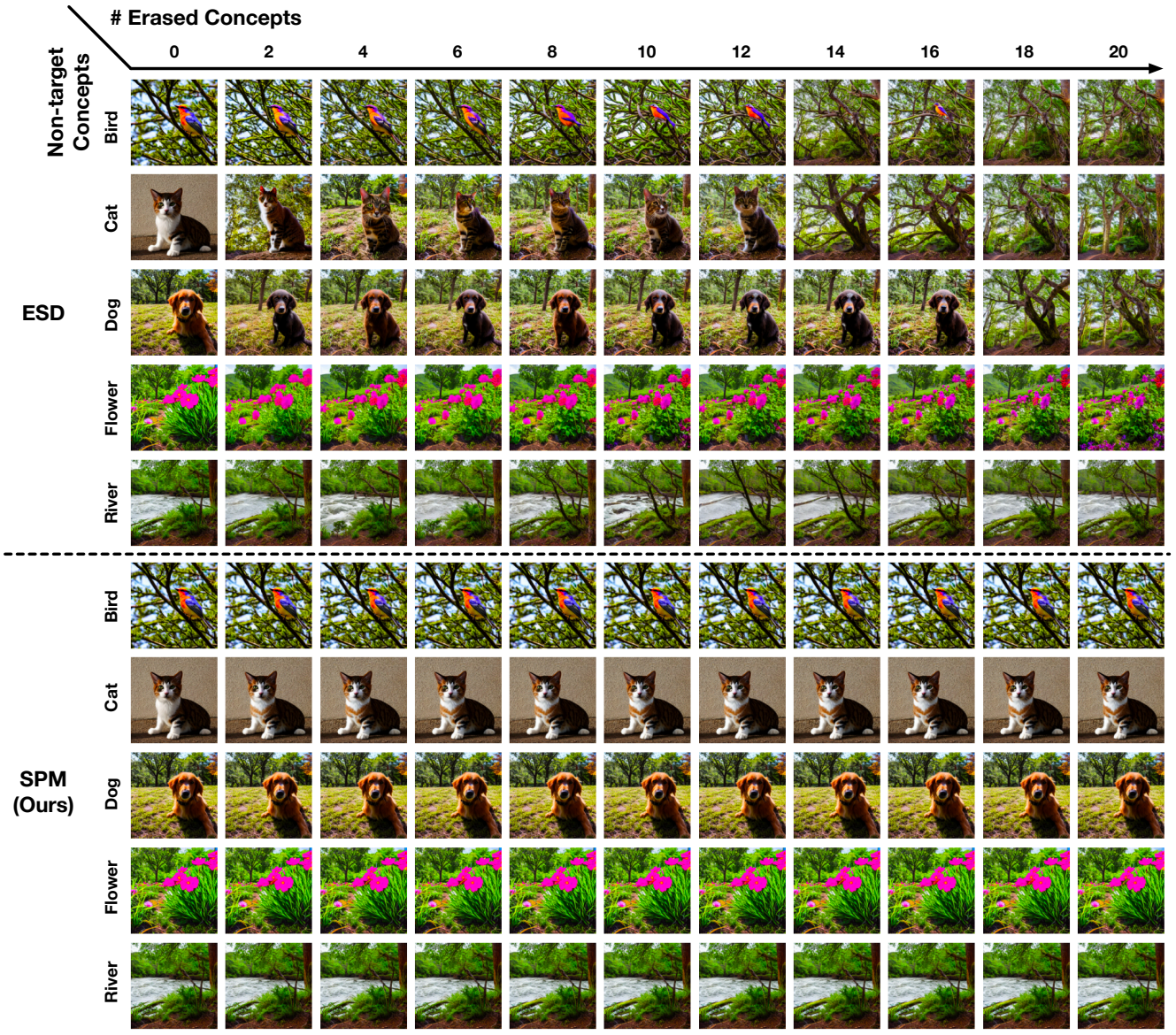

Figure 20. **Additional samples and their generations as 20 concepts are incrementally erased.** With empty prompt an the surrogate concept, the object-centric generation outputs of ESD would be erased towards a few background images, while our results are robust with multiple SPMs overlaid.

