Supplementary Material for Learning to Generate Inversion-Resistant Model Explanations

1 Baseline Defense

In §5.3, we evaluate the efficacy GNIME by comparing it with two baseline defenses: Random Noise Defense (RND) and Optimized Noise Defense (OND). Each defense is set up to perturb explanations of the target model against an explanation-aware model inversion (ExpMI) adversary.

1.1 Random Noise Defense

As the primitive defense technique, we apply pixel-wise Gaussian noise to the explanations and assess its efficacy in defending ExpMI. For fair comparisons, the perturbation size must be the same with it of GNIME. In this regard, we generate the noise filter of RND from a normal distribution of a variance, which is same as the pixel variance of GNIME perturbations. Formally, the RND-perturbed explanation e' is defined as:

$$e' = e + n',\ where\ n' \sim N(0, \sigma^2(n)). \tag{1}$$

1.2 Optimized Noise Defense

The other baseline, OND, attempts to iteratively update the explanations in the direction where the ExpMI inversion quality is degraded. We adopt the idea of Projected Gradient Descent (PGD) devised for adversarial training [4]. The general idea of PGD attack is to iteratively update the input image in the desired direction of the loss function. For the targeted PGD attack, let $\bar{y}$ denote the desired misclassification label, then the PGD iteration is defined as:

$$x^{t+1} := x^t + \alpha\ sign[\bigtriangledown_{x^t} \mathcal{L}_{ce}(f(x^t), \bar{y})]. \tag{2}$$

The image will move to the direction where the cross-entropy loss $\mathcal{L}_{ce}$ is reduced. However, we make two changes to equation 2: (1) we update the explanation instead of the image, and (2) we update in the direction to increase the image reconstruction loss $\mathcal{L}_{re}$. Formally, a single OND step is defined as:

$$e^{t+1} := e^t + \alpha\ sign[\bigtriangleup_{e^t} \mathcal{L}_{re}(f_{OND}(e^t, \hat{y}), x)]. \tag{3}$$

