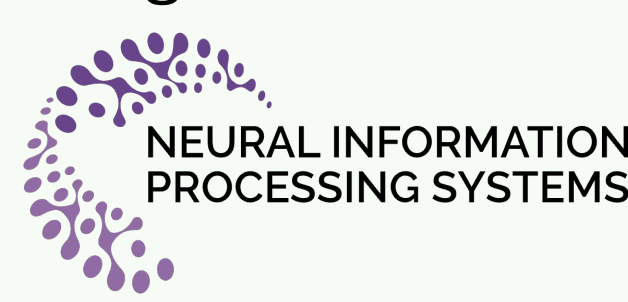


Introduction

- Watermarks ensure authenticity of AI generated content.
- But current methods are vulnerable to removal attacks
- NeurIPS 2024 WAVES challenge stress-tests their robustness under attacks.

Challenge Setup

- Beige-box track: Attacker knows watermarking methods.
- Black-box track: Attacker knows nothing about the watermark.



Methodology

For each track we tailored a suitable attack, both achieve near-perfect watermark removal (95.7%).

Beige-box Attack

StegaStamp Watermark

- Dataset Generation:**
 - Generate images from stable Diffusion 2-1.
 - Embed watermark message and its inverse via a StegaStamp

- Variational Autoencoder (VAE) based Removal:**

- A finetuned VAE minimizes the MSE loss between pairs.

$$L(\theta, \phi) = ||D_{\phi}(E_{\theta}(x_w)) - x_i||_2^2$$

- Post-processing:**

- Test-time VAE optimization.

$$L(\theta, \phi) = L_{mse} + L_{lips} + 0.5(1 - ssim(D_{\phi}(E_{\theta}(x_r)), x_w))$$

- Color and contrast transfer in CIELAB space.

TreeRing Watermark

- Simple spatial translation in the image domain.** $x_{shifted} = T(x_w, \Delta x)$
- Restore the leftmost Δx pixels.** $x_{final}(i, j) = \begin{cases} x_w(i, j) & \text{if } j < \Delta x \\ x_{shifted}(i, j) & \text{otherwise} \end{cases}$

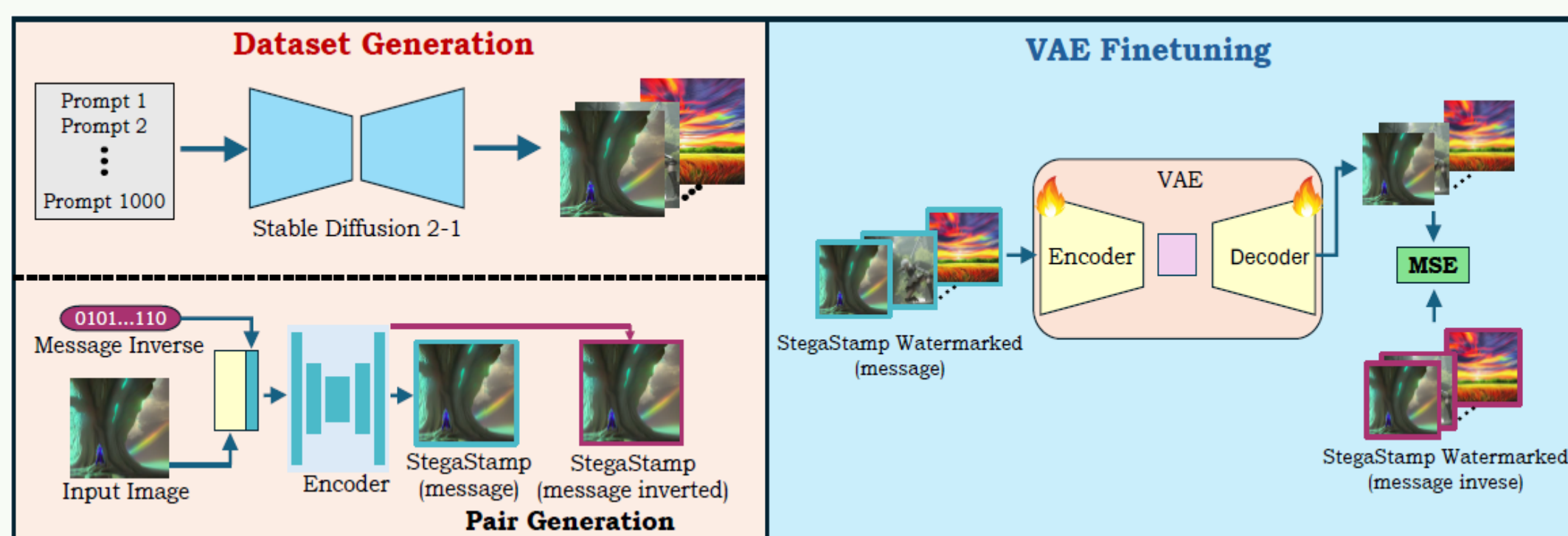


Figure1: Overview of our dataset generation and VAE-based watermark removal pipeline for StegaStamp.

Method	Detection	Quality	Overall
VAE Finetune	0.023	0.192	0.193
+ Test-Time Optimization	0.033	0.161	0.165
+Color/Contrast	0.037	0.153	0.157

Table1: Performance comparison of our watermark removal pipeline.

Black-box Attack

- Image-to-image Diffusion Model**

Applied stable diffusion Refiner with semantic prior to erase watermark without harming content.

- Forward diffusion (add noise)
- Reverse diffusion (remove noise to reconstruct clean image)

- Clustered the target images into groups**

Developed adaptive attacks for each cluster.

- Cluster 1: image-to-image diffusion with high strength parameter
- Cluster 2,3: three-stage pipeline like in Fig. 1
- Cluster 4: hybrid approach (image-to-image diffusion with lower strength, 7-pixel translation and boundary pixel restoration).

