DIA: The Adversarial Exposure of Deterministic Inversion in Diffusion Models

Supplementary Material

A. Disrupting Deterministic Inversion with Differentiable Trajectory: DIA-MT

We aimed to attack the inversion process by maximizing the process trajectory (PT), which is derived from the difference between the initial point x_0 and the final point x_T. However, it is also valid to target only the model's predicted trajectory (MT). This MT is derived as shown in Equation 7, and can be understood as exclusively capturing the contribution of the model's predictions to the inversion process. We propose an attack on this MT, called DIA-MT, which maximizes the residual image signal, defined as (x_T − decayed x_0), away from an isotropic Gaussian. DIA-MT is formulated as follows:

$$\delta_{\text{DIA-MT}} = \underset{||\delta|| \leq \epsilon}{\arg\max} \|\hat{x}_T(x_0 + \delta) - \sqrt{\bar{\alpha}_T}(\mathcal{E}(x_0 + \delta))\|_2^2 \tag{1}$$

Same as DIA-PT, $x_0 + \delta$ is detached from the computational graph used to calculate the gradient. Additionally, as an ablation study for DIA-MT, we compared its background preservation and prompt-image consistency with those of DIA-PT in Table 4. Here, the "Natural Edit" represents the natural outcome of an image editing process without any disruption, and it is used as a reference point in our experiments.

Inversion	**DDIM Inversion**				**Null-Text Inversion**		**Negative-Prompt Inversion**		**Direct Inversion**
Edit	**DDIM**	**MasaCtrl**	**PnP**	**P2P**	**P2P**	**Proximal-Guidance**	**P2P**	**Proximal-Guidance**	**P2P**
Natural Edit	25.7100	24.9504	26.1414	25.9123	25.5750	24.8495	25.4566	25.2090	25.8333
DIA-PT	**23.4614**	**18.3076**	**20.7749**	26.0381	**23.1999**	**20.0266**	**17.4938**	**17.3992**	26.0563
DIA-MT	23.7177	21.8592	23.4419	**25.7381**	24.4444	22.6471	21.4318	21.2247	**25.4861**

Table 3. CLIP similarity between the edited image and the prompt: Under a combination of different image inputs (clean or disrupted) and an inversion-editing method pairing, we show the CLIP similarity for images in the PIE-Bench dataset. Lower CLIP similarity indicates better immunization.

Metrics	**Structure**	**Background Preservation**			
Method	**Distance ↑**	**PSNR ↓**	**LPIPS ↑**	**MSE ↑**	**SSIM ↓**
Natural Edit	0.0249	24.3767	0.0914	0.0071	0.8124
DIA-PT	**0.1059**	**18.2202**	**0.3410**	**0.0237**	**0.5653**
DIA-MT	0.0514	22.0443	0.2447	0.0107	0.6856

Table 4. Average background and structure preservation metric for 9 editing techniques. This metric assesses how well the unedited regions are preserved.

According to Table 3 and Table 4, DIA-MT showed that attacking model trajectories are indeed effective, supporting our main argument that trajectories should be taken into account during attacks. However, it is observed that the performance of DIA-MT, which excludes the scaling of x_0 during the inversion process, is slightly weaker compared to DIA-PT, which includes it. This suggests that considering the scaling of x_0 leads to a more effective attack.

B. More experimental results

B.1. Step Generalizability

The number of steps used in DDIM varies according to the user's preference and budget. However, in DIA-PT and DIA-R, we execute the attack with trajectories sampled with 10 DDIM steps during the inversion and reconstruction process. Therefore, it is important to verify that our method works in editing environments using different timestep spacings. In Table 5, we compared the performance by setting DDIM steps to 20, 50, 200, and 1000 in an editing environment DDIM-to-DDIM for 140 randomly selected images from PIE-Bench [11].