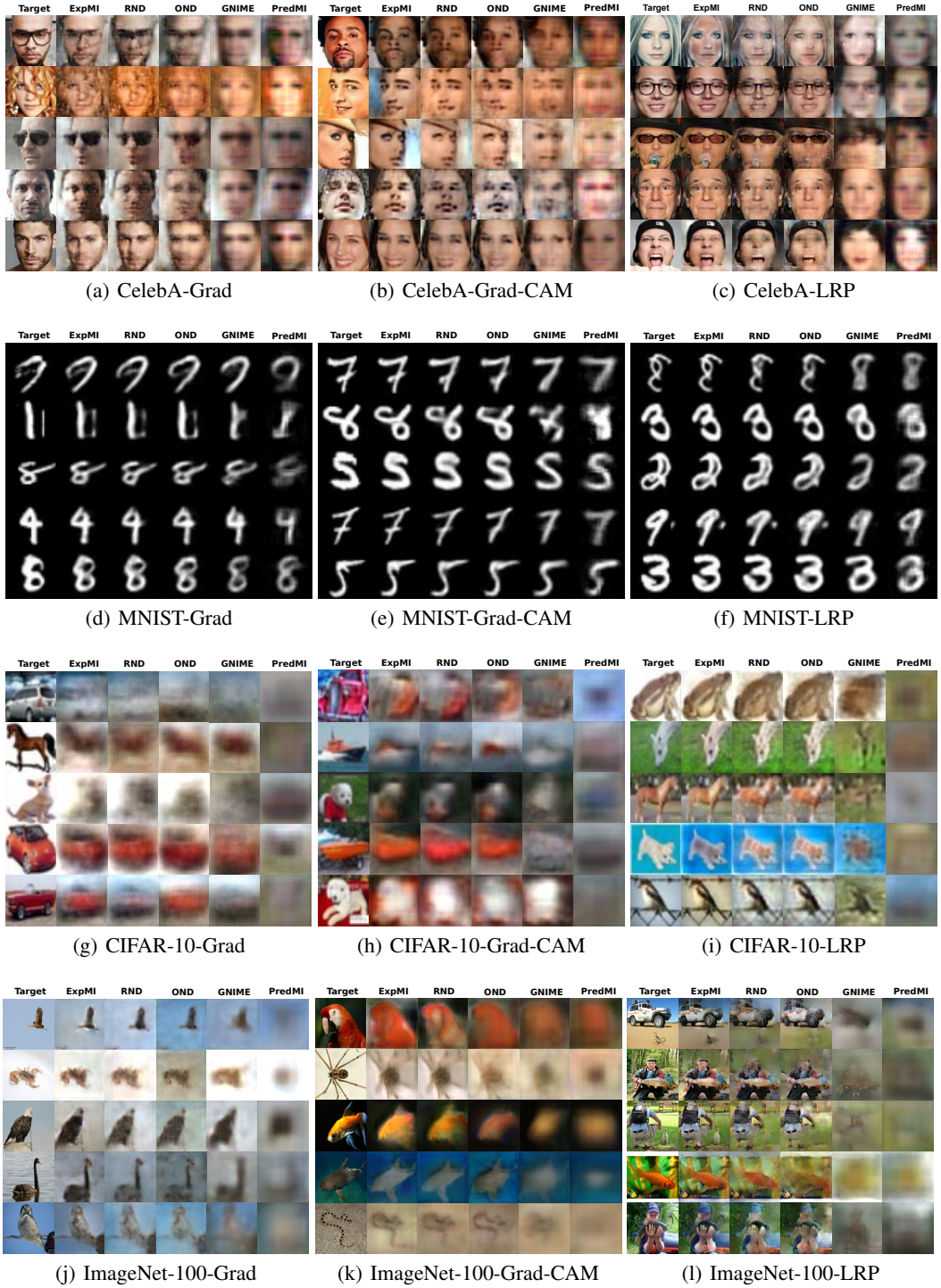
Note that the defender must train a dedicated ExpMI inversion network $f_{OND}(e, \hat{y}) \rightarrow \hat{x}$ in advance to apply such optimization. Leveraging f_{OND}, we iterate Equation 2 30 steps with $\alpha = 0.01$. After 30 iterations, we clip the perturbation magnitude to match GNIME.

2 Qualitative Analysis

2.1 Inversion Results

Figure 1 confirms that GNIME outperforms the baseline techniques (RND and OND) and degrades MI attack the performance close to the ideal lower bound (PredMI). With reconstructed facial images from ExpMI attacks against no defense, OND, and RND (columns 2–4), we can easily identify most

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) CelebA-Grad (b) CelebA-Grad-CAM (c) CelebA-LRP

(d) MNIST-Grad (e) MNIST-Grad-CAM (f) MNIST-LRP

(g) CIFAR-10-Grad (h) CIFAR-10-Grad-CAM (i) CIFAR-10-LRP

(j) ImageNet-100-Grad (k) ImageNet-100-Grad-CAM (l) ImageNet-100-LRP

Figure 1: For each dataset and explanation type, we compare ExpMI inversion results against RND, OND, and GNIME (columns 2–5), along with the results of the worst-case (column 1 – ExpMI against no defense) and best-case scenario (column 6 – PredMI; which does not utilize explanation).

(a) CelebA-Grad (b) CelebA-Grad-CAM (c) CelebA-LRP

Figure 2: Qualitative comparison of original image x, model explanation e, and perturbed model explanation $\hat{e}$. For LRP, we increased brightness by 20% for better visualization.

facial features such as gender, facial expression, posture, hair style, mustache, and accessories. By contrast, GNIME renders difficult to reconstruct original images against ExpMI attacks.

One may question GNIME 's effectiveness on MNIST given that the fifth columns of figure 1 (d-f) shows perfect digits. Please note that the limited complexity of handwritten digits causes even PredMI attacks to produce identifiable inversion results. Still, one can acknowledge from the figures that only GNIME succeeds in concealing subtle handwriting features – observe that the cross strokes of '7' and the small dot next to '9' disappear in figure 1 (c) and (f), respectively.

