

Silence is Golden: Leveraging Adversarial Examples to Nullify Audio Control in LDM-based Talking-Head Generation

Supplementary Material

A. More Experiments

A.1. More Implementation Details

The resolution of our input portrait is 512×512 . The audio used for training in our experiment is a four-second clip. For testing on CelebA-HQ, the audio length is seven seconds. In the case of TalkingHead-1KH, the audio length varies between three and seven seconds. In our experiment, the DDIM inversion step is set to 20. Due to the limitation of GPU memory, we optimize only the inverted latent feature from the final step. All experiments can be conducted using a single NVIDIA A40 GPU.

A.2. Evaluating the Transferability of Silencer

To evaluate the transferability of **Silencer (S-I and S-II)**, we performed a cross-model evaluation. Adversarial noise was optimized on the Hallo model and subsequently tested on other LDM-based talking-head generation models. Specifically, we randomly selected 20 portraits from the TalkingHead-1KH dataset and generated talking-head videos using the publicly available EchoMimic [6] and Hallo2 [1]. As shown in Table 5, the synchronization values of the generated videos demonstrate that Silencer maintains a significant adversarial effect even when applied to models different from the one used for optimization. Although **Silencer** is designed as a white-box attack, these results highlight its notable generalization capability across various LDM-based talking-head models. This cross-model robustness suggests the potential for broader applicability and further validates the effectiveness of our method. A likely explanation for the observed cross-model effectiveness of Silencer is a combination of factors. First, these LDM-based talking-head models share similar architectural designs. Second, and perhaps more crucially, they are all fine-tuned upon Stable Diffusion. This common foundation could introduce common weaknesses or biases that Silencer is able to exploit, even across different models.

A.3. Efficiency Analysis

We evaluated the computational efficiency on an NVIDIA A40 GPU. The results, shown in Table 6, demonstrate a significant difference in Silencer-I and Silencer-II. Silencer-I exhibits superior efficiency, requiring considerably less computational time compared to Silencer-II. This difference in efficiency stems primarily from the architectural design of Silencer-II. Unlike Silencer-I, Silencer-II incorporates an optimization step within the latent space of an

| Method | GT | AdvDM(+) | Mist | SDST(-) | S-I | S-II |
|---------------|--------|----------|--------|---------|--------|--------|
| EchoMimic [6] | 4.0365 | 1.8252 | 1.7839 | 2.2228 | 1.4601 | 0.9973 |
| Hallo2 [1] | 5.6661 | 3.2136 | 3.0679 | 3.9238 | 1.5952 | 2.0783 |

Table 5. **Evaluating the Transferability of Silencer.** Synchronization scores demonstrating cross-model transferability of Silencer (S-I and S-II). Videos were generated by EchoMimic [6] and Hallo2 [1] using original (GT) and adversarial inputs. Lower scores signify greater disruption. Despite being optimized on Hallo, Silencer significantly impacts both models.

| | AdvDM(+) | PhotoGuard | Mist | SDS(-) | SDST(-) | S-I | S-II |
|------|----------|------------|------|--------|---------|-----|------|
| time | 59 | 34 | 59 | 22 | 40 | 64 | 241 |

Table 6. **Efficiency Analysis.** Average time (seconds/image) required for different protection methods.

| DiffPure timesteps | 50 | 100 | 150 |
|--------------------|--------------|--------------|--------------|
| Silencer-I | 30.65/0.2606 | 29.26/0.2540 | 28.13/0.2691 |
| Silencer-II | 27.80/0.4057 | 27.26/0.3909 | 26.82/0.3504 |

Table 7. **Ablation on Timesteps of DiffPure [3].** We present I-PSNR/LPIPS scores for Silencer-I and Silencer-II after applying DiffPure with varying timesteps. Red values highlight greater robustness.

