Adapter Shield: A Unified Framework with Built-in Authentication for Preventing Unauthorized Zero-Shot Image-to-Image Generation

Supplementary Material

1. More Implementation Details

In our experiments, the budget ϵ is set to $11/255$ for the facial identity protection task and $21/255$ for the artwork anti-plagiarism task across all compared methods, respectively. The budget for the artwork anti-plagiarism task is larger than that for the facial identity protection task, as the embedding dimensions used in the former (1024 and 1280) exceed those in the latter (512). A larger embedding dimension increases the difficulty of optimizing similarity for protected images, necessitating a larger perturbation budget. Likewise, a lower similarity threshold is adopted for the artwork anti-plagiarism task. It is worth noting that due to the threshold constraint, the set budgets are redundant for most of the samples that need protection. This setting is to balance the protection effect and visual fidelity.

2. Supplementary Results

Table 1. Experiments on DiT-based models.

		SD-3.5-IP-Adapter			Flux-IP-Adapter		
	PSNR↑	ESM↓	Enc.div↓	Dec↑	ESM↓	Enc.div↓	Dec↑
Ours	31.47	0.1079	0.0890	0.9308	0.1648	0.0681	0.9375

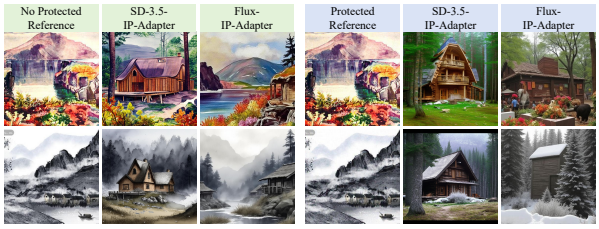

Figure 1. The visualization results on DiT-based models.

2.1. More Results of DiT-based Models

To demonstrate the generalization ability to DiT-based models [1], we evaluate our method on two DiT-based zero-shot image generation models: SD-3.5-Large-IP-Adapter[1] and Flux-IP-Adapter[2]. Both models employ SigLip [2] as the vision encoder, sharing an embedding dimension of 1152 but selecting different hidden states. All experimental configurations are kept identical to those reported in the main text. Quantitative results and visual results are summarized in Table 1 and Figure 1, respectively. In Table 1, ESM denotes the cosine similarity of embeddings between protected and original images, Enc.div is the similarity measuring encryption diversity, and Dec represents the decryption similarity.

[1]https://huggingface.co/InstantX/SD3.5-Large-IP-Adapter
[2]https://huggingface.co/XLabs-AI/flux-ip-adapter

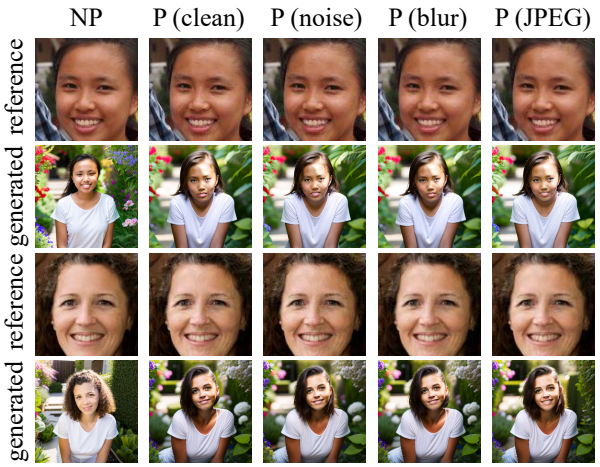

Figure 2. More visualization results of robustness evaluation. "NP" represents the original images without protection.

2.2. More Visualization Results of Robustness

We present more visualization results of robustness evaluation in Figure 2.

2.3. More Visualization Results of Ablation Study

We present more visualization results of ablation study in Figure 3 and Figure 4. Due to space constraints in the main text, Figure 3 presents a high-resolution version of Figure 7 in the main text, along with a detailed caption for enhanced clarity and reference. In Figure 3, the results of the same item across different images are obtained by the same passwords. For instance, "enc1" across "img1" and "img2" are encrypted by the same passwords. "enc1" and "enc2" denote the encrypted results using two distinct passwords. "dec1" and "dec2" denote the decrypted results using two distinct and random passwords. As shown in Figure 3 (b), the loss function $\mathcal{L}_{div}$ enhances the diversity of encrypted and decrypted results using different passwords ("enc1" vs "enc2", "dec1" vs "dec2"). The loss function $\mathcal{L}_{div\_s}$ improves the diversity of encrypted and decrypted results using the same password for different images ("enc1"/"enc2"/"dec1"/"dec2" across "img1" and

Figure 3. More visualization results of ablation study on loss functions of encryption and decryption. The results of the same item across different images are obtained by the same passwords. For instance, "enc1" across "img1" and "img2" are encrypted by the same passwords. "enc1" and "enc2" denote the encrypted results using two distinct passwords. "dec1" and "dec2" denote the decrypted results using two distinct and random passwords.

