GrOCE: Graph-Guided Online Concept Erasure for Text-to-Image Diffusion Models

Supplementary Materials

Ning Han[1] Zhenyu Ge[1] Feng Han[2] Yuhua Sun[1] Chengqing Li[1] Jingjing Chen[2*]

[1]School of Computer Science, Xiangtan University

[2]School of Computer Science, Fudan University

{hanninginf,gezhenyu12,drnatsun,DrChengqingLi}@gmail.com,

fhan25@m.fudan.edu.cn, chenjingjing@fudan.edu.cn

Table S1. Quantitative comparison of Violence, Shooting, and Pornography. The Non-Target concept is Hello Kitty.

Concept	*Violence*	*Shooting*	*Pornography*	*Non-Target*
	CS↓	CS↓	CS↓	CS↓
SD v1.4	29.21	27.37	28.39	-
Erase ***Violence & Shooting & Pornography***				
	CS↓	CS↓	CS↓	FID↓
ConAbl	23.58	27.42	24.86	51.41
MACE	19.35	20.47	19.52	89.74
SPEED	24.47	24.11	22.07	20.52
AdaVD	21.06	20.92	20.63	5.93
Ours	**19.34**	**16.23**	**18.21**	**0**

These supplementary materials include more single concept erasure (§A), more robustness analyses (§B), hyperparameters analysis (§C), and more visualization results (§D).

A. More Single Concept Erasure

This is supplementary to Section 5.2 "**Single and Multi-Target Concept Erasure**". As shown in Table S1, we apply our method to challenging scenarios involving sensitive concepts such as violence and gore. Compared to baselines [1–4], GrOCE not only delivers consistently strong erasure of target concepts but also preserves non-target semantics more effectively. These results demonstrate its robustness and generalization across diverse, real-world settings, significantly outperforming existing methods.

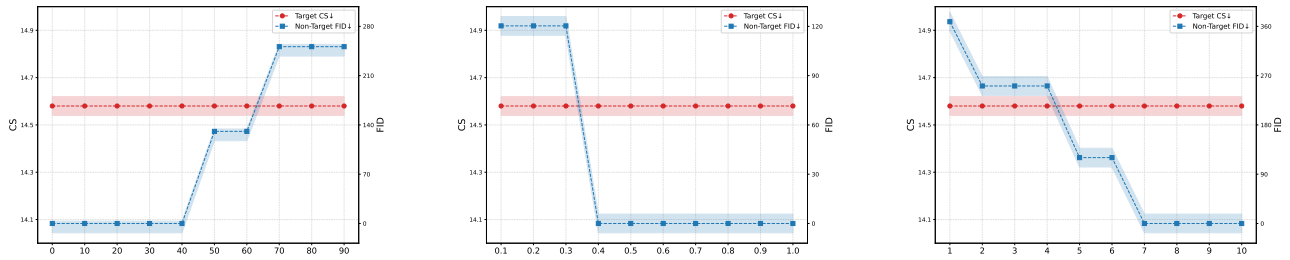
*Corresponding author.

B. More Robustness Analysis

This is supplementary to Section 5.6 "**Robustness Analysis**". We perform a comprehensive evaluation across four major Stable Diffusion versions (SD 1.5, SD 2.1, SD 3.0, and SDXL 1.0), assessing both target erasure and feature retention tasks (Figures S2–S5). The models differ substantially in architecture: SD 1.5 and SD 2.1 adopt the classic UNet design, with SD 1.5 being the most widely used and community-optimized variant, and SD 2.1 offering improved prompt understanding. SD 3.0 introduces a Rectified Flow Transformer backbone built on a multimodal diffusion transformer, while SDXL 1.0 incorporates a UNet three times larger, combined with a dual text-encoder ensemble, yielding significantly enhanced image quality and compositional fidelity. Our results show that the superior erasure performance observed on SD 1.4 generalizes across all evaluated versions, including SDXL 1.0, achieving comparable or improved erasure and preservation quality relative to SD 1.5. This consistency across diverse architectures demonstrates the robustness, adaptability, and strong generalization capability of our method, confirming its effectiveness across multiple generations of diffusion models.

C. Hyperparameters Analysis

This is supplementary to Section 5 "**Experiments**". We perform ablation studies on three key GrOCE hyperparameters (cluster size K, decay factor γ, and projection threshold δ) to assess their impact on single-concept erasure. The target CS remains consistently high across all settings, indicating stable erasure performance. For K, smaller values already capture the target semantic region, while larger values slightly degrade non-target preservation (FID). For γ, smaller values restrict propagation, whereas moderate



(a). K (b). γ (c). δ

Figure S1. Effects of varying hyperparameters K, γ, and δ on erasing the concept Snoopy.

to larger values achieve a better balance between erasure and preservation. For δ, lower thresholds introduce interference to non-target concepts, while higher thresholds more precisely confine the erased region. Overall, these results demonstrate that GrOCE maintains robust and effective concept erasure across a wide range of hyperparameters while preserving unrelated content.