Inference Step	Method	CLIP↓	Distance ↑	PSNR ↓	LPIPS ↑	MSE ↑	SSIM ↓
20	Natural Edit	25.1773	0.02102	25.6405	0.07350	0.00460	0.83539
	Photoguard	22.2119	0.08438	20.0121	0.26683	0.01327	0.65680
	Glaze	24.5483	0.04575	21.8093	0.20853	0.01063	0.67463
	AdvDM	23.1290	0.08768	20.4630	0.27519	0.01470	0.60280
	SDS	22.9042	0.06344	21.0721	0.26125	0.01203	0.62612
	DIA-PT (ours)	**20.2686**	**0.13860**	**17.1762**	**0.38523**	0.02803	**0.51990**
	DIA-R (ours)	21.2578	0.11058	17.5010	0.28155	**0.04004**	0.61981
50	Natural Edit	25.5855	0.02488	24.7736	0.08818	0.00566	0.82333
	Photoguard	23.7524	0.08060	19.4990	0.24705	0.01432	0.68406
	Glaze	25.3976	0.04158	22.1291	0.18450	0.00979	0.71290
	AdvDM	24.4719	0.07046	21.0579	0.23121	0.01242	0.66378
	SDS	24.0583	0.06431	21.1631	0.22512	0.01174	0.66120
	DIA-PT (ours)	23.0969	**0.10368**	**19.0090**	**0.31325**	0.02019	**0.59189**
	DIA-R (ours)	**22.9501**	0.08432	19.2288	0.22481	**0.02689**	0.68966
1000	Natural Edit	25.7686	0.02837	23.9500	0.10169	0.00668	0.81311
	Photoguard	24.4224	**0.07838**	**19.5179**	0.24314	0.01427	0.70117
	Glaze	25.8500	0.03955	21.9824	0.16913	0.00955	0.73640
	AdvDM	25.2232	0.06341	21.2205	0.20878	0.01198	0.69459
	SDS	24.6955	0.06262	21.2171	0.20686	0.01153	0.69147
	DIA-PT (ours)	**24.2379**	0.07757	20.1859	**0.25300**	0.01587	**0.65586**
	DIA-R (ours)	24.2615	0.06724	19.8701	0.19693	**0.01990**	0.72183

Table 5. Attack Performance Across Different Editing Steps. This table shows the performance of various attack methods using 20, 50, and 1000 DDIM steps for inversion and reconstruction on 140 images from the PIE-Bench dataset. Key metrics include CLIP Similarity (CLIP), Distance, PSNR, LPIPS, MSE, and SSIM.

A notable observation from these results is that the attack performance decreases as the number of steps increases, which is evident across all metrics. Our experiments include assessments with 1000 steps, the maximum step size typically used in the diffusion process, where we observe the poorest attack performance. However, it is crucial to note that the performance remains consistently lower than that of natural edits performed without any attack. This demonstrates the efficacy of our method across all step sizes and supports the stability of our approach.

B.2. Comparing Performance Through Noise Purification

In this section, we compare the robustness of different methods against cleaning approaches known as 'purification' for adversarial noise. We provide performance measurements after applying JPEG Compression, Crop & Resize, and AdverseCleaner to 700 immunized images across all methods. Details for each purification method are as follows:

- JPEG Compression: The simplest and fastest image compression algorithm for purifying adversarial noise. Compression quality can be selected between 0 and 100, where lower values cause more image degradation. We provide results with quality values of 70, 80, and 90.
- Crop & Resize: A naturally occurring and effective purification technique. We cropped 10% of each image and then resized it to match the model's input requirements.
- Adverse Cleaner [33]: An algorithmic approach capable of purifying high-frequency noise patterns.
- Gaussian Noising: The purification method that adds random Gaussian noise on immunized images. We provide results with σ=0.1.
- Noisy Upscaling [9]: A two-stage purification method proposed by Shan et al. [23], which applies Gaussian Noising (σ=0.1) followed by Stable Diffusion Upscaler [20].

As shown in Table 6, all baselines demonstrate robustness to purification when compared to Natural Edit. Notably, our method maintains superior performance while remaining robust to most purification methods. In some experiments, SDS shows sub-optimal performance, which appears to be due to its low-frequency pattern and higher degradation scale.