The wide variety of CIFAR-10 images makes the overall inversion task challenging. Accordingly, even the inversion results of ExpMI with no defense (second columns) are blurry. Nevertheless, the contours of the target images can be identified from columns 2–4, whereas GNIME contributes to significantly blurring the contours.

2.2 Perturbation Magnitute

Figure 2 compares the original model explanations of CelebA to their perturbed versions generated by GNIME. Note that the perturbations are hardly perceptable, which is a rough indicator that GNIME has successfully injected minimal perturbation.

3 Impact of Perturbation Magnitude on Defense

$$e' = e + \frac{n}{\sqrt{\sigma^2(n)/\gamma}},\ where\ n = \hat{e} - e. \tag{4}$$

Recall from §4.1 that one can vary γ in Equation 4 to control the perturbation magnitude. We thus evaluate the trade-off between (1) defending against MI attacks and (2) preserving XAI functionality with varying magnitudes of perturbations. Specifically, we compute the former via the MSE between x and $\hat{x}$, and the latter using the DeePSiM between $x \odot e$ and $x \odot \hat{e}$.

Figure 3 illustrates how the MSE and DeePSiM vary over different perturbation sizes. The inversion model for each magnitude was trained separately to ensure the optimal MI adversary for the MSE value of each case. We observed a linear trade-off relationship between the defensive capability and the explanation functionality of GNIME. As the figure shows, the defensive capability increases when

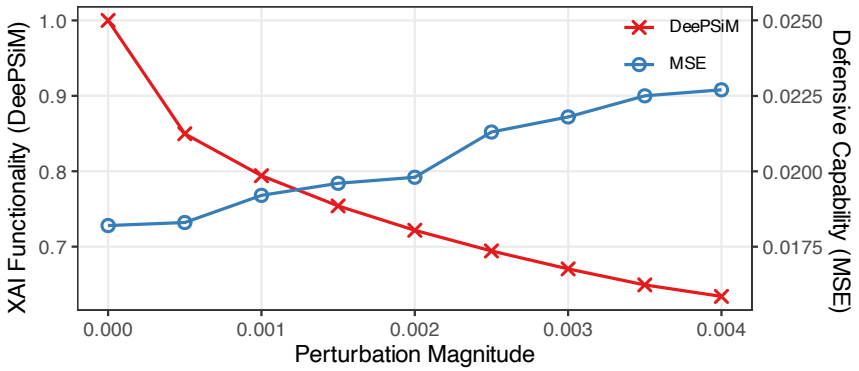


Figure 3: Impact of perturbation magnitude of GNIME on Grad-CAM explanations of CelebA to MI defense and XAI functionality preservation.

the perturbation size becomes larger. On the other hand, the XAI functionality gradually degrades, as larger noise is added to the explanation. We also observed a similar trade-off correlation between these two factors when we measure the defensive capability using SSIM, TCA, and DeePSiM instead of MSE, which demonstrates the generality of this trade-off correlation. Hence, the GNIME service providers can use γ of their choice with this in mind. For instance, if they choose to use 0.0040 as the perturbation magnitude, GNIME achieves the 0.6340 DeePSiM. However, when limiting the perturbation size to 0.0005, the DeePSiM score increases to 0.8497. Note that changing γ does not require any retraining.

4 Defense against ExpMI using a Surrogate Model

Zhoa *et al.* introduced a variant ExpMI attack (ExpMI-3), which is applicable even when a target DNN model does not provide model explanations [5]. In ExpMI-3, two additional models (a surrogate target model and an explanation inversion model) should be trained before the training of the actual ExpMI inversion model. The sanity of the two inversion models (i.e., explanation inversion and ExpMI inversion models) depends on the degree to which the surrogate model copies the original task of the target model. Therefore, a successful model extraction (ME) attack is the key to its success. We argue that the threat due to ExpMI-3 should be addressed by mitigating the ME threat, rather than the MI threat. To support this, we implemented a simple ME defense mechanism against ExpMI-3 and compared its inversion quality to ExpMI against GNIME.