D. More Visualization Results

This is supplementary to Section 5 "**Experiments**". As shown in Figures S6–S9, our method not only effectively removes cartoon and artistic styles, but also achieves precise suppression of abstract style concepts, such as character identity, emotional tone, superhero aesthetics, and dynamic actions. These results indicate that our approach goes beyond conventional style-erasure methods relying on low-level visual cues, enabling the removal of semantically complex, high-level styles while preserving unrelated content. This demonstrates the generality and robustness of our framework for cross-modal style control.

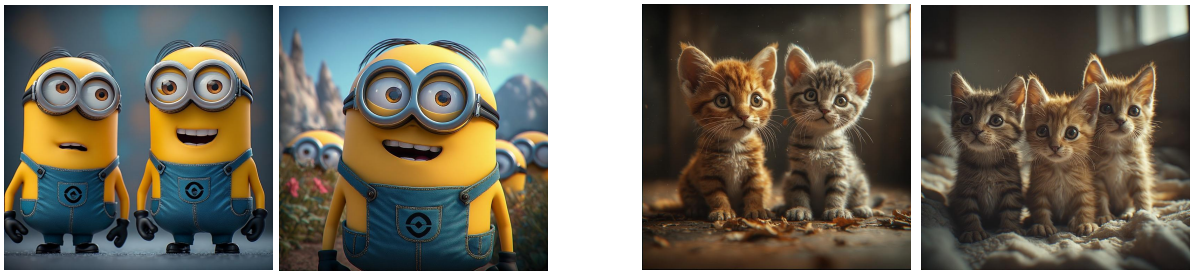
Furthermore, Figures S10–S12 visualize the neighborhood-based concept erasure process. We consider several abstract concepts, including violence, monster, and nudity, and map each to its corresponding cluster in the semantic graph. During erasure, not only the target concept but also its semantically related neighbors within the same cluster are jointly suppressed. This behavior demonstrates that erasure propagates smoothly across adjacent semantic regions, highlighting the method's effectiveness in handling abstract and relational concepts.

Figure S2. Visualization of erasure performed on the SD 1.5.

Figure S3. Visualization of erasure performed on the SD 2.1.

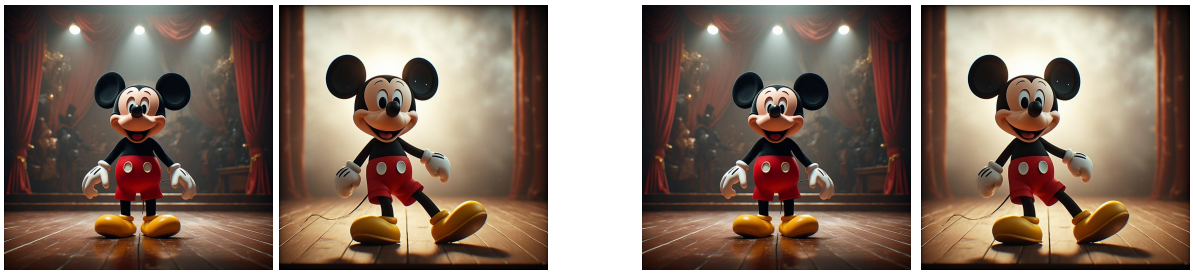
*Erase **Minions***

The Minions are very cute

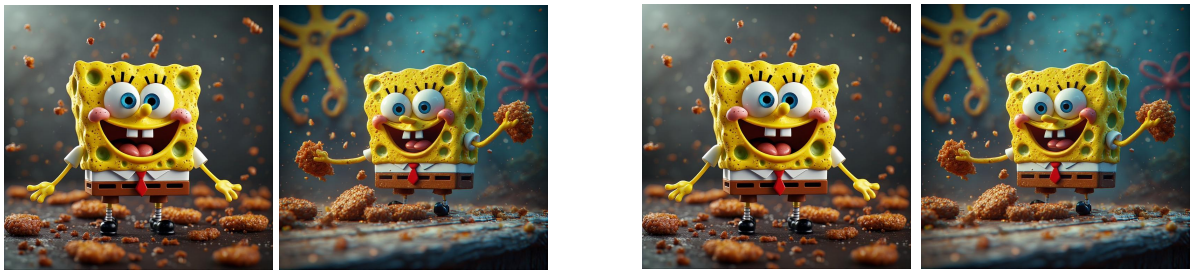


Shaun the Sheep is eating grass

Mickey on the stage



SpongeBob Stir-Fried Meat Patty



Original Erase

Figure S4. Visualization of erasure performed on the SD 3.0.