Purification Method	Attack Method	CLIP↓	Distance ↑	PSNR ↓	LPIPS ↑	MSE ↑	SSIM ↓
-	Natural Edit	25.7100	0.02613	23.8400	0.09933	0.00639	0.80723
JPEG Compression (90)	Photoguard	25.6936	0.05174	21.4936	0.22142	0.00962	0.70957
	Glaze	25.8862	0.03472	22.4310	0.15541	0.00829	0.74462
	AdvDM	24.5583	0.07191	21.0165	0.21653	0.01233	0.67113
	SDS	24.1685	0.06742	20.8744	0.21799	0.01181	0.67212
	DIA-PT (Ours)	24.2789	0.07655	20.0105	**0.27472**	0.01610	**0.63854**
	DIA-R (Ours)	**23.7255**	**0.08374**	**19.2542**	0.21633	**0.02637**	0.67706
JPEG Compression (80)	Photoguard	26.0247	0.04463	22.0655	0.18531	0.00880	0.73312
	Glaze	26.0196	0.03095	23.0004	0.13806	0.00741	0.76862
	AdvDM	24.4738	0.07044	21.1526	0.21349	0.01209	**0.67737**
	SDS	24.1725	0.06773	21.0680	0.21292	0.01159	0.67927
	DIA-PT (Ours)	24.8818	0.05645	20.9928	**0.23660**	0.01259	0.68420
	DIA-R (Ours)	**24.2000**	**0.07318**	**19.9076**	0.20011	**0.02179**	0.69623
JPEG Compression (70)	Photoguard	26.0953	0.04212	22.3060	0.16901	0.00836	0.74815
	Glaze	26.0306	0.03010	23.1220	0.13043	0.00712	0.77931
	AdvDM	24.7252	**0.06771**	21.3877	0.20781	0.01172	0.68590
	SDS	**24.2350**	0.06743	21.1680	**0.21060**	0.01171	**0.68238**
	DIA-PT (Ours)	25.5055	0.04608	21.6928	0.20653	0.01036	0.71623
	DIA-R (Ours)	25.0112	0.06244	**20.6276**	0.18408	**0.01673**	0.71640
Crop & Resize	Photoguard	25.7733	0.08424	17.3266	0.25859	0.02362	0.61375
	Glaze	25.8354	0.06285	17.3832	0.23109	0.02411	0.61677
	AdvDM	25.1026	0.07810	17.1060	0.27175	0.02563	**0.57034**
	SDS	**24.5399**	0.07795	16.9971	0.26720	0.02630	0.57432
	DIA-PT (Ours)	24.8340	0.07498	16.9972	**0.29035**	0.02618	0.57572
	DIA-R (Ours)	24.8310	**0.08518**	**16.5469**	0.25361	**0.03056**	0.59598
Adverse Cleaner	Photoguard	25.3614	0.06022	21.7390	**0.19018**	0.00939	0.75646
	Glaze	25.7053	0.03406	22.8134	0.14303	0.00768	0.78250
	AdvDM	24.6748	0.04763	22.4885	0.16196	0.00926	0.75834
	SDS	**24.1779**	0.05513	22.1588	0.16882	0.01001	0.74709
	DIA-PT (Ours)	25.3572	0.03936	21.9392	0.18543	0.00971	0.75839
	DIA-R (Ours)	24.6166	**0.06166**	**20.6104**	0.18917	**0.01714**	**0.73489**
Gaussian Noising	Photoguard	26.1351	0.0426	21.5181	**0.2908**	0.0094	0.5805
	Glaze	26.1265	0.0364	22.1301	0.2591	0.0084	0.6048
	AdvDM	25.5851	**0.0528**	21.5729	0.2772	0.0104	0.5920
	SDS	**25.3543**	0.0515	21.7844	0.2720	0.0101	0.6031
	DIA-PT (Ours)	26.2993	0.0423	21.4033	0.2809	0.0099	**0.5788**
	DIA-R (Ours)	25.9461	0.0449	**21.1744**	0.2750	**0.0115**	0.5918
Noisy Upscaling	Photoguard	25.4812	**0.0381**	23.0208	0.1561	0.0076	0.7654
	Glaze	**25.4772**	0.0351	22.9662	0.1506	0.0077	0.7658
	AdvDM	25.5246	0.0344	22.8962	0.1582	0.0076	0.7553
	SDS	25.5639	0.0355	22.8581	**0.1609**	**0.0079**	**0.7516**
	DIA-PT (Ours)	25.5119	0.0374	**22.8195**	0.1568	0.0078	0.7615
	DIA-R (Ours)	25.6470	0.0351	22.8716	0.1511	**0.0079**	0.7662

Table 6. Immunization performance across purification methods. This table demonstrates the robustness of various immunization methods against JPEG Compression, Crop & Resize, and Adverse Cleaner attacks, evaluated on 700 images from the PIE-Bench dataset.

B.3. Considerations for Selecting Hyperparameters

We provide an analysis of the hyperparameters of DIA-PT and DIA-R: attack iteration and trajectory length. Attack iteration is the number of PGD updates needed for optimization, while trajectory length is the length of the differentiable trajectory used in DDIM inversion and sampling during a single update.

Through Table 7, we noted that both DIA-PT and DIA-R converge in disruption performance with just 20 attack iterations, which is likely because we precisely target the chained trajectory. Additionally, Table 8 reveals a difference between DIA-PT and DIA-R, with their best values found at trajectory lengths of 10 and 20, respectively. This indicates that for DIA-PT, trajectories beyond a certain length may have a negative impact since its loss is calculated based on the latent code z_0. Instead, DIA-R's performance improves with more detailed trajectories as it computes loss through x_0. To ensure a consistent inversion trajectory environment across all our experiments, we set the trajectory length to 10.

Method	Attack Iteration	CLIP↓	Distance ↑	PSNR ↓	LPIPS ↑	MSE ↑	SSIM ↓
DIA-PT	5	25.6048	0.0482	21.5865	0.2094	0.0103	0.6992
	10	24.3525	0.0751	20.0086	0.2693	0.0155	0.6366
	15	23.7979	0.0913	19.2879	0.2949	0.0188	0.6078
	20	**23.4575**	**0.1006**	**18.7744**	**0.3124**	**0.0208**	**0.5874**
DIA-R	5	24.6790	0.0547	20.8372	0.1791	0.0133	0.7274
	10	24.3205	0.0670	19.9336	0.2038	0.0186	0.6967
	15	23.8511	0.0796	19.3068	0.2190	0.0239	0.6818
	20	**23.4670**	**0.0882**	**18.7633**	**0.2307**	**0.0288**	**0.6666**

Table 7. Attack Performance Across Different Attack Iterations. This table shows the performance of DIA-PT and DIA-R attacks using 5, 10, 15, and 20 attack iterations on the PIE-Bench dataset. The bold values represent the best performance across different attack iterations for each method.

Method	Trajectory Length	CLIP↓	Distance ↑	PSNR ↓	LPIPS ↑	MSE ↑	SSIM ↓
DIA-PT	5	25.6181	0.0506	21.1142	0.2163	0.0107	0.6967
	10	**23.4575**	**0.1006**	**18.7744**	**0.3124**	**0.0208**	**0.5874**
	20	24.1361	0.0782	20.0423	0.2835	0.0154	0.6209
DIA-R	5	24.3258	0.0676	19.7006	0.2118	0.0179	0.6899
	10	23.4670	0.0882	18.7633	0.2307	0.0288	0.6666
	20	**22.0941**	**0.1101**	**17.5972**	**0.2540**	**0.0432**	**0.6451**

Table 8. Attack Performance Across Different Trajectory Steps. This table shows the performance of DIA-R and DIA-PT attacks using 5, 10, and 20 trajectory steps on the PIE-Bench dataset. The bold values represent the best performance across different trajectory lengths for each method.

C. Transferability to Black-Box Models

The Diffusion model is constantly updated and has an active developer community, resulting in many variants. As a result, the model used for attacks and the model used for editing the disrupted images may differ, potentially leading to attack performance degradation. This concept is referred to as model transferability, which indicates how well the disrupting performance is maintained across different scenarios. We conducted an experiment to test whether images disrupted using the initial stable diffusion model, version 1.4 (SD v1.4), retain their resistance when edited with black-box models, specifically stable diffusion versions 2.0 (SD v2.0) and 2.1 (SD v2.1). The experiment utilized a simple editing method DDIM-to-DDIM, and the hyperparameters used for editing were identical to those employed with SD v1.4, with the experiment conducted on the PIE-Bench.

Diffusion Ver.	Method	CLIP↓	Distance ↑	PSNR ↓	LPIPS ↑	MSE ↑	SSIM ↓
SD v1.4	Natural Edit	25.7100	0.02613	23.8400	0.09933	0.00639	0.80723
	DIA-PT (Ours)	**23.4613**	**0.10042**	18.7803	**0.31218**	0.02074	**0.58770**
	DIA-R (Ours)	23.4626	0.08821	**18.7655**	0.23087	**0.02877**	0.66656
SD v2.0	Natural Edit	25.7983	0.04129	23.2952	0.12510	0.00708	0.79470
	DIA-PT (Ours)	25.1747	0.05941	21.0996	0.22989	0.01152	0.70174
	DIA-R (Ours)	**24.1616**	**0.07028**	**20.1148**	**0.20655**	**0.01826**	**0.71447**
SD v2.1	Natural Edit	24.5758	0.05082	21.7974	0.16001	0.01027	0.76361
	DIA-PT (Ours)	23.4423	**0.08422**	19.2938	**0.26466**	0.01734	**0.65941**
	DIA-R (Ours)	**22.5146**	0.08270	**19.1470**	0.23906	**0.02164**	0.68034

Table 9. Disrupting Performance Comparison Across Stable Diffusion Model Versions. The table illustrates the robustness of images disrupted using the early version of Stable Diffusion (SD v1.4) when attempting editing attempts using different versions of the model (SD v2.0 and SD v2.1). The provided metrics (CLIP, Distance, PSNR, LPIPS, MSE, SSIM) evaluate various aspects of the edited images, showing that the immunized retain some immunity despite the difference in model versions. The arrow next to each metric name indicates the direction of better performance.

In Table 9, we observe that images immunized with the early version of Stable Diffusion (SD v1.4) retain a substantial disruptive signal when edited with different versions of Stable Diffusion. These results are crucial, as SD v2.1 and SD v2.0, along with the earlier SD v1.4, serve as the foundational models for most community-driven developments.

Interestingly, our experiments consistently show that the attack is less disruptive when using different versions of Stable Diffusion (SD v2.0 and SD v2.1), but it remains a consistent disruption. Additionally, the qualitative result in Fig. 1 enables visual understanding. Overall, the results support the generalizability of our approach, demonstrating that even with advancements in model versions, the disrupted images continue to exhibit strong resistance to editing attempts.

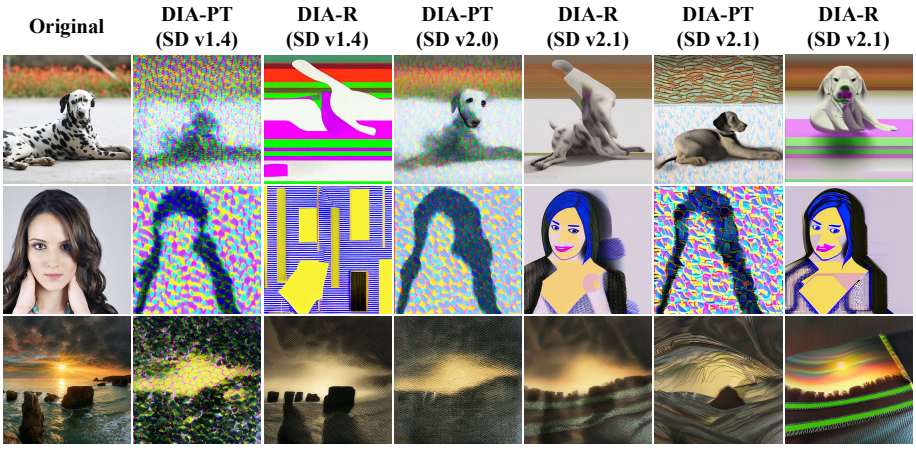

Figure 1. Quality comparison across Stable Diffusion Model Version. In this figure, DIA-PT and DIA-R visualize the results of editing images immunized in SD v1.4 across SD v1.4, SD v2.0, and SD v2.1. Editing in different versions reduces the disruptive performance, but still shows considerable effectiveness.

D. Observation on Over-Editing Scenarios

We extensively report over-edited images observed in DDIM-to-P2P and Direct-to-P2P. In Fig. 2, DDIM-to-P2P produces text-familiar images through P2P's aggressive attention map handling, which causes the failure to preserve the integrity of the original image during editing. In Fig. 3, Direct-to-P2P shows a similar performance to Natural Edit as it corrects the target diffusion trajectory.

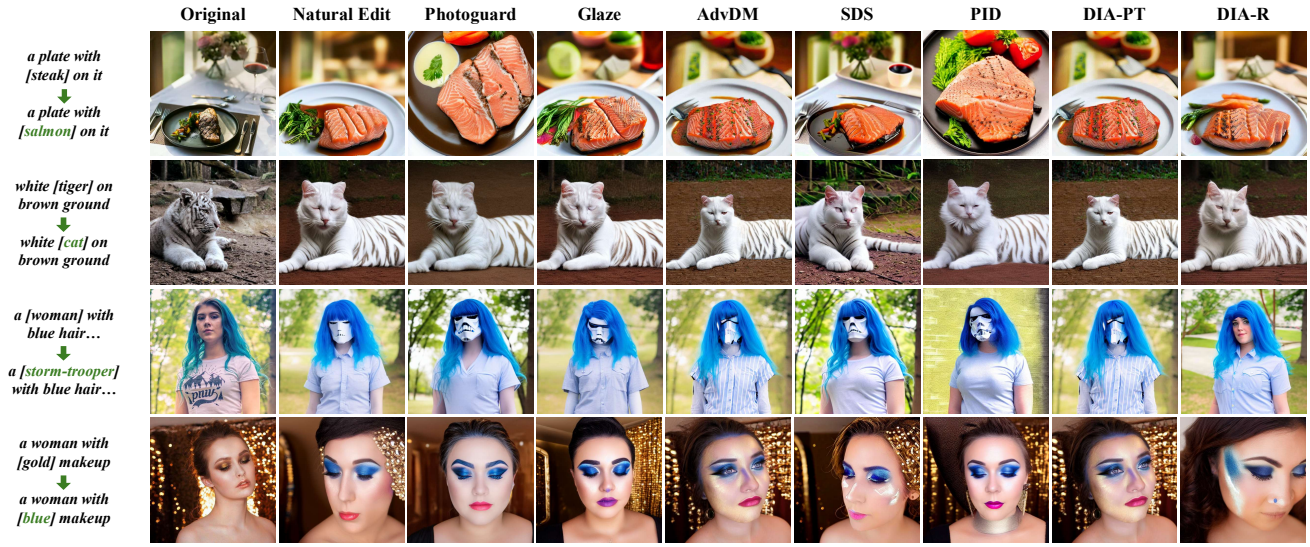

Figure 2. Quality comparison of images generated by DDIM-to-P2P across different immunization methods. The words in green indicate the parts to be edited from the original image. We visualize the failure to preserve the integrity of the original image.

Figure 3. Quality comparison of images generated by Direct-to-P2P across different immunization methods. The words in green indicate the parts to be edited from the original image. We visualize that Direct-to-P2P robustly edits against immunization methods.

E. Limitation

Our proposed DIA-PT takes approximately 40 seconds, while DIA-R takes around 1 minute and 50 seconds. Although the required VRAM of 6-7GB is not overly demanding, there is room for improvement. Additionally, our method focuses on current image inversion methods and prominent image generation models. Should future image inversion methods evolve with operations orthogonal to the current DDIM inversion process, or the image modeling paradigm is subjected to changes, our method may undergo performance decay. We believe that analyzing our approach to address these limitations will help guide future research on the problem of image editing immunity through disruption.