Table 1: Comparison between ExpMI against GNIME and ExpMI-3 against RSP in terms of inversion performance

Dataset	Metric	ExpMI	GNIME	ExpMI-3 w/ RSP
CelebA	MSE (↑)	.0067	.0254 (.0187↑)	**.0249 (.0182↑)**
	SSIM (↓)	.8306	.4706 (.3600↓)	**.4928 (.3378↓)**
	TCA (↓)	.2190	.0333 (.1857↓)	**.0350 (.1840↓)**
	DeePSiM (↓)	.5066	.1923 (.3143↓)	**.1926 (.3140↓)**

In this evaluation, we used reverse sigmoid perturbation [1] (RSP) as the simple ME defense mechanism. The intuition of RSP is to transform the final output of a victim classifier so that it becomes difficult to infer the internal function. Specifically, it maps different values to the same output, making it hard to determine which value it originally belongs to in the context of ME. Table 1 compares the inversion performance degradation effect of GNIME and RSP. Notice that the inversion quality is degraded to the level of GNIME even with such a simple defense approach.

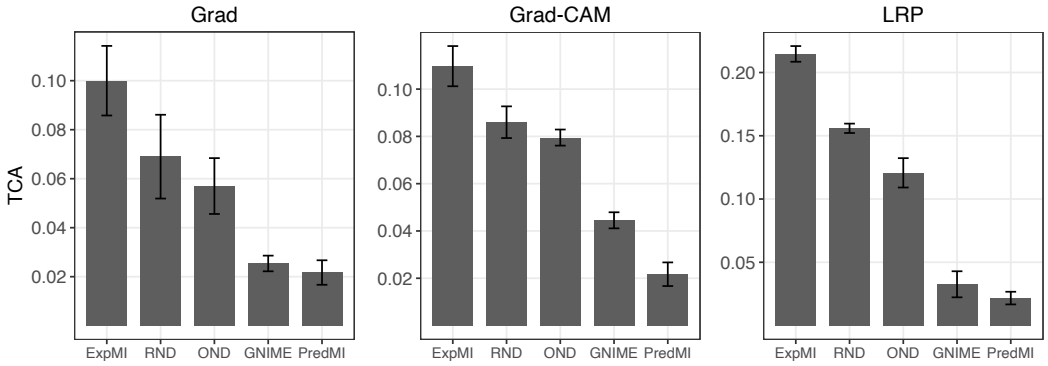


Figure 4: TCA measured on GNIME and the other defenses for CelebA. Note that the values are identical from those of Table 1 in the main paper.

5 Statistical Significance of the Differences between GNIME and the Other Defenses

In Table 1 of the main paper, we reported the quantitative comparison results between the defensive capability of GNIME and the other defenses. We now demonstrate that these differences are statistically significant. Figure 4 illustrates the same results from Table 1 of the main paper. We show TCA for CelebA, which shows the most prominent differences across different defenses. Note from the figure that GNIME significantly outperforms other defenses, and the differences between them are statistically significant.

We further performed a statistical test on results reported in Table 1 of the main paper. Specifically, we conducted two-tailed Mann Whitney U tests on results reported by GNIME and another defense. As shown in Table 2, all the differences between GNIME and another defense are statistically significant with p-values less than 0.05 except three cases. These results highlight that the superior performance of GNIME over the other defenses is statistically significant.

6 Defense against an Adversary Knowing the Presence of GNIME

In this section, we discuss the threat of an adversary who knows that GNIME is adopted by the target service. Such an adversary can come up with three different strategies: (1) avoiding the use of perturbed explanations (i.e., utilize only the prediction vectors) for training the inversion model; (2) training a surrogate model in mimicry of the target model to generate clean explanations; or (3) training the inversion model to reconstruct inputs from perturbed explanations.

Conducting an MI attack without explanations. This case coincides with PredMI. Considering that the PredMI inversion performance is even lower than that of GNIME, the adversary has little incentive to take this strategy. Furthermore, GNIME does not focus on addressing the inversion risk due to PredMI.