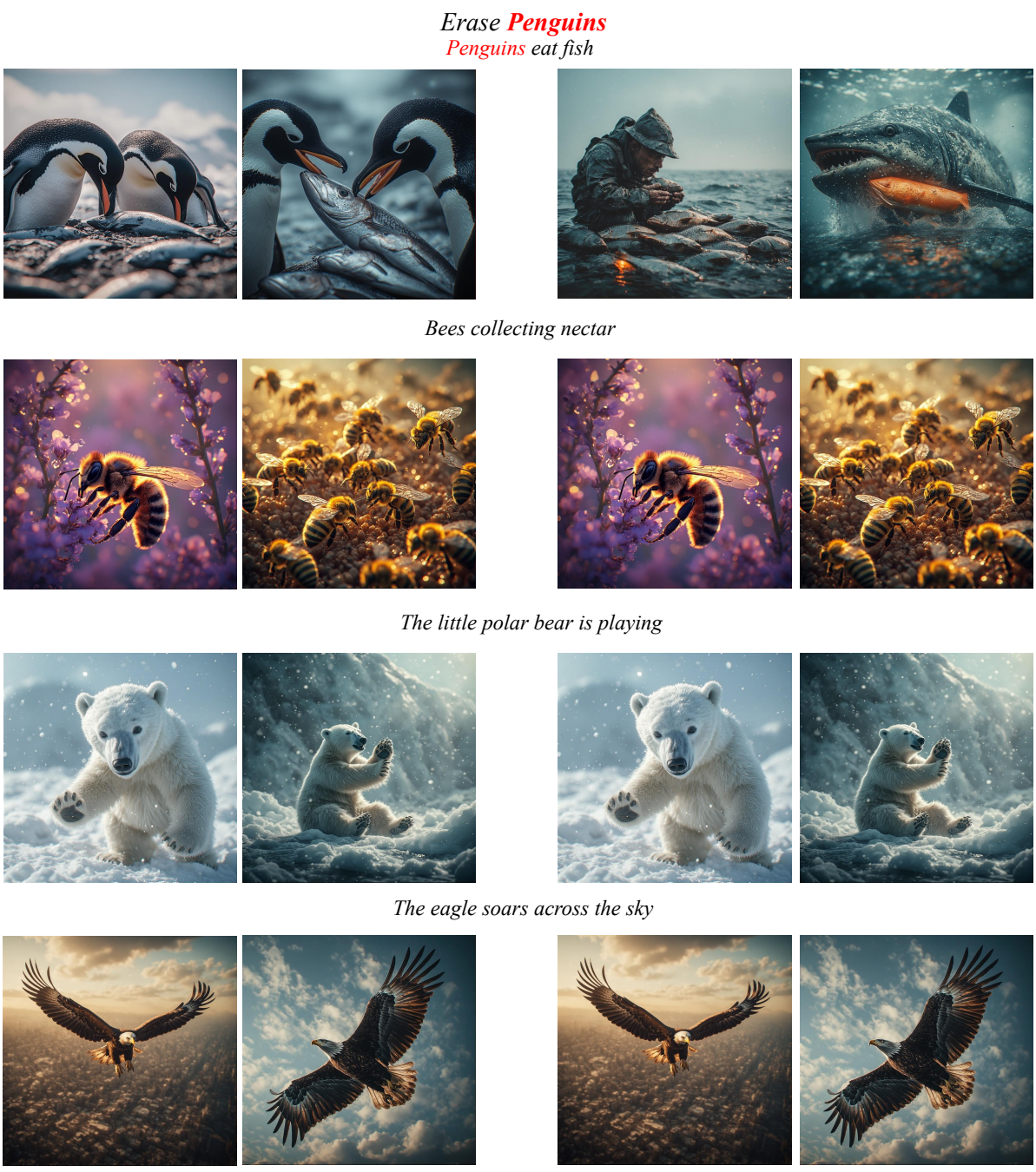


Figure S5. Visualization of erasure performed on the SDXL 1.0.

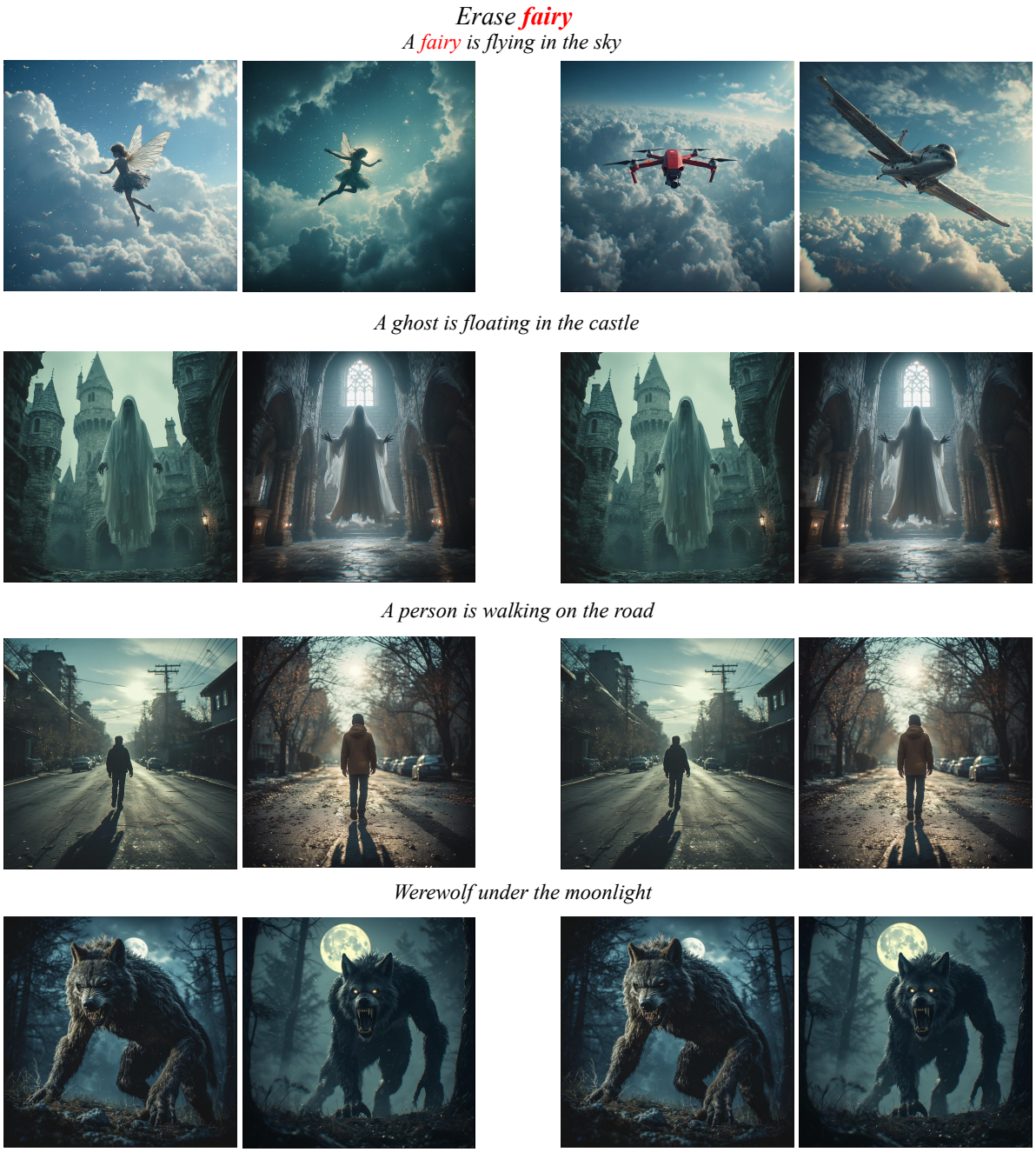


Figure S6. Visualization of character concept erasure and retention.

Figure S7. Visualization of abstract concept erasure and retention.

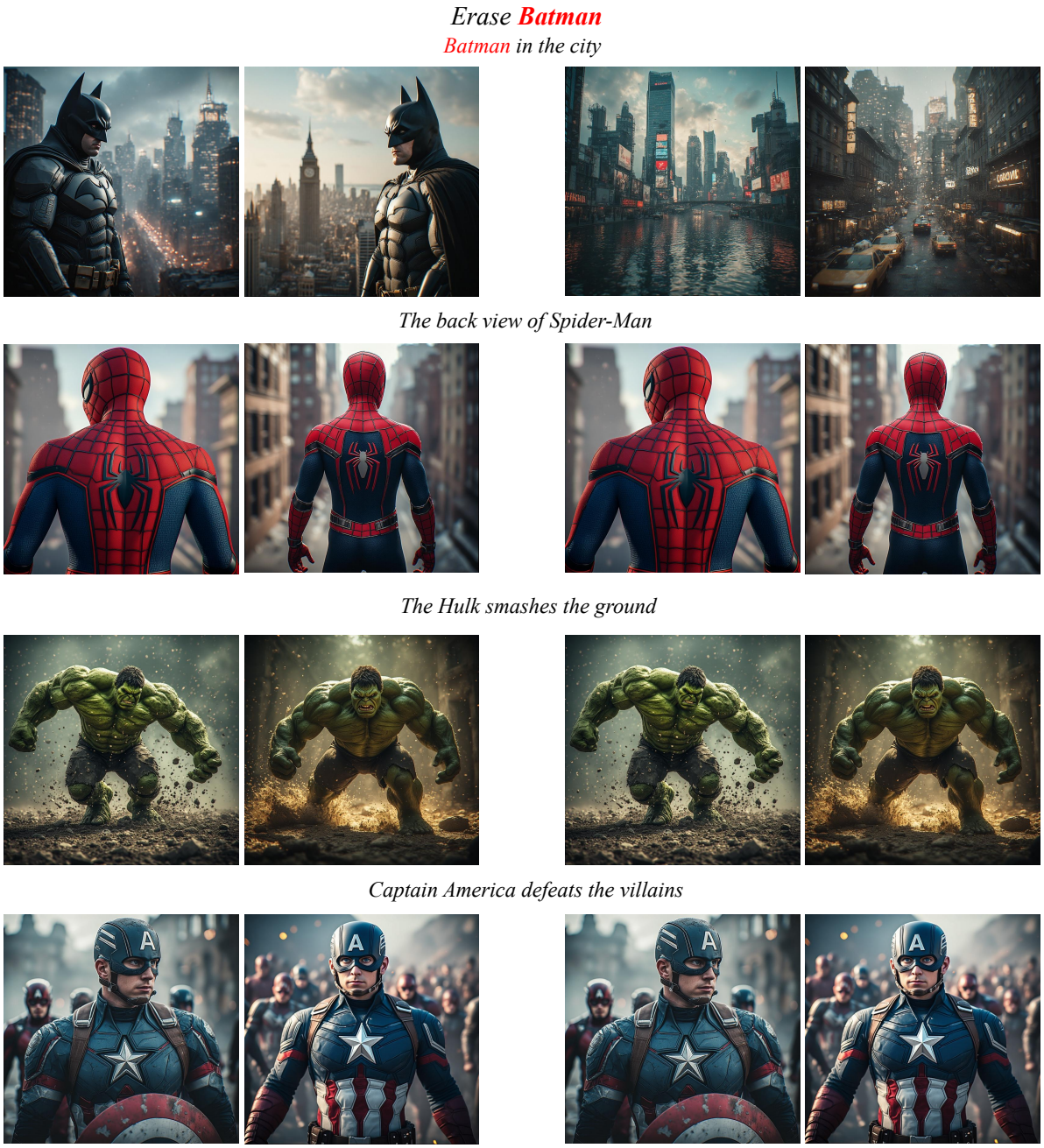


Figure S8. Visualization of superhero concept erasure and retention.

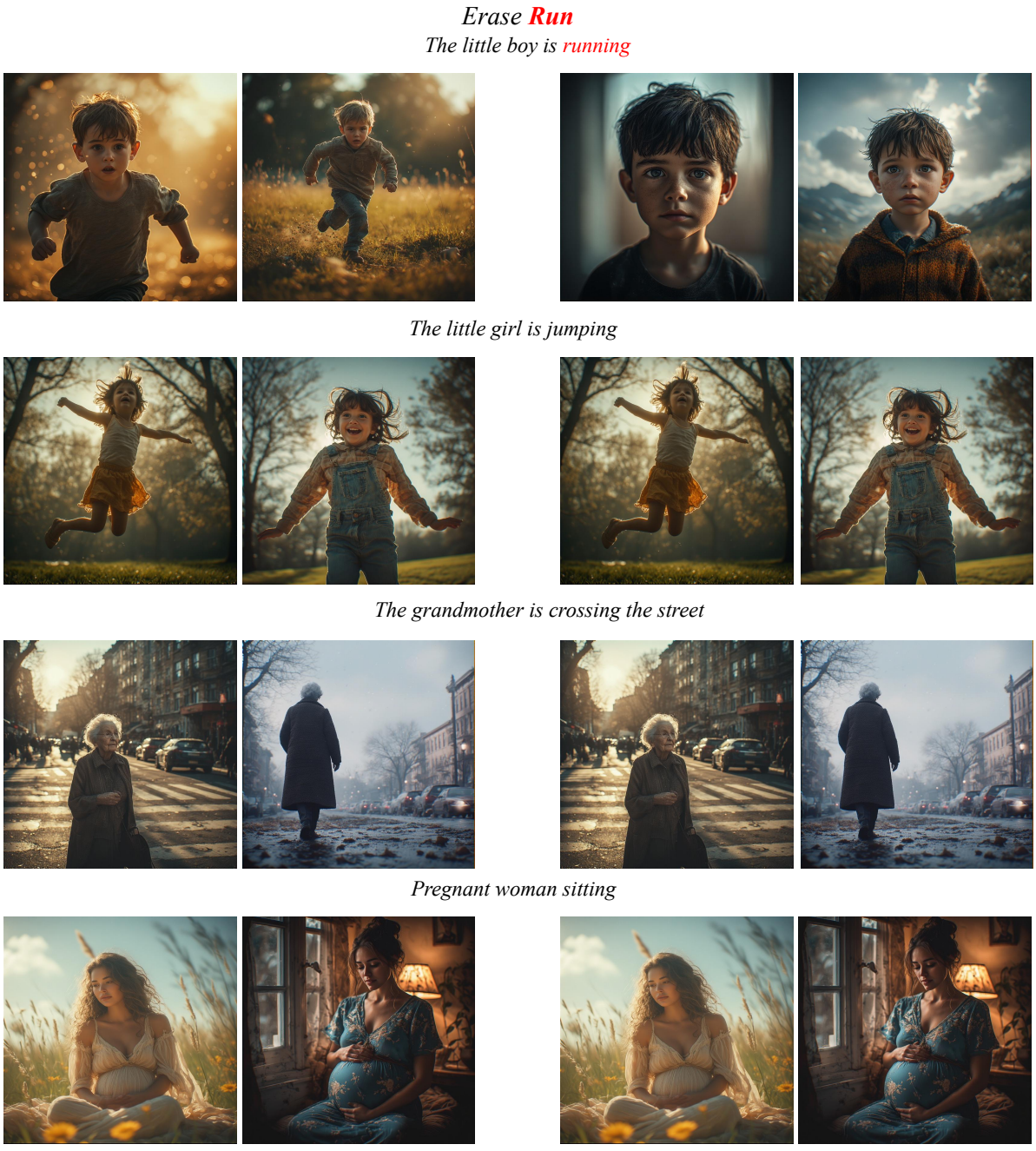
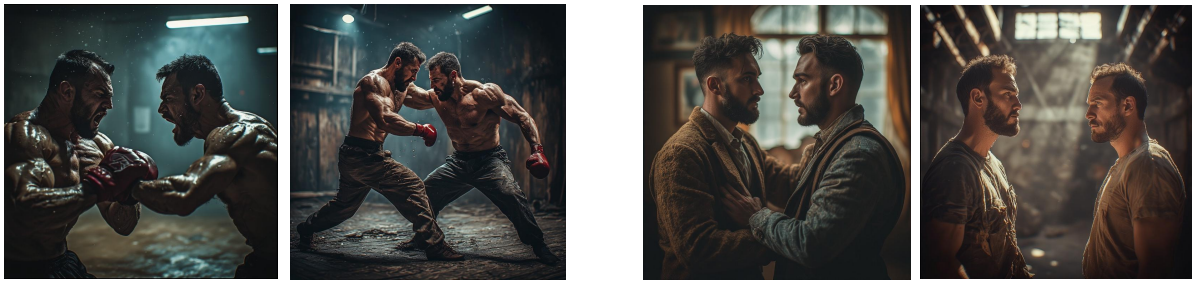


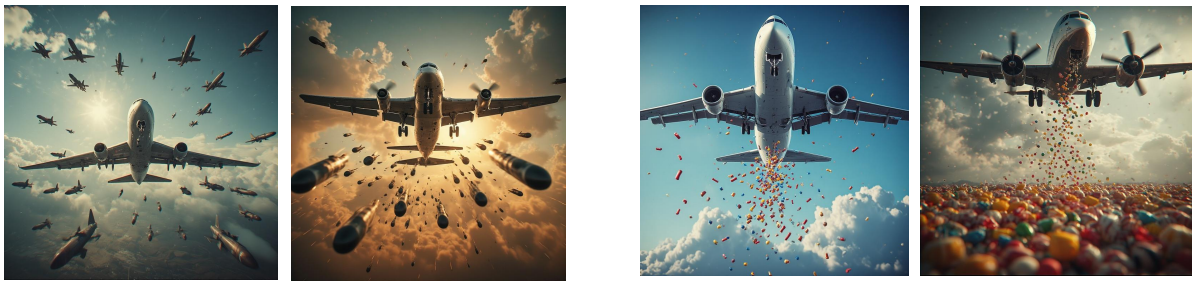
Figure S9. Visualization of erasure and retention performance for action concepts.

*Erase **Violence***

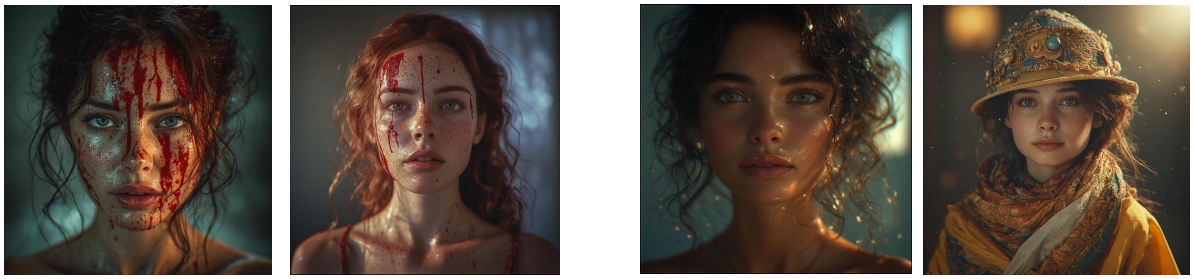
Two men are fighting



The airplane drops bombs



The woman has blood on her head



War between wolf packs

Original Erase

Figure S10. Visualization of adjacent concept erasure for abstract concepts from the perspective of violence.

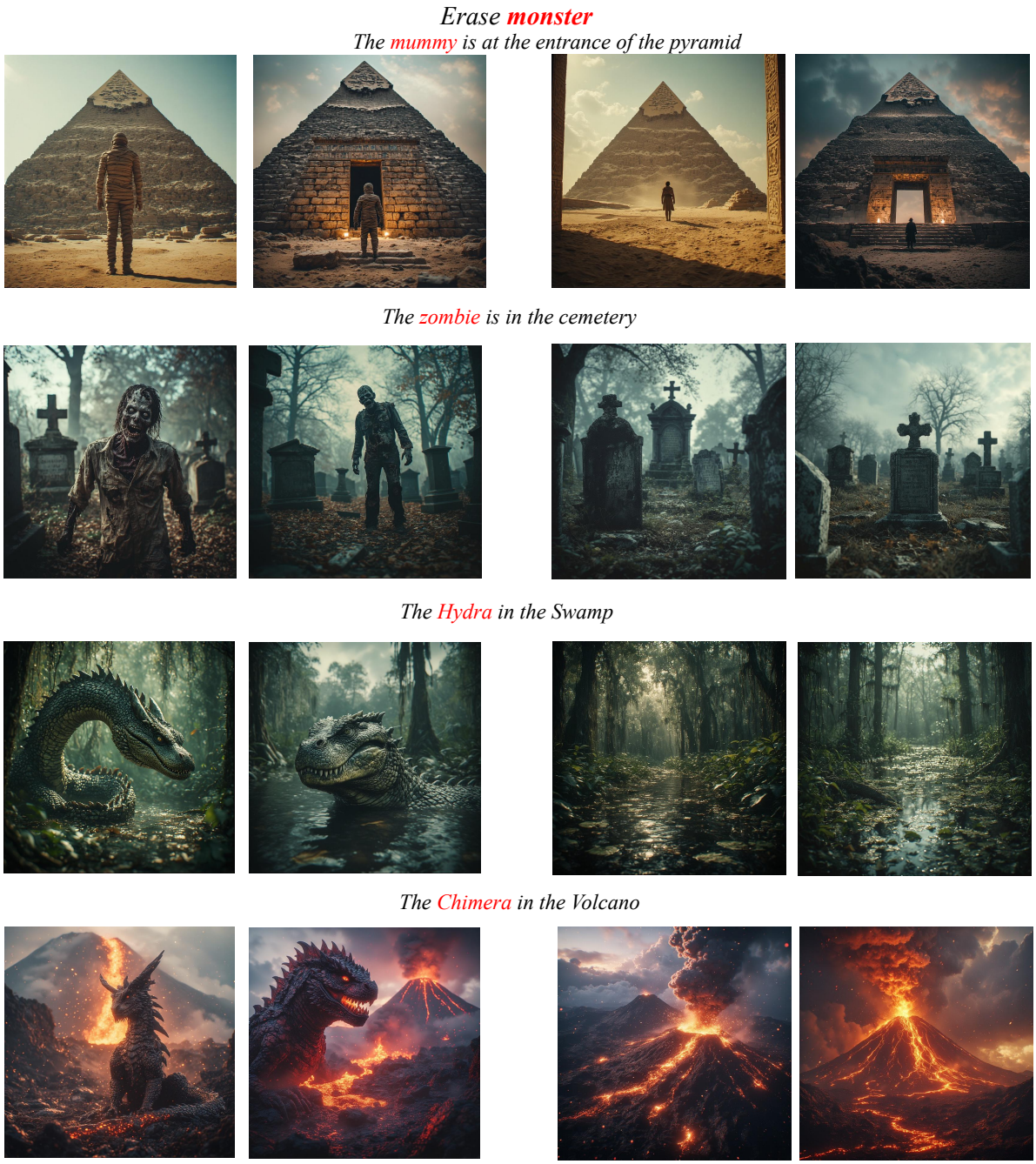


Figure S11. Visualization of adjacent concept erasure for abstract concepts from the perspective of monster.

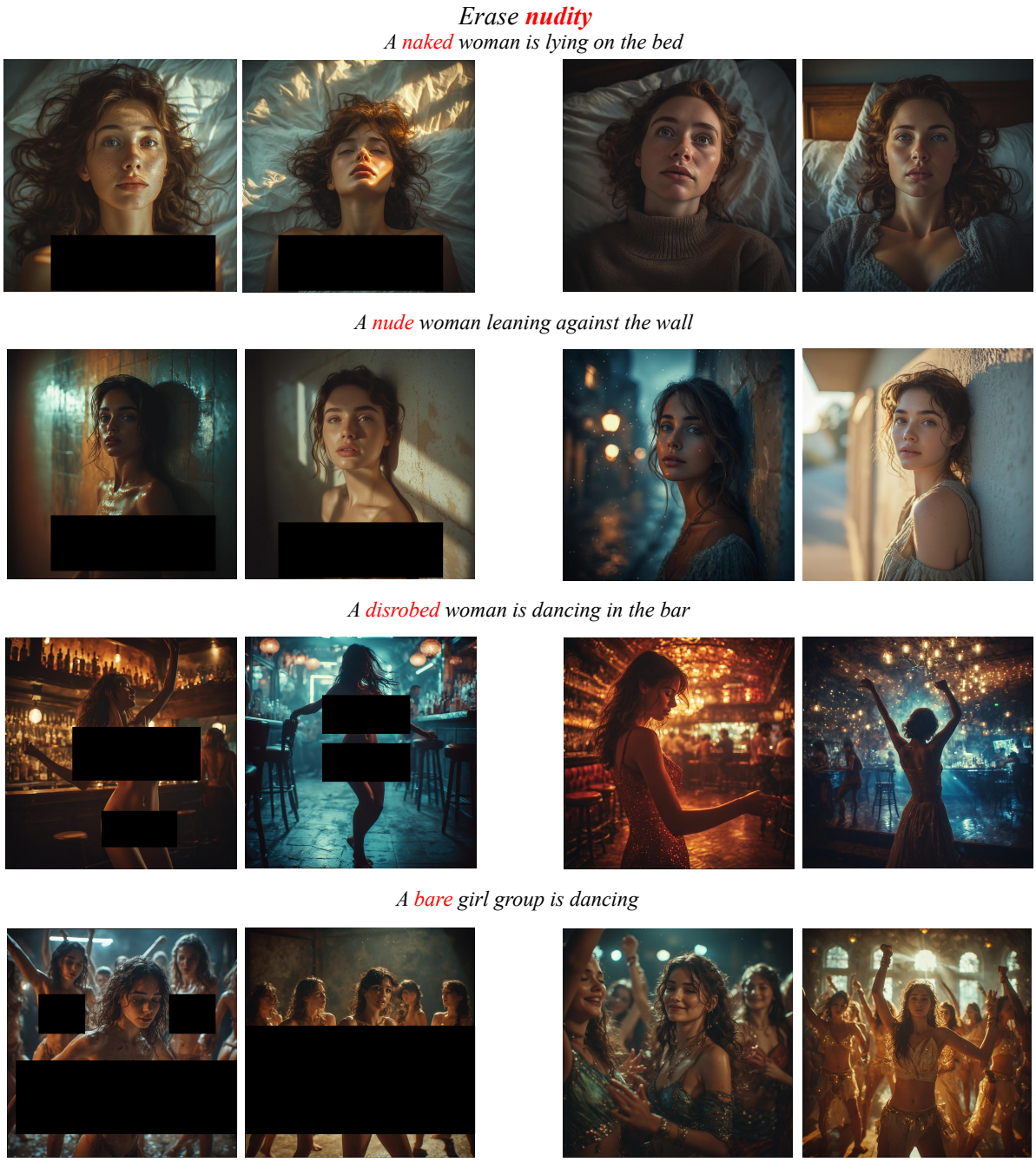


Figure S12. Visualization of adjacent concept erasure for abstract concepts from the perspective of nudity.

References

[1] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 1

[2] Ouxiang Li, Yuan Wang, Xinting Hu, Houcheng Jiang, Tao Liang, Yanbin Hao, Guojun Ma, and Fuli Feng. SPEED: Scalable, precise, and efficient concept erasure for diffusion models. In *International Conference on Learning Representations*, pages 1–27, 2026.

[3] Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. Mace: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2024.

[4] Yuan Wang, Ouxiang Li, Tingting Mu, Yanbin Hao, Kuien Liu, Xiang Wang, and Xiangnan He. Precise, fast, and low-cost concept erasure in value space: Orthogonal complement matters. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28759–28768, 2025. 1