Table 5. Extended transferability results on normally trained models for the remaining surrogates (MobileNet-v2, ViT-B, Swin-B). Same setup as Table 2.

		Target models				Avg	FID
Surrogate	**Attack**	**Res-50**	**MN-v2**	**ViT-B**	**Swin-B**		
MN -v2	Boundary Attack	5.0	89.5	4.5	6.2	5.2	124.3
	HSJA	4.5	90.6	4.4	5.3	4.7	117.2
	ColorFool	30.9	95.2	15.8	10.1	18.9	70.1
	NCF	**70.4**	92.7	40.7	22.2	44.4	70.4
	ACA	62.0	93.8	48.8	48.5	53.1	**63.2**
	DiffAttack	69.0	96.3	76.6	56.2	**67.3**	64.2
	AdvDiffEdit (ours)	45.8	95.6	**81.0**	**57.4**	61.4	64.4
ViT -B	Boundary Attack	6.5	7.8	91.7	2.4	5.5	124.6
	HSJA	3.6	8.2	89.5	3.2	5.0	118.3
	ColorFool	35.7	45.4	83.8	10.1	30.4	69.2
	NCF	30.1	59.0	82.6	23.9	37.7	67.8
	ACA	60.1	64.3	87.9	**55.3**	**59.9**	65.3
	DiffAttack	**60.8**	42.3	88.0	25.2	42.8	**61.3**
	AdvDiffEdit (ours)	52.4	**67.0**	90.5	46.3	55.2	68.5
Swin -B	Boundary Attack	8.5	15.6	4.4	92.3	9.5	126.3
	HSJA	7.3	12.7	5.3	91.5	8.4	117.8
	ColorFool	40.6	46.5	32.1	67.1	51.4	70.0
	NCF	50.5	55.5	37.1	63.1	47.7	71.2
	ACA	45.2	60.6	53.1	88.7	53	66.4
	DiffAttack	56.7	56.6	58.4	90.1	57.2	**65.2**
	AdvDiffEdit (ours)	**58.2**	**67.1**	**67.5**	92.5	**64.3**	74.7

7. Proof of Lemma 1

proof 1 *Fix ε and let*

$$\psi(t) := m(P_t(z_t(\varepsilon))).$$

By assumption, ψ is A-Lipschitz in t. Hence for any $t' \in [t-\delta, t]$,

$$\psi(t') \;\geq\; \psi(t) - A|t - t'| \;\geq\; \psi(t) - A\delta.$$

Using the buffer condition at time t,

$$\psi(t) \;\geq\; \Delta(t) + \rho,$$

we obtain

$$\psi(t') \;\geq\; \Delta(t) + \rho - A\delta.$$

If $\delta \leq \rho/(2A)$, then $A\delta \leq \rho/2$, so

$$\psi(t') \;\geq\; \Delta(t) + \rho/2.$$

Since $\Delta(\cdot)$ is nonincreasing and $t' \leq t$, we have

$$\Delta(t') \leq \Delta(t).$$

Therefore,

$$\psi(t') \;\geq\; \Delta(t) + \rho/2 \;\geq\; \Delta(t') + \rho/2 \;\geq\; \Delta(t').$$

Recalling that $\psi(t') = m(P_{t'}(z_{t'}(\varepsilon)))$, this gives

$$m(P_{t'}(z_{t'}(\varepsilon))) \geq \Delta(t'),$$

as claimed. Hence the same ε remains a seed for every finer strength $t' \in [t-\delta, t]$ with $\delta \leq \rho/(2A)$.

Algorithm 2 GREEDY NOISE REFINEMENT

Require: Adversarial seed ε with $\mathcal{O}(I(\varepsilon)) = 1$; Oracle $\mathcal{O}$; Metric $d(\cdot)$; Budget B; Step δ

$D \leftarrow d(I(\varepsilon))$; $\quad q \leftarrow 0$
while $q < B$ **do**
  Sample direction $u \sim \mathcal{N}(0, I)$
  found $\leftarrow$ **false**
  for $s \in \{+1, -1\}$ **do**
    $\varepsilon_{\text{cand}} \leftarrow \varepsilon + s \cdot \delta \cdot u$
    $y \leftarrow \mathcal{O}(I(\varepsilon_{\text{cand}}))$; $\quad q \leftarrow q + 1$
    if $y = 1$ **and** $d(I(\varepsilon_{\text{cand}})) < D$ **then**
      $\varepsilon \leftarrow \varepsilon_{\text{cand}}$; $\quad D \leftarrow d(I(\varepsilon))$
      $\delta \leftarrow \delta \cdot \gamma_{\text{grow}}$
      found $\leftarrow$ **true**; $\quad$ **break**
    end if
  end for
  if not found then
    $\delta \leftarrow \max(\delta \cdot \gamma_{\text{shrink}}, \delta_{\min})$
  end if
end while
return ε

8. Refinement Algorithm

9. ℓ_2 Interpretation

Let an image of size $h \times w \times 3$ be represented in $[0, 255]$ per channel and let d denote the ℓ_2 distance measured on the

Figure 6. Additional examples of AdvDiffEdit. In each pair, the original image is shown on top and the adversarially edited image on the bottom. The examples illustrate that our method can flip decisions while preserving natural visual appearance.

rescaled $[0, 1]$ image. A convenient interpretation of d is an average per-pixel bit change:

$$\mathrm{bpp}(d) \approx \left\lceil \frac{d}{\sqrt{3hw}} \cdot 255 \right\rceil, \tag{6}$$

i.e., a perturbation of size d on $[0, 1]$ corresponds on average to $\mathrm{bpp}(d)$ integer levels per pixel on $[0, 255]$.

10. More Transferability

Table 5 reports extended transferability for the remaining surrogates (MobileNet-v2, ViT-B, Swin-B). The same trends as in the main table hold: Boundary Attack and HSJA achieve very low off-diagonal transfer and high FID, while AdvDiffEdit consistently outperforms color-based black-box attacks and remains competitive with diffusion-based white-box attacks across all targets. For brevity, we only report FID here; LPIPS is shown in the main ResNet-50 table.

11. More Examples