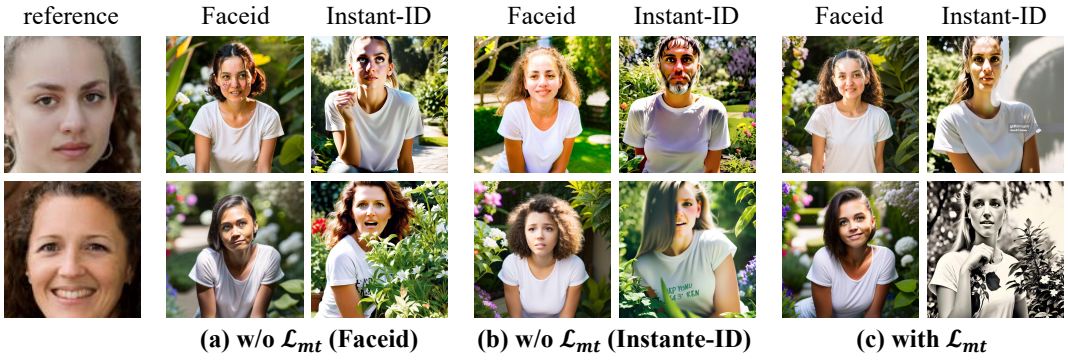

Figure 4. More visualization results of ablation study on multi-targeted loss function. (a) Minimizes cosine similarity only in the Faceid embedding domain. (b) Minimizes cosine similarity only in the Instant-ID embedding domain. (c) Minimizes cosine similarity in both Faceid and Instant-ID embedding domains.

"img2"). As shown in Figure 4, the loss function $\mathcal{L}_{mt}$ enhances the universality across various zero-shot image-to-image generation methods.

2.4. More Visualization Results of Encryption and Decryption

We present more visualization results of encryption and decryption in Figure 5 and Figure 6.

Figure 5. More visualization results of encryption and decryption performance. Text prompts: "*a young woman in white T-shirt in a garden*".

References

[1] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 1

[2] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1

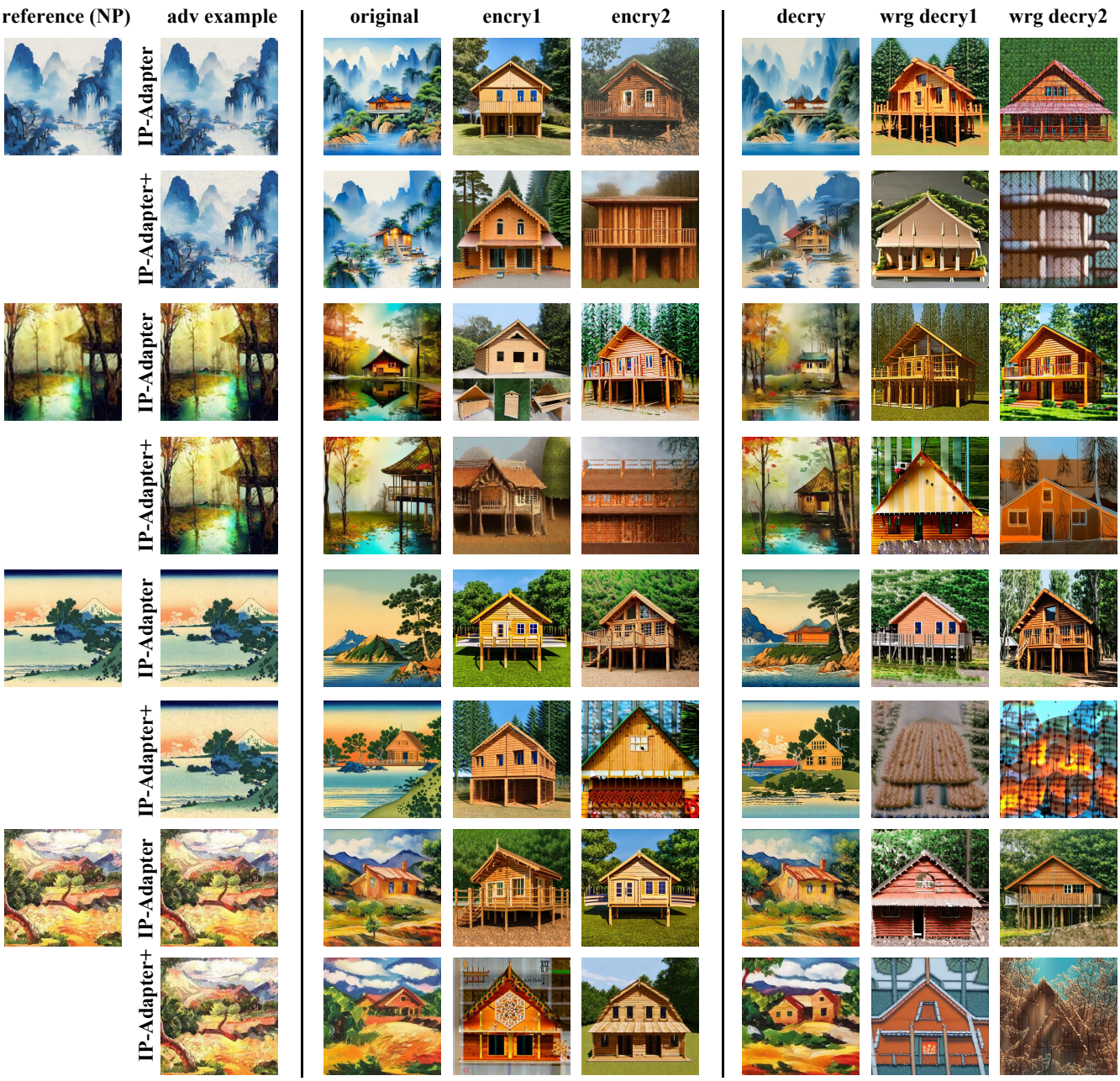

Figure 6. More visualization results of encryption and decryption performance. Text prompts: "*best quality, high quality, a wooden house in forest*".