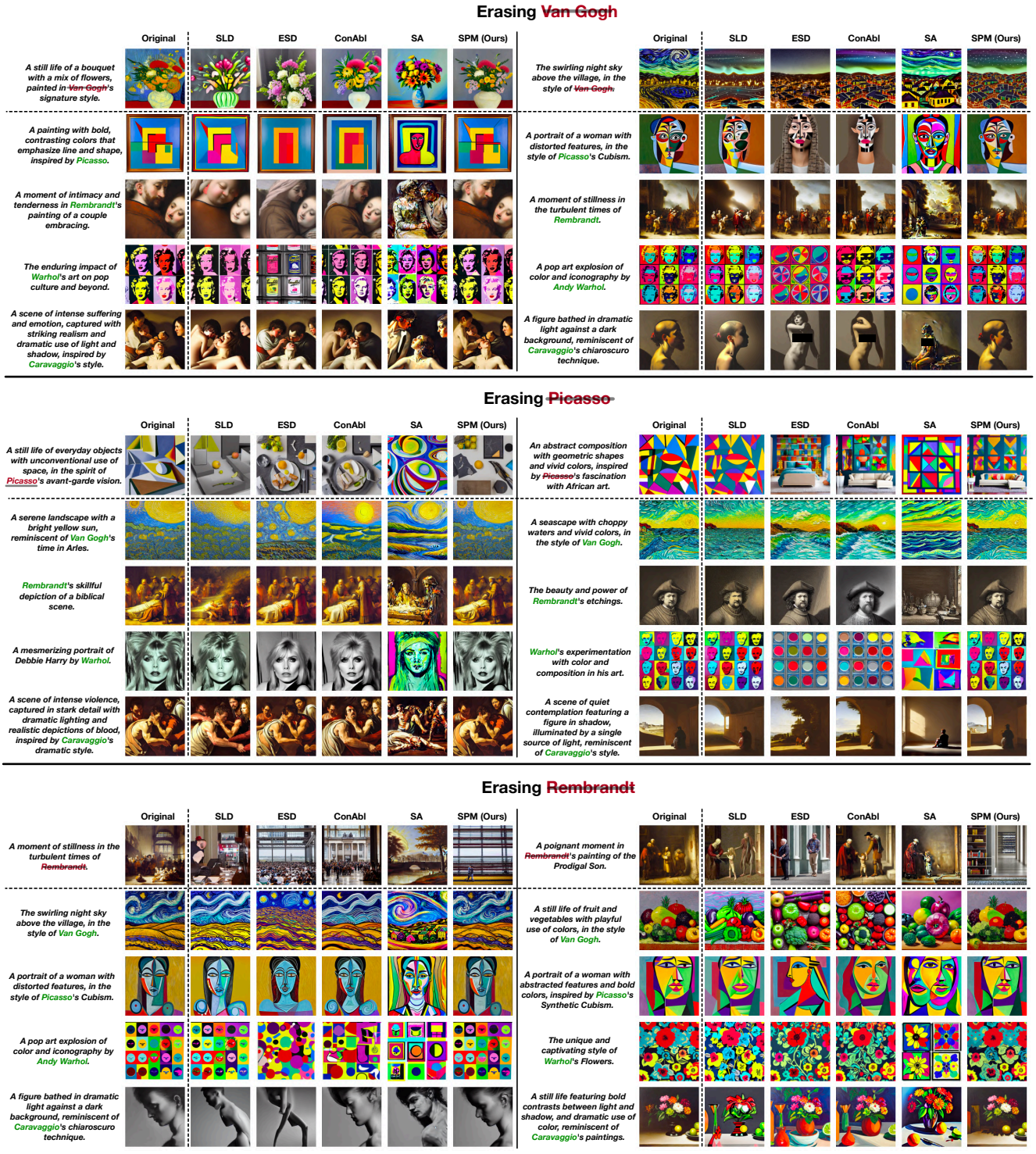

Figure 21. **Additional samples of artistic style erasure** with *Van Gogh* (top), *Picasso* (middle), and *Rembrandt* (bottom), as the targets. Previous studies show deterioration in non-targeted artistic styles under investigation, or underperform with respect to the targeted style. In contrast, SPM gently diminishes the expression of the targeted style while preserving the content, as well as generation consistency of non-targeted styles.

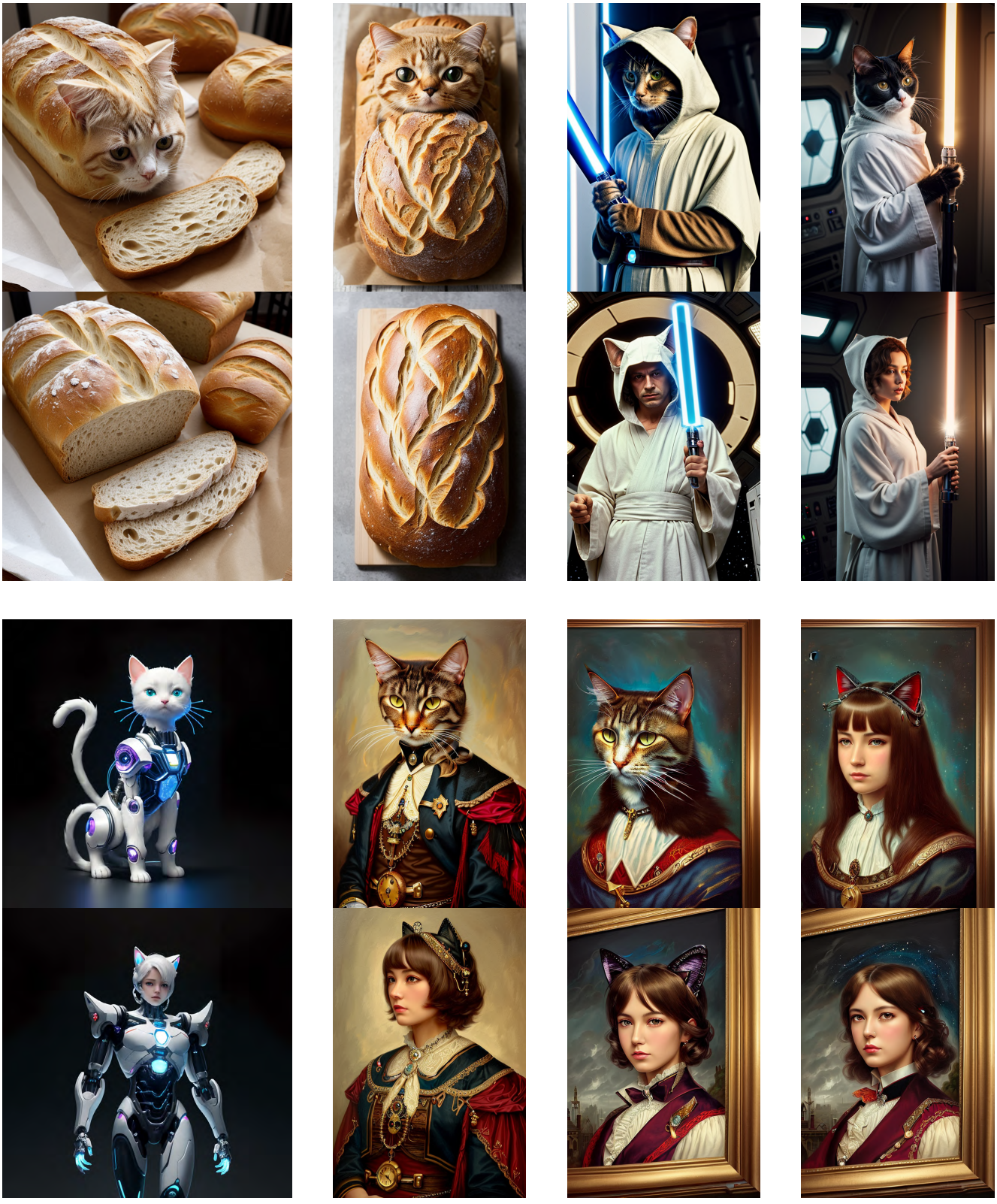

Figure 22. **Training-free transfer samples with a SPM to erase *cat*.** In each pair, the top images show the original results obtained from the community models (1-4 with RealisticVision, 5 with Dreamshaper and 6-8 with ChillOutMix), and the bottom ones are results with the SPM.

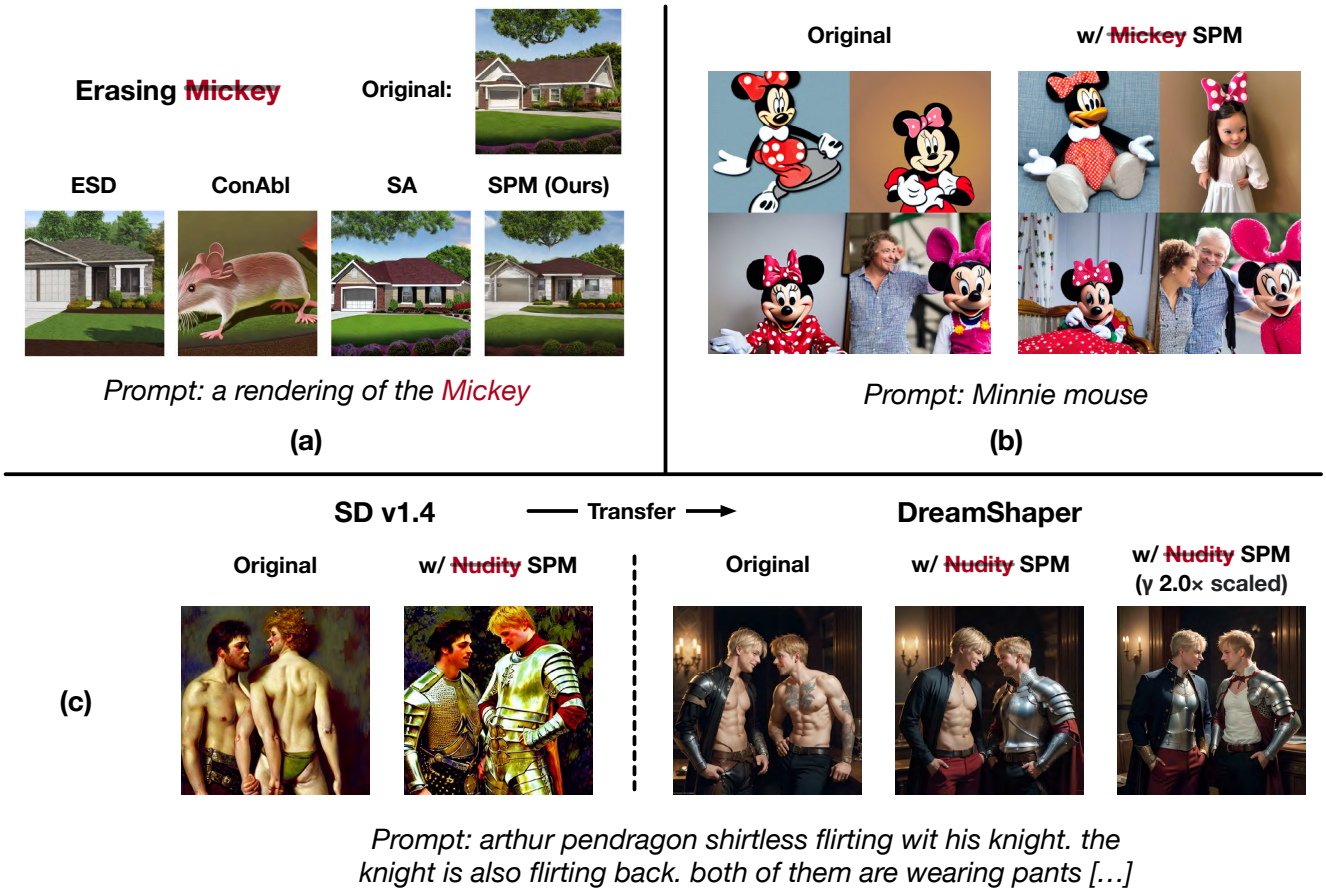

Figure 23. **Suboptimal and failure cases** of (a, b) *Mickey* erasure and (c) *Nudity* erasure.