Training a surrogate model. The adversary may attempt to create a surrogate model by conducting model extraction attacks. The adversary can directly generate clean model explanations from this surrogate model, as shown in [5]. Otherwise, the adversary can also use the surrogate model to train a denoiser that removes perturbations in our model explanations. However, in both cases, model extraction should successfully copy the victim model's functionality into the surrogate model in the first place. Therefore, we believe that this case should be addressed by mitigating the model extraction threat. We have already discussed this case in detail in the limitations section and supplementary material. In §4 of the supplementary material, we showed that a simple model extraction defense of leveraging reverse sigmoid perturbation can mitigate this threat to the degree to which GNIME can protect the target model.

Training an inversion model with perturbed explanations. The adversary may try to train an inversion model that can directly reconstruct inputs from perturbed model explanations. Note that all

Table 2: p-values reported by Mann-Whitney U tests between GNIME and the other defenses. We marked results in bold when the difference between GNIME and the other defense is statistically significant.

	Metric	Grad			Grad-CAM			LRP		
		vs. ExpMI	vs. RND	vs. OND	vs. ExpMI	vs. RND	vs. OND	vs. ExpMI	vs. RND	vs. OND
CelebA	MSE↑	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	SSIM↓	**.01208**	.06010	.09492	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	TCA↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	DeePSiM↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
MNIST	MSE↑	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	SSIM↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	TCA↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	DeePSiM↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
CIFAR-10	MSE↑	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	SSIM↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	TCA↓	**.01208**	**.03662**	.06010	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	DeePSiM↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
ImageNet-100	MSE↑	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	SSIM↓	**.01208**	**.02144**	**.03662**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.02144**
	TCA↓	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**
	DeePSiM↓	**.01208**	**.01208**	**.03662**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**	**.01208**

evaluation in this paper already assumes this adversary. We demonstrated that GNIME successfully mitigates the model inversion threat from such an attacker.

7 Convergence Analysis

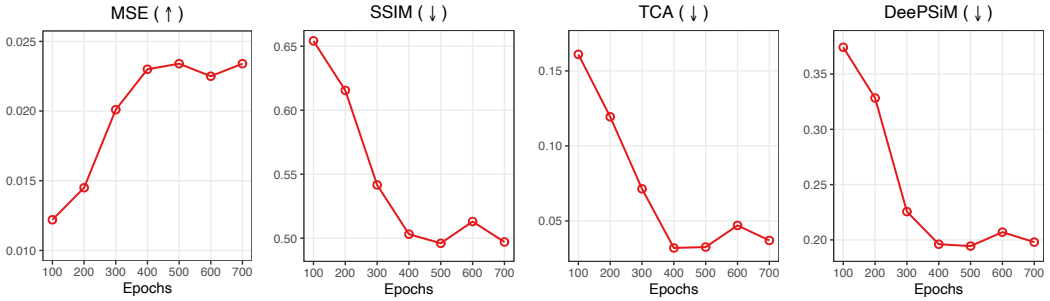


Figure 5: Empirical convergence analysis of CelebA-LRP.

Similar to previous works [2, 3] demonstrating their empirical loss convergence, we conduct an empirical analysis. We show that GNIME becomes better at injecting noise that undermines the inversion performance as the epoch in training the NG network increases. Specifically, we deploy GNIME of each hundredth epoch and evaluate the performance of ExpMI. Note that, for fair comparison, we clip the perturbation magnitude of every other epoch to match it of $epoch = 500$. From Figure 5, we confirm that f_{NG} becomes better at deteriorating the inversion performances (i.e., MSE, SSIM, TCA, and DeepPSiM) as the epoch increases and starts to plateau after 500 epochs.

References

[1] Taesung Lee, Benjamin Edwards, Ian Molloy, and Dong Su. Defending against neural network model stealing attacks using deceptive perturbations. In *IEEE Security and Privacy Workshops*, pages 43–49, 2019.

[2] Yan Li and Jