Concept Replacer: Replacing Sensitive Concepts in Diffusion Models via Precision Localization

Supplementary Material

6. Ablation Study

In this section, we present the ablation studies on concept location and replacing, which illustrate the impact of each design.

Ablation on Concept Localizer. In Table 5, we present the ablation study results on the CelebA-Mask-HQ datasets using 10-shot training. ”Ours-L” denotes the concept localizer that employs low-resolution cross-attention layers with spatial dimensions under 32. ”Ours-H” signifies the concept localizer incorporating both refined low and high-resolution cross-attention layers. ”Ours-T” represents our final concept localizer, combining the ”Ours-H” setup with average timesteps. We utilize the average of $T = 5, 50, 100$ timesteps for real image segmentation. For concept localization during the denoising process, we calculate the average over $T = 666, 726, 766$ timesteps. This choice is guided by the requirement that concept replacement must occur during the early stages of the denoising process, as illustrated in Figure 9.

Ablation on Replacing Timestep. In Figure 9, we demonstrate different timesteps utilized for replacing the concept ”Brad Pitt” with ”Leonardo DiCaprio.” Among 1000 timesteps, we picked specific points for this replacing process. For $T = 0$, it refers to the initial image generated using the prompt ”a photo of Brad Pitt.” From $T = 0$ to $T = 900$, there is relatively minimal semantic change, whereas for T exceeding 900, the semantic alteration becomes significant. This suggests that low-frequency semantic information is established early in the denoising process when T is large, while high-frequency details emerge when T is small. In our experiments, we replaced the concept at $T = 666$ to achieve a balance between the replacement effect and the preservation of the overall structure.

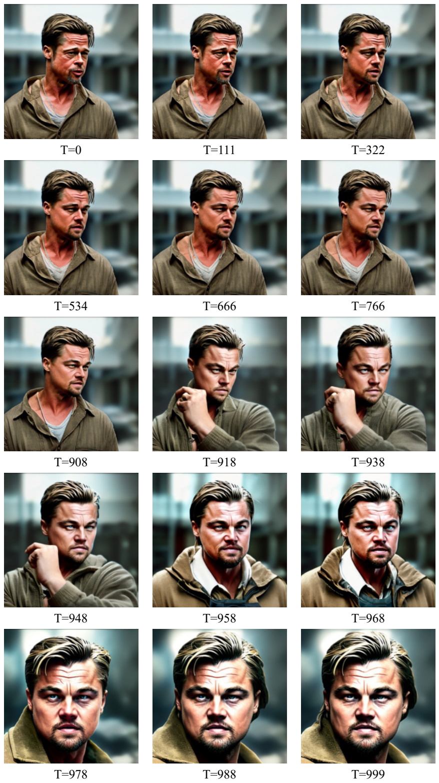

Figure 9. Replacing Brad Pitt with Leonardo DiCaprio at Varying Timesteps.

7. Inference Efficiency

In Figure 10, we evaluate inference efficiency on 512×512 image generation using 100 diffusion steps (NVIDIA RTX A6000). Since our localizer only activates at predefined timesteps [766, 726, 666], we maintain near-baseline throughput with little computational overhead.

8. More Comparing Results

We added Forget-Me-Not[49] and SALUN[10] for more comparison. Our method focuses on localization and replacement during inference, while those two remove harmful concepts via retraining. Figure 11 demonstrates our superior nudity replacement, outperforming Forget-Me-Not that either incompletely remove harmful concepts. SALUN compromises consistency with the original SD model and exhibits a certain degree of overfitting, often generating images containing a man without considering the context.

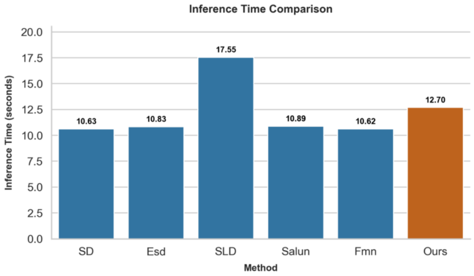

Figure 10. Inference Time Comparing with Other Methods.

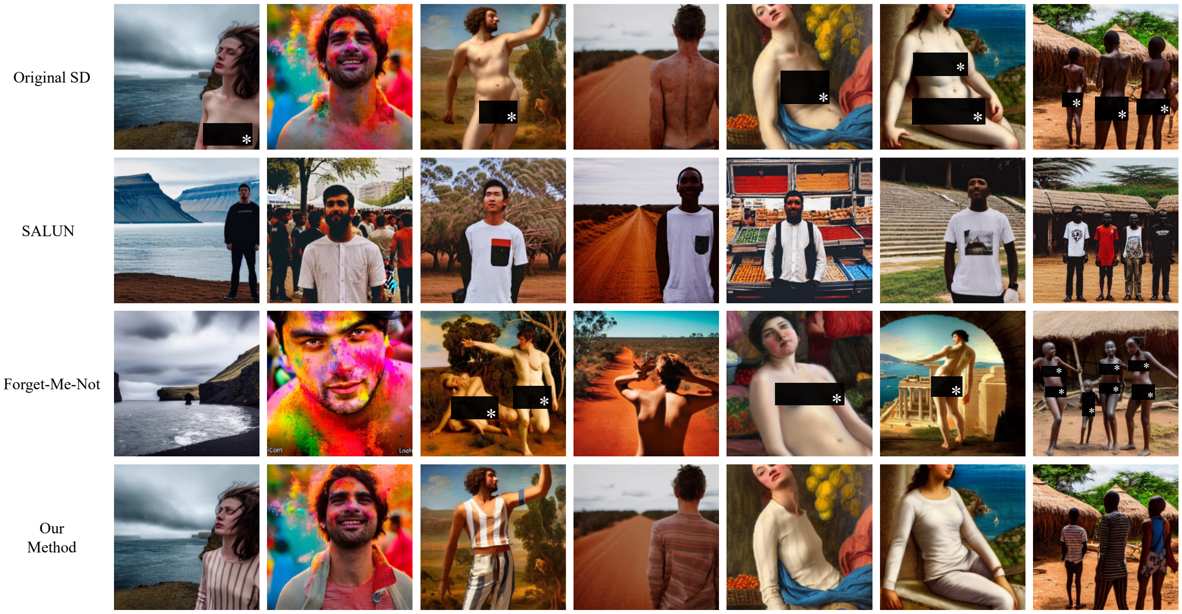

Figure 11. Additional Comparison Results with SALUN and Forget-Me-Not.

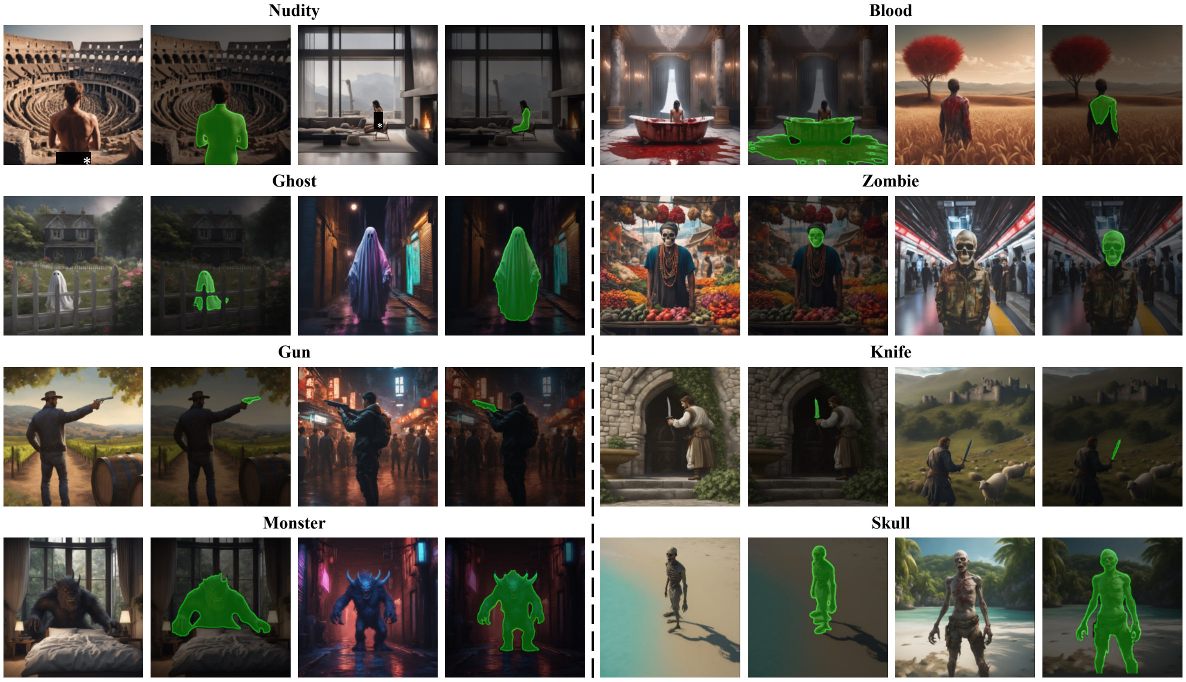

Figure 12. Additional Results of Concept Location.

9. Additinoal Results on Concept Location and Replacing

In Figure 12, we showcase the location results for various concepts, highlighting our method's capability to accurately identify and localize concepts across a wide range of sizes and complexities. This demonstrates the adaptability and robustness of our approach in handling diverse concepts effectively during the image generation process. Figure 13 depicts replacing ”Elon Musk” with various faces. Figure 14 demonstrates the substitution of the nudity concept with multiple alternatives. Collectively, these results validate the

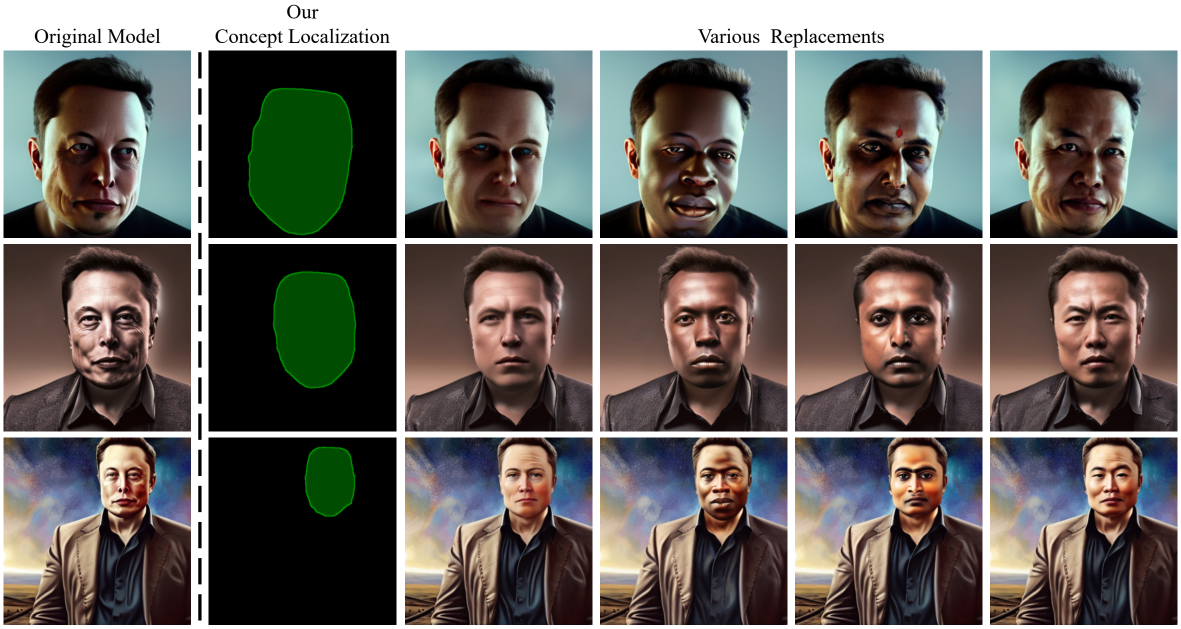

Figure 13. Additional Results of Elon Musk Replacing with Different Alternatives.

	Cloth	Eyebrow	Ear	Eye	Hair	Mouth	Neck	Nose	Face	Background	Average
Ours-L	64.5	63.8	65.6	72.0	86.2	83.3	81.0	82.0	90.1	86.6	77.5
Ours-H	66.8	63.4	64.4	73.2	85.9	83.0	82.1	81.7	90.1	87.7	77.8
Ours-T	67.1	63.7	65.7	72.6	86.4	83.0	82.5	81.0	90.0	87.9	78.1

Table 5. Ablation Study on Concept Localizer with Different Configurations.

effectiveness of our approach in achieving accurate localization and harmonious replacement, reinforcing its potential for targeted concept manipulation in diffusion models.

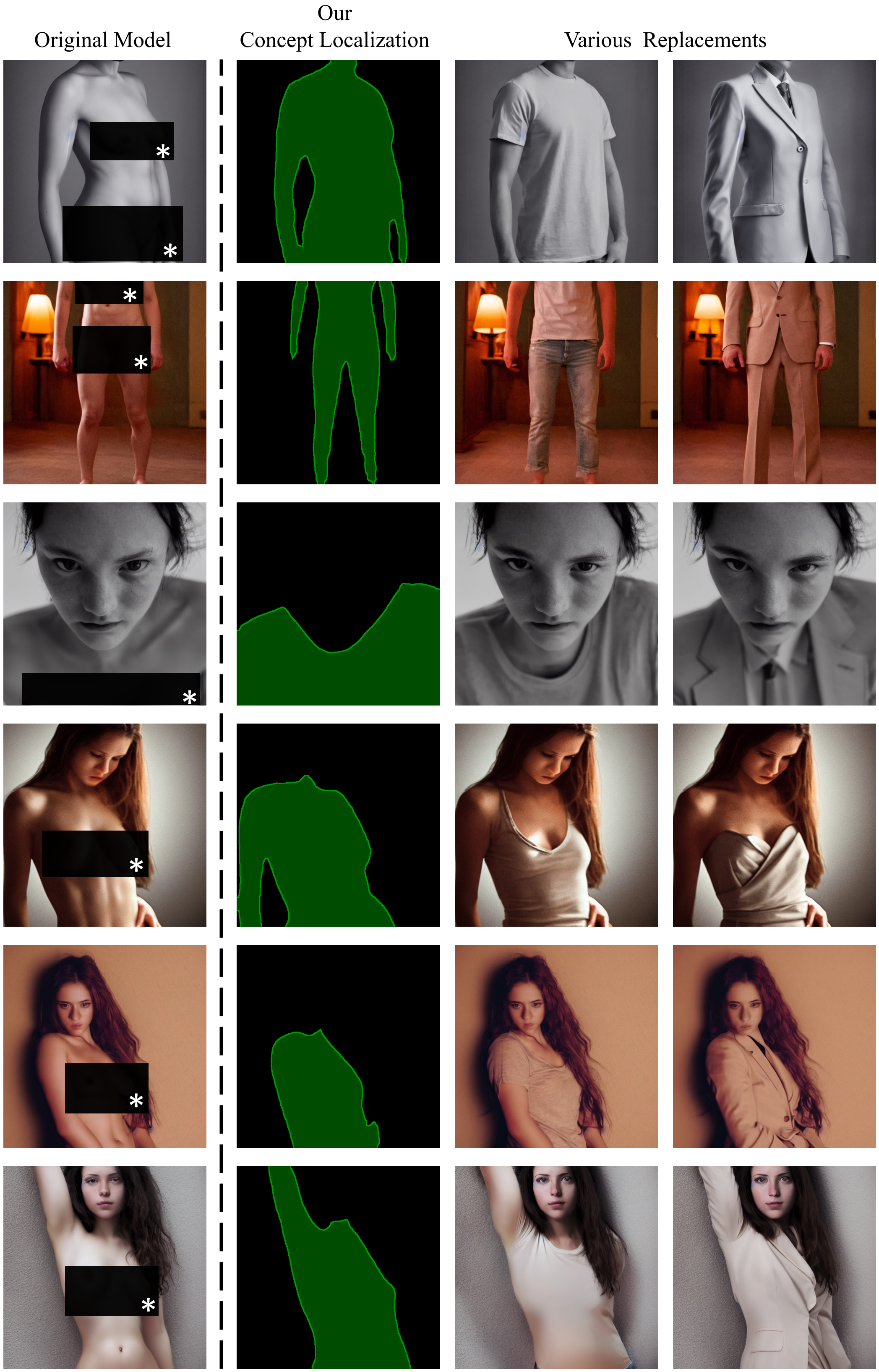

Figure 14. Additional Results of Nudity Replacing with Various Replacements.