A PROOF OF PROPOSITION 1

Proposition 1. *Let $d_{max}(\phi) := \max_{x \sim P_{\mathcal{D}}} |\phi^T x|$. If $\varepsilon \geq 1 + d_{max}(\phi)$, then $\Delta x = (1 + d_{max}(\phi))\phi$ solves Eq. (4), and $f_\phi(x + \Delta x) > 0 \; \forall \; x \sim G_0$.*

Proof. Let ϕ be a data-compliant key and let x be sampled from $P_{\mathcal{D}}$. First, from the KKT conditions for Eq. (4) we can show that the solution Δx^* is proportional to ϕ:

$$\Delta x^* = \phi / \mu^*, \tag{1}$$

where $\mu^* \geq 0$ is the Lagrange multiplier. To minimize the objective, we seek μ such that

$$1 - (x + \Delta x^*)^T \phi = 1 - x^T \phi - 1/\mu^* \leq 0, \tag{2}$$

for all x. Since $x^T \phi < 0$ (data compliance), this requires $1/\mu^* = 1 + d_{\max}(\phi)$. Therefore, when $\varepsilon \geq 1 + d_{\max}(\phi)$, $\Delta x^* = (1 + d_{\max}(\phi))\phi$ solves Eq. (4). And $f_\phi(x + \Delta x^*) = \phi^T(x + (1 + d_{\max}(\phi))\phi) = \phi^T x + 1 + d_{\max}(\phi) > 0$.

□

B PROOF OF THEOREM 1

Theorem 1. *Let $d_{max}(\phi) = \max_{x \in \mathcal{D}} |\phi^T x|$, $\sigma^2(\phi) = \phi^T \Sigma \phi$, $\delta \in [0, 1]$, and ϕ be a data-compliant key. $D(G_\phi) \geq 1 - \delta/2$ if*

$$\gamma \geq \sigma(\phi) \sqrt{\log\left(\frac{1}{\delta^2}\right)} + d_{max}(\phi) - \phi^T \mu. \tag{3}$$

Proof. We first note that due to data compliance of keys, $\mathbb{E}_{x \sim P_{\mathcal{D}}} \left[\mathbb{1}(\phi^T x < 0)\right] = 1$. Therefore $D(G_\phi) \geq 1 - \delta/2$ iff $\mathbb{E}_{x \sim P_{G_\phi}} \left[\mathbb{1}(\phi^T x > 0)\right] \geq 1 - \delta$, i.e., $\Pr(\phi^T x > 0) \geq 1 - \delta_d$ for $x \sim P_{G_\phi}$. We now seek a lower bound for $\Pr(\phi^T x > 0)$. To do so, let x and x_0 be sampled from P_{G_ϕ} and P_{G_0}, respectively. Then we have

$$\begin{aligned} \phi^T x &= \phi^T \left(x_0 + \gamma\phi + \epsilon\right) \\ &= \phi^T x_0 + \gamma + \phi^T \epsilon, \end{aligned} \tag{4}$$

and

$$\Pr(\phi^T x > 0) = \Pr\left(\phi^T \epsilon > -\phi^T x_0 - \gamma\right). \tag{5}$$

Since $d_{\max}(\phi) \geq -\phi^T x_0$, we have

$$\Pr(\phi^T x > 0) \geq \Pr\left(\phi^T \epsilon > d_{\max}(\phi) - \gamma\right) = \Pr\left(\phi^T(\epsilon - \mu) \leq \gamma - d_{\max}(\phi) + \phi^T \mu\right). \tag{6}$$

The latter sign switching in equation 6 is granted by the symmetry of the distribution of $\phi^T(\epsilon - \mu)$, which follows $\mathcal{N}(0, \phi^T \Sigma \phi)$. A sufficient condition for $\Pr(\phi^T x > 0) \geq 1 - \delta$ is then

$$\Pr\left(\phi^T(\epsilon - \mu) \leq \gamma - d_{\max}(\phi) + \phi^T \mu\right) \geq 1 - \delta. \tag{7}$$

Recall the following tail bound of $x \sim \mathcal{N}(0, \sigma^2)$ for $y \geq 0$:

$$\Pr(x \leq \sigma y) \geq 1 - \exp(-y^2/2). \tag{8}$$

Compare equation 8 with equation 7, the sufficient condition becomes

$$\gamma \geq \sigma(\phi) \sqrt{\log\left(\frac{1}{\delta^2}\right)} + d_{\max}(\phi) - \phi^T \mu. \tag{9}$$