additional LDM. This additional optimization process introduces a substantial computational overhead, increasing the overall time required for Silencer-II to generate adversarial examples. While this optimization contributes to more robust perturbations, it comes at the cost of reduced computational efficiency. Silencer-I, by contrast, avoids this extra optimization step, leading to a more streamlined and faster process. While Silencer-I takes 64 seconds per image, its runtime is comparable to other methods like AdvDM(+) and Mist (59 seconds). This makes Silencer-I a more practical choice in scenarios where computational resources are limited or where rapid generation of adversarial examples is critical. Notably, SDS(-) demonstrate significantly faster runtimes, due to skipping the UNet portion of the gradient calculation. However, whether such an optimization can be effectively and reliably applied within an LDM-based talking-head network to improve efficiency remains an open challenge for future research.

| GrIDPure timesteps | 5 | 10 | 15 |
|--------------------|---------------------|---------------------|---------------------|
| Silencer-I | 28.35/0.1672 | 28.16/0.1698 | 27.93/0.2016 |
| Silencer-II | 25.81/0.3451 | 25.72/0.3511 | 25.59/0.3610 |

Table 8. **Ablation on Timesteps of GrIDPure [5].** We present I-PSNR/LPIPS scores for Silencer-I and Silencer-II after applying GrIDPure purification. GrIDPure was run for 20 iterations with initial timesteps of 5, 10, and 15. Red values highlight greater robustness.

| | DiffAudio | SameAudio |
|--------------|-----------|-----------|
| Silencer-II | 3.9685 | 2.4926 |
| Ground Truth | 6.4041 | 5.7509 |

Table 9. **Impact of Audio Consistency on Silencer-II while Training and Testing with CelebA-HQ.** "DiffAudio" denotes using different audio for training and testing, while "SameAudio" uses the same audio. Lower Sync value is better.

| l_{inf} | V-PSNR/SSIM↓ | FID↑ | Sync↓ | M-LMD↑ |
|-----------|---------------------|---------------|---------------|---------------|
| 8/255 | 19.59/0.5768 | 78.78 | 4.8368 | 2.0444 |
| 16/255 | 19.02/0.5104 | 124.07 | 4.0644 | 2.2008 |

Table 10. **Ablation Study of l_{inf} Perturbation Budgets in Silencer-I on CelebA-HQ.**

| Inverted Timesteps | V-PSNR/SSIM↓ | FID↑ | Sync↓ | M-LMD↑ |
|--------------------|---------------------|---------------|---------------|---------------|
| the last one | 19.01/0.5111 | 156.99 | 3.9685 | 2.2108 |
| the last two | 19.30/0.5402 | 111.99 | 4.4579 | 2.1731 |

Table 11. **Ablation Study of Inverted Timesteps in Silencer-II on CelebA-HQ.**

A.4. More Ablation Study

Ablation Study on Timesteps in Purification Methods. Our anti-purification experiments are conducted using the publicly available implementation¹. For DiffPure, we set the diffusion timestep to 100, while for GrIDPure, we use a timestep of 10 with 20 iterations. We conduct the ablation experiments on different settings of diffusion-based purification. Table 7 and Table 8 illustrate the effectiveness of Silencer-I and Silencer-II against image purification techniques, specifically DiffPure and GrIDPure, across different timesteps. The tables compare I-PSNR and LPIPS scores for images processed by both Silencer versions. While larger timesteps in these purification methods improve the smoothness of the resulting images, they fail to completely remove the perturbations introduced by Silencer-II. This

¹<https://github.com/zhengyuezhao/gridpure>

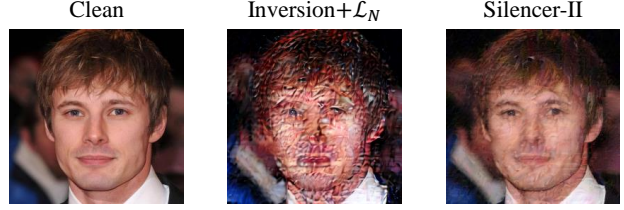


Figure 7. **Ablation Study on \mathcal{L}_T in Silencer-II.** Without the assistance of \mathcal{L}_T , the generated perturbation becomes highly noticeable, significantly compromising the facial identity.

| s | 50 | 75 | 100 | 125 | 200 |
|---------|---------------|---------------|---------------|---------------|---------------|
| I-SSIM↑ | 0.7125 | 0.6998 | 0.6918 | 0.6844 | 0.6704 |
| FID↑ | 136.88 | 166.10 | 173.18 | 171.08 | 193.46 |
| Sync↓ | 5.5725 | 5.0413 | 4.0602 | 4.1832 | 4.0791 |
| M-LMD↑ | 1.8559 | 2.1371 | 2.2053 | 2.3748 | 2.3563 |

Table 12. **Ablation Study on the Initial Iteration s without Mask.** Larger iterations without the face mask lead to better protection performance with lower image quality.

highlights the robustness of our approach.

Ablation Study on Audio and Portrait in the Training and Testing of CelebA-HQ. For audio, We investigated the effect of using the same versus different audio inputs during the training and testing phases. This tests whether Silencer is overly sensitive to specific audio characteristics or if it can generalize to unseen audio. As shown in Table 9, both scenarios resulted in a reduction of the synchronization value compared to the ground truth. The decrease in synchronization demonstrates that Silencer effectively disrupts synchronization regardless of whether the audio input is consistent between training and testing. This finding highlights the robustness of the Silencer method to variations in audio input, suggesting that it is not overfitting to specific audio features.

For the starting portrait, we conducted experiments on 50 different portraits of CelebA-HQ in Table 1. The average sync value is 3.9685 and the standard deviation is 1.5607. Our findings indicate that the effectiveness of adversarial perturbations varies across different facial identities, suggesting variations in inherent robustness. We intend to investigate the factors contributing to this variability in future research.

Ablation Study on Perturbation Budget in Silencer-I.

To understand the influence of the perturbation budget on the effectiveness of Silencer-I, we conducted an ablation study on the CelebA-HQ dataset. Specifically, we investigated the performance of Silencer-I under constrained l_{inf}

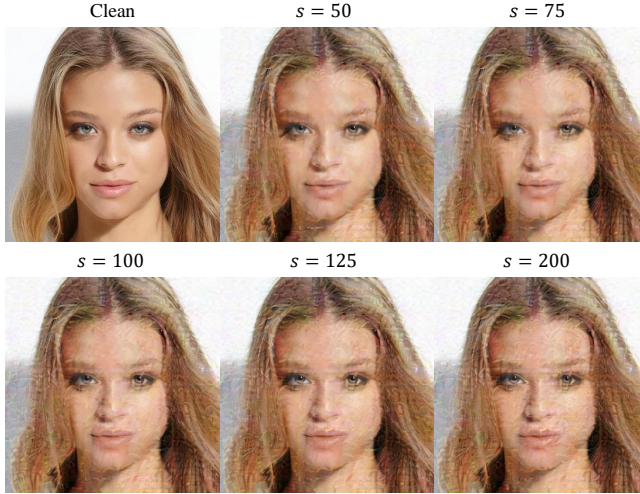


Figure 8. **Visualization Results with Different Iteration s .** The quality of the portrait decreases with the growth of s .

perturbation budgets. The l_{inf} limits the maximum change allowed for any single pixel value in the input image. A smaller budget implies a more subtle, less perceptible adversarial perturbation. As shown in Table 10, we evaluated Silencer-I with two different l_{inf} budget: 8/255 and 16/255. The results demonstrate that decreasing the perturbation budget leads to a reduction in Silencer-I’s performance. This is because a smaller budget restricts the degree to which Silencer-I can modify the input image to disrupt synchronization. However, even with a stricter budget, Silencer-I still achieves a notable level of protection performance compared with existing methods in Table 1. This suggests that Silencer-I is more effective, achieving considerable protection with fewer changes to the input portrait.

Ablation Study on Inverted Timesteps in Silencer-II.

We conducted an ablation study on the inverted latent space timesteps used in Silencer-II. Due to memory constraints, we investigated the impact of optimizing the latent feature for the final timestep versus optimizing for the final two timesteps specifically in the context of DDIM inversion. As shown in Table 11, optimizing the latent feature at only the final timestep yielded superior performance while consuming fewer resources compared to optimizing the last two steps. Consequently, we opted for the single-timestep optimization strategy. Further exploration is needed to improve the efficiency and effectiveness of latent feature optimization, addressing potential vulnerabilities to purification methods.

Ablation Study on \mathcal{L}_T in Silencer-II. We perform an ablation study to evaluate the effectiveness of \mathcal{L}_T in optimizing the inverted latent representation. As shown in Fig. 7,

while the nullifying loss \mathcal{L}_N still produces disturbed results, it achieves this by distorting the portrait, compromising the output’s quality and identification. It is mainly because the talking-head model fails to operate effectively when it cannot detect a face, rendering it unable to function as intended. This highlights the necessity of exploring optimized solutions that protect privacy without sacrificing visual integrity. With the assistance of \mathcal{L}_T , we can effectively reduce noise in the facial region while achieving our intended objectives. This approach strikes a balance between minimizing distortions and achieving the desired outcomes, enhancing the overall effectiveness of Silencer.

Ablation Study on the Initial Iteration s without Mask in Silencer-II.

To prevent facial blurring, we incorporate a face mask during the training process of Silencer-II. We begin by training the entire image without a mask for s iterations. Subsequently, a face mask is applied to exclude the facial region from further optimization. To verify the effect of s , we conduct an ablation study on a subset of CelebA-HQ, as shown in Fig. 8 and Table 12. The results indicate that as the number of iterations s increases, face quality deteriorates while protection performance improves. Therefore, we set $s = 100$ in our main experiments as it offers a balanced trade-off between maintaining facial clarity and achieving effective protection.

A.5. Additional Visual Results

Additional qualitative comparisons are presented in Fig. 9 and Fig. 10. These figures illustrate that our Silencer consistently achieves superior protection performance across various datasets. These video results can be found in our supplementary video.

References

- [1] Jiahao Cui, Hui Li, Yao Yao, Hao Zhu, Hanlin Shang, Kaihui Cheng, Hang Zhou, Siyu Zhu, and Jingdong Wang. Hallo2: Long-duration and high-resolution audio-driven portrait image animation. *arXiv preprint arXiv:2410.07718*, 2024. 1
- [2] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018. 4
- [3] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022. 1
- [4] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *CVPR*, 2021. 4
- [5] Zhengyue Zhao, Jinhao Duan, Kaidi Xu, Chenan Wang, Rui Zhang, Zidong Du, Qi Guo, and Xing Hu. Can protective perturbation safeguard personal data from being exploited by stable diffusion? In *Proceedings of the IEEE/CVF Conference*

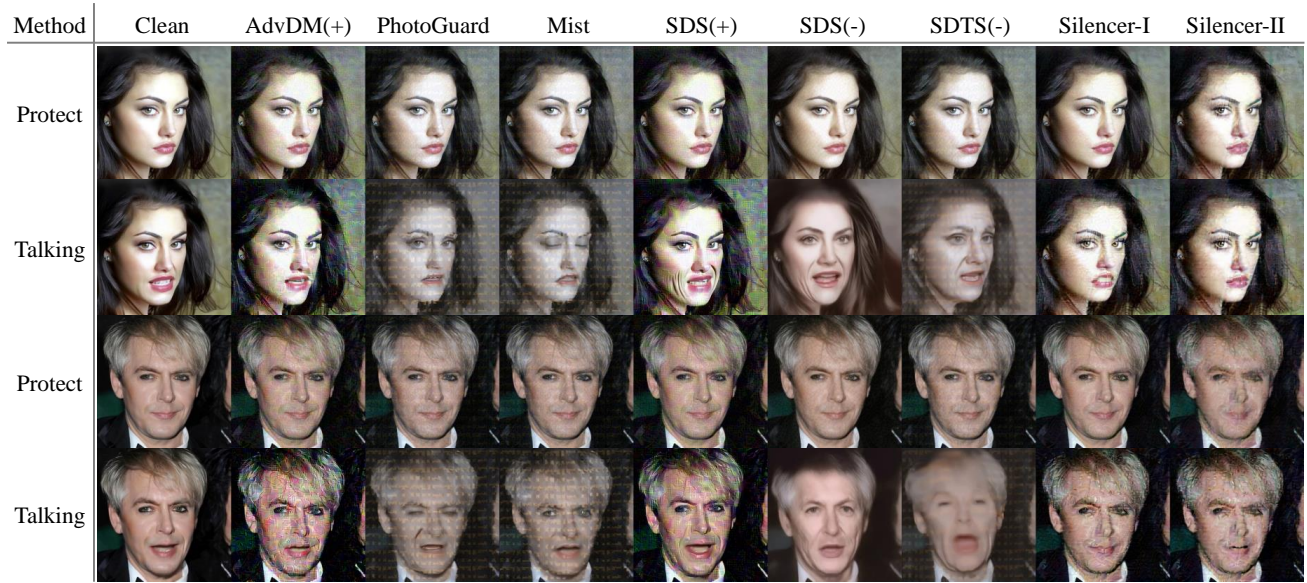


Figure 9. Additional Visualization Comparison with Image Protection Methods in CelebA-HQ [2].

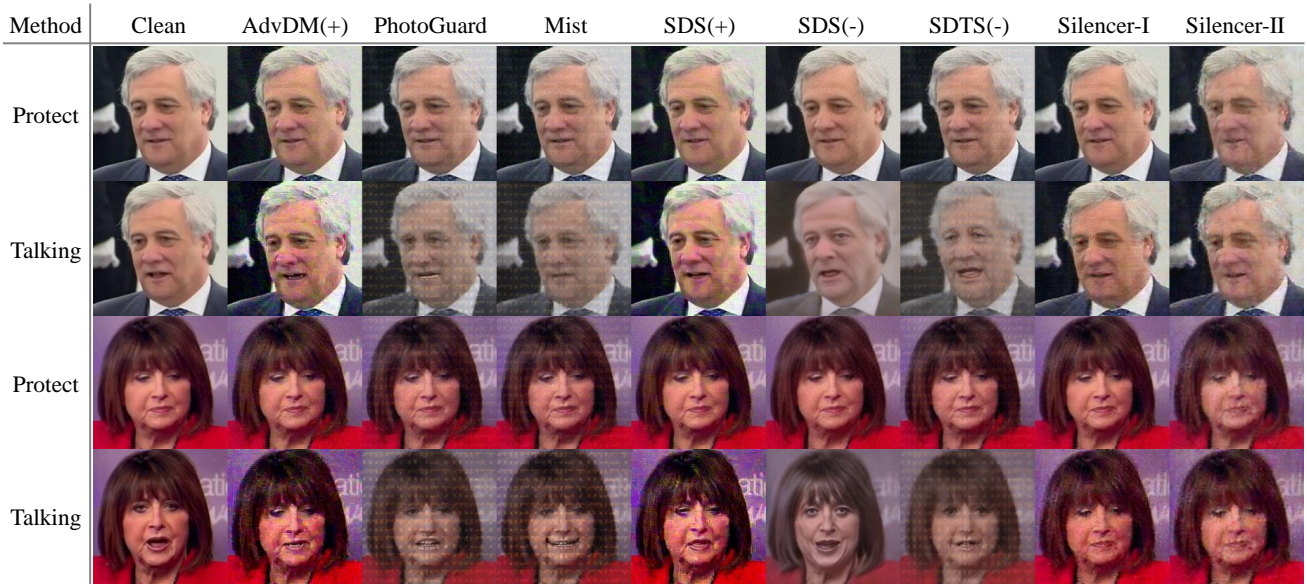


Figure 10. Additional Visualization Comparison with Image Protection Methods in TalkingHead-1KH [4].

on *Computer Vision and Pattern Recognition*, pages 24398–24407, 2024. [2](#)

- [6] Zhiqian Chen Yuming Li Chenguang Ma Zhiyuan Chen, Jia-jiong Cao. Echomimic: Lifelike audio-driven portrait animations through editable landmark conditioning, 2024. [1](#)