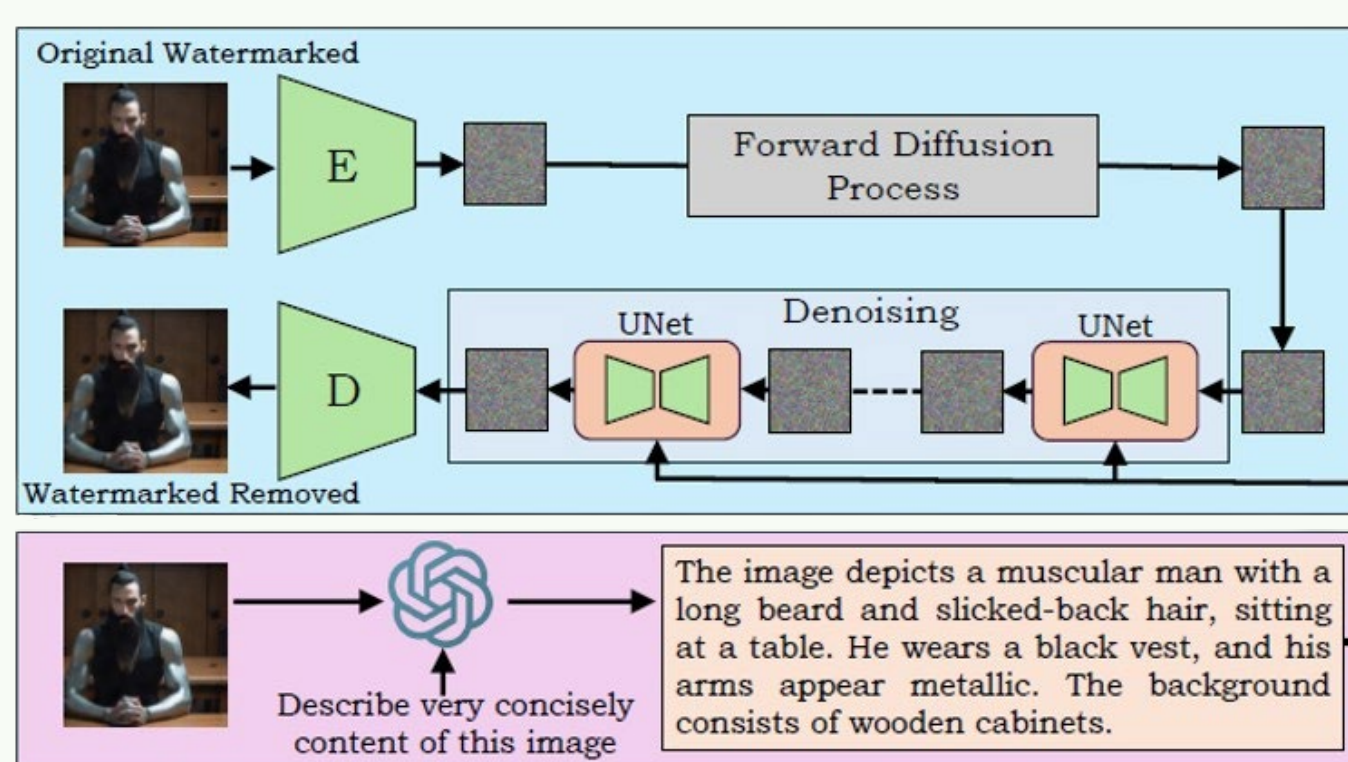
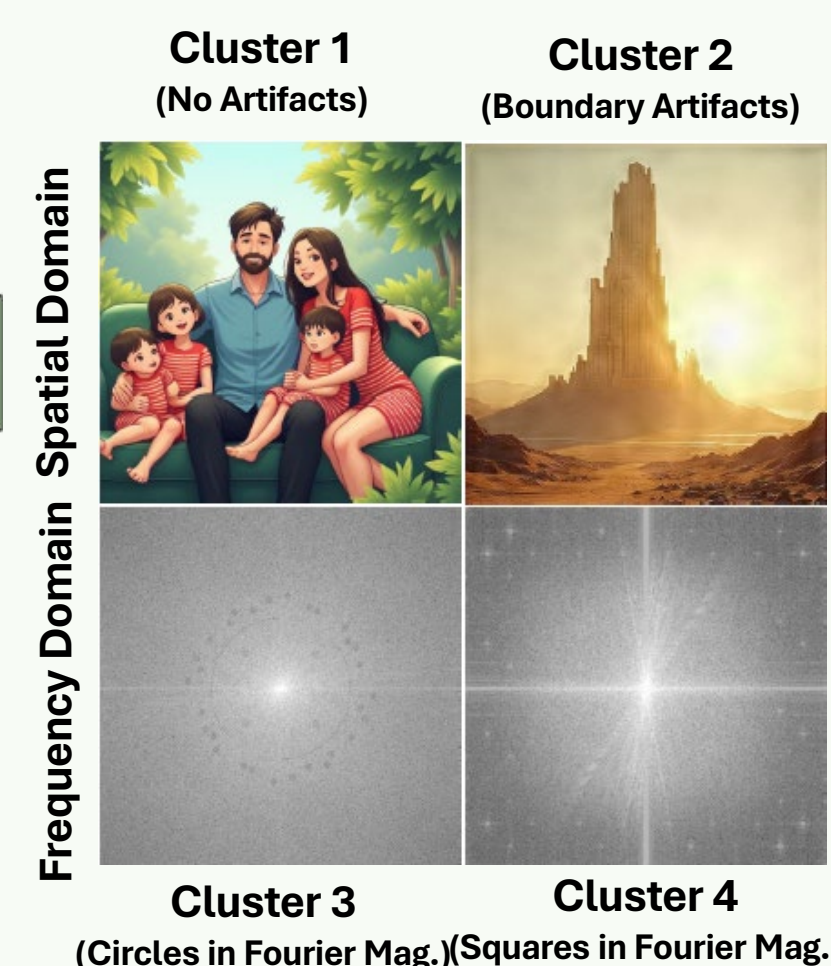


Figure2: Proposed image-to-image diffusion-based pipeline for watermark removal.



Results

Beige-box Track				
Rank	Participant	Detection	Quality	Total
1	Ours	0.037	0.153	0.157
2	Team-Ask	0.050	0.176	0.183
3	Team-Jafari	0.127	0.222	0.256
Beige-box Track				
Rank	Participant	Detection	Quality	Total
1	Ours	0.043	0.136	0.143
2	Team-Jafari	0.063	0.158	0.170
3	Team-Yepenh	0.087	0.177	0.197

Table2: Black-box and Beige-box tracks leaderboard



Figure3: Top row: Original watermarked images. Bottom row: Images after our attack.

Conclusion

- Our attack exposed vulnerabilities in current watermarking method.
- Our findings highlight the urgent need for stronger watermarking defenses.