□

C PROOF OF THEOREM 2

Theorem 2. *Let* $d_{min} = \min_{x\in\mathcal{D}} |\phi^T x|$*,* $d_{max} = \max_{x\in\mathcal{D}} |\phi^T x|$*,* $\sigma^2(\phi) = ||\phi||^2_\Sigma$*,* $\delta \in [0,1]$*.* $A(\mathcal{G}) \geq 1 - N\delta$ *if* $D(G) \geq 1-\delta$ *for all* $G_\phi \in \mathcal{G}$ *and for any pair of data-compliant keys* ϕ *and* ϕ'*:*

$$\phi^T \phi' \leq -1 + \frac{d_{max}(\phi') + d_{min}(\phi') - 2\phi'^T \mu}{\sigma(\phi')\sqrt{\log\left(\frac{1}{\delta^2}\right)} + d_{max}(\phi') - \phi'^T \mu}. \tag{10}$$

Proof. Let ϕ and ϕ' be any pair of keys. Let x and x_0 be sampled from P_{G_ϕ} and P_{G_0}, respectively. We first derive the sufficient conditions for $\Pr(\phi'^T x < 0) \geq 1-\delta$. Since $x = x_0 + \gamma\phi + \epsilon$ for $x \in G_\phi$, we have

$$\begin{aligned} \phi'^T x &= \phi'^T \left(x_0 + \gamma\phi + \epsilon\right) \\ &= \phi'^T x_0 + \gamma\phi^T\phi' + \phi'^T\epsilon. \end{aligned} \tag{11}$$

Then

$$\begin{aligned} \Pr(\phi'^T x < 0) &= \Pr\left(\phi'^T\epsilon < -\phi'^T x_0 - \gamma\phi^T\phi'\right) \\ &\geq \Pr\left(\phi'^T(\epsilon - \mu) < d_{\text{min}}(\phi') - \gamma\phi^T\phi' - \phi'^T\mu\right), \end{aligned} \tag{12}$$

where $d_{\text{min}}(\phi') := \min_{x\in\mathcal{D}} |\phi'^T x|$ and $\phi'^T(\epsilon-\mu) \sim \mathcal{N}(0, \sigma^2(\phi'))$. Using the same tail bound of normal distribution and Theorem 1, we have $\Pr(\phi^T x < 0) \geq 1-\delta$ if

$$\begin{aligned} & -\gamma\phi^T\phi' \geq \sigma(\phi')\sqrt{\log\left(\frac{1}{\delta^2}\right)} - d_{\text{min}}(\phi') + \phi'^T\mu \\ \Rightarrow \quad & \phi^T\phi' \leq -1 + \frac{d_{\text{max}}(\phi') + d_{\text{min}}(\phi') - 2\phi'^T\mu}{\sigma(\phi')\sqrt{\log\left(\frac{1}{\delta^2}\right)} + d_{\text{max}}(\phi') - \phi'^T\mu} \end{aligned} \tag{13}$$

Note that $\Pr(A=1, B=1) = 1 - \Pr(A=0) - \Pr(B=0) + \Pr(A=0, B=0) \geq 1 - \Pr(A=0) - \Pr(B=0)$ for binary random variables A and B. With this, it is straight forward to show that when $\Pr(\phi'^T x < 0) \geq 1-\delta$ for all $\phi' \neq \phi$, and $\Pr(\phi^T x > 0) \geq 1-\delta$ for all ϕ, then $\Pr(\phi^T x > 0, \phi'^T x < 0\,\forall\phi' \neq \phi) \geq 1 - N\delta$ and $A(\mathcal{G}) \geq 1 - N\delta$.

□

D TRAINING DETAILS

D.1 METHOD

We trained user-end models based on the objective function (Eq.(10) in the main text). For datasets where the root models follow DCGAN and PGAN, the user-end models follow the same architecture. For the FFHQ dataset where StyleGAN is used, we introduce an additional shallow convolutional network as a residual part, which is added to the original StyleGAN output to match with the perturbed datasets $\mathcal{D}_{\gamma,\phi}$. In this case, the training using Eq.(10) is limited to the additional shallow network, while the StyleGAN weights are frozen. More specifically, denoting the combination of convolution, ReLU, and max-pooling by Conv-ReLU-Max, the shallow network consists of three Conv-ReLU-Max blocks and one fully connected layer. All of the convolution layers have 4 x 4 kernels, stride 2, and padding 1. And all of the max-pooling layers have 3 x 3 kernels and stride 2.

D.2 PARAMETERS

We adopt the Adam optimizer for training. Training hyper-parameters are summarized in Table 1.

Table 1: Hyper-parameters to train keys (ϕ) and generators (G_ϕ).

GANs	Dataset	Batch Size	Learning Rate	β_1	β_2	Epochs
DCGAN	MNIST	16	0.001	0.9	0.99	10
DCGAN	CelebA	64	0.001	0.9	0.99	2
StyleGAN	FFHQ	8	0.001	0.9	0.99	5

D.3 TRAINING TIME

All experiments are conducted on V100 Tesla GPUs. Table 2 summarizes the number of GPUs used and the training time for the non-robust models (Eq.(10) in the main text) and robust models (Eq.(12) in the main text). Recall that we chose Eq.(10) for training the non-robust user-end models for consistency with the theorems, although Eq.(12) can be used to achieve attributability in practice, as is shown in the robust attribution study. Therefore, the non-robust training takes longer to due the iteration of γ in Alg. 1.

Table 2: Training time (in minute) of one key (Eq.(9) in main text) and one generator (Eq.(10) in main text). DCGAN_M: DCGAN for MNIST, DCGAN_C: DCGAN for CelebA.

GANs	GPUs	Key	Non-robust	Blurring	Cropping	Noise	JPEG	Combination
DCGAN_M	1	1.77	14	4.12	3.96	4.19	5.71	5.12
DCGAN_C	1	5.31	15	10.33	9.56	10.35	10.25	10.76
PGAN	2	50.89	141.07	140.05	131.90	133.46	132.46	135.07
CycleGAN	1	20.88	16.04	16.26	15.43	15.71	15.98	16.41
StyleGAN	1	54.23	3.12	-	-	-	-	-

E ABLATION STUDY

Here we conduct an ablation study on the hyper-parameter C for the robust training formulation (Eq.(12)). Training with larger C focuses more on generation quality, thus sacrificing distinguishability and attributability. These effects are reported in Table 3 and Table 4. Due to limited time, the results here are averaged over five models for each C and data-model pairs.

Table 3: Distinguishability (top), attributability (btm) before (Bfr) and after (Aft) robust training. DCGAN$_\text{M}$: DCGAN for MNIST, DCGAN$_\text{C}$: DCGAN for CelebA.

Model	C	`Blurring`		`Cropping`		`Noise`		`JPEG`		`Combination`	
-	-	Bfr	Aft	Bfr	Aft	Bfr	Aft	Bfr	Aft	Bfr	Aft
DCGAN$_\text{M}$	10	0.49	**0.97**	0.51	**0.99**	0.84	**0.99**	0.53	**0.99**	0.50	**0.63**
DCGAN$_\text{M}$	100	0.49	0.61	0.51	0.98	0.76	0.98	0.53	0.99	0.50	0.52
DCGAN$_\text{M}$	1K	0.49	0.50	0.51	0.81	0.69	0.91	0.53	0.97	0.50	0.51
DCGAN$_\text{C}$	10	0.49	**0.99**	0.49	**0.99**	0.96	**0.99**	0.50	**0.99**	0.49	**0.85**
DCGAN$_\text{C}$	100	0.50	0.96	0.49	0.99	0.92	0.93	0.50	0.99	0.49	0.61
DCGAN$_\text{C}$	1K	0.50	0.62	0.49	0.97	0.88	0.91	0.50	0.99	0.49	0.51
PGAN	100	0.50	**0.98**	0.50	**0.99**	0.96	**0.99**	0.96	**0.99**	0.50	**0.81**
PGAN	1K	0.50	0.89	0.49	0.95	0.94	0.95	0.88	0.99	0.50	0.60
PGAN	10K	0.50	0.61	0.50	0.76	0.89	0.90	0.76	0.98	0.50	0.51
CycleGAN	1K	0.49	**0.92**	0.50	**0.87**	0.98	**0.99**	0.55	**0.99**	0.49	**0.62**
CycleGAN	10K	0.49	0.70	0.50	0.66	0.94	0.96	0.52	0.98	0.50	0.51
DCGAN$_\text{M}$	10	0.02	**0.94**	0.03	**0.88**	0.77	**0.95**	0.16	**0.98**	0.00	**0.26**
DCGAN$_\text{M}$	100	0.00	0.87	0.00	0.85	0.73	0.90	0.10	0.95	0.00	0.13
DCGAN$_\text{M}$	1K	0.00	0.75	0.00	0.80	0.63	0.80	0.10	0.91	0.00	0.05
DCGAN$_\text{C}$	10	0.00	**0.98**	0.00	**0.99**	0.89	**0.93**	0.07	**0.98**	0.00	**0.70**
DCGAN$_\text{C}$	100	0.00	0.95	0.00	0.93	0.82	0.85	0.02	0.93	0.00	0.61
DCGAN$_\text{C}$	1K	0.00	0.90	0.00	0.89	0.77	0.81	0.00	0.88	0.00	0.43
PGAN	100	0.26	**1.00**	0.21	**1.00**	0.99	**0.99**	0.99	**0.99**	0.00	**0.99**
PGAN	1K	0.21	0.99	0.00	0.99	0.97	0.98	0.98	0.99	0.00	0.54
PGAN	10K	0.00	0.51	0.00	0.90	0.90	0.92	0.83	0.99	0.00	0.22
CycleGAN	1K	0.00	**0.99**	0.00	**0.97**	0.97	**0.99**	0.45	**0.99**	0.00	**0.77**
CycleGAN	10K	0.00	0.87	0.00	0.77	0.95	0.96	0.30	0.99	0.00	0.31

Table 4: $||\Delta x||$ (top) and FID score (btm). Standard deviations in parenthesis. DCGAN$_\text{M}$: DCGAN for MNIST, DCGAN$_\text{C}$: DCGAN for CelebA, `Combi.`: Combination attack. *Lower is better.*

Model	C	Baseline	`Blurring`	`Cropping`	`Noise`	`JPEG`	`Combi.`
DCGAN$_\text{M}$	10	5.05(0.09)	15.96(2.18)	9.17(0.65)	5.93(0.34)	6.48(0.94)	17.08(1.86)
DCGAN$_\text{M}$	100	4.09(0.53)	12.95(4.47)	7.62(1.55)	4.57(0.78)	4.70(1.02)	12.70(3.37)
DCGAN$_\text{M}$	1K	**3.88(0.60)**	**7.17(2.10)**	**7.43(1.37)**	**4.22(0.77)**	**5.12(1.94)**	**7.56(1.41)**
DCGAN$_\text{C}$	10	5.63(0.11)	11.83(0.65)	9.30(0.31)	4.75(0.17)	6.01(0.29)	13.69(0.59)
DCGAN$_\text{C}$	100	3.08(0.27)	10.00(1.61)	7.80(0.58)	3.20(0.45)	4.26(0.59)	11.65(1.48)
DCGAN$_\text{C}$	1K	**2.55(0.36)**	**7.68(1.53)**	**7.13(0.47)**	**2.65(0.24)**	**3.39(0.58)**	**9.23(1.22)**
PGAN	100	9.29(0.95)	18.49(2.04)	21.27(0.81)	10.20(0.81)	10.08(1.03)	24.82(2.33)
PGAN	1K	6.52(1.85)	14.79(4.15)	18.88(1.96)	6.40(1.48)	7.09(1.62)	22.09(2.12)
PGAN	10K	**5.04(1.63)**	**10.19(2.87)**	**18.23(0.94)**	**5.13(1.14)**	**5.67(1.62)**	**17.26(1.39)**
CycleGAN	1K	55.85(3.67)	68.03(3.62)	80.03(3.59)	55.47(1.60)	57.42(2.00)	83.94(4.66)
CycleGAN	10K	**49.66(5.01)**	**58.64(3.70)**	**66.05(3.47)**	**53.14(0.44)**	**54.52(2.30)**	**66.24(5.29)**
DCGAN$_\text{M}$	10	5.36(0.12)	41.11(20.43)	21.58(2.44)	5.79(0.19)	6.50(1.70)	68.16(24.67)
DCGAN$_\text{M}$	100	5.32(0.11)	23.83(14.29)	18.39(3.70)	5.41(0.18)	5.46(0.11)	36.05(16.20)
DCGAN$_\text{M}$	1K	**5.23(0.12)**	**10.85(4.28)**	**18.08(1.77)**	**5.37(0.14)**	**5.30(0.96)**	**21.86(4.16)**
DCGAN$_\text{C}$	10	53.91(2.20)	73.62(6.70)	98.86(9.51)	59.51(1.60)	60.35(2.57)	87.29(9.29)
DCGAN$_\text{C}$	100	45.02(3.37)	73.12(11.03)	85.50(12.25)	47.60(2.57)	50.48(4.58)	78.11(12.95)
DCGAN$_\text{C}$	1K	**40.85(3.41)**	**55.63(7.97)**	**72.11(13.81)**	**40.87(3.03)**	**45.46(5.03)**	**57.13(7.20)**
PGAN	100	21.62(1.73)	28.15(3.43)	47.94(5.71)	25.43(2.19)	22.86(2.06)	45.16(7.87)
PGAN	1K	19.05(3.14)	25.19(5.26)	43.48(12.24)	19.20(2.96)	19.05(2.82)	35.07(8.72)
PGAN	10K	**16.75(1.87)**	**18.96(2.65)**	**37.01(8.74)**	**16.94(1.89)**	**17.39(2.33)**	**26.63(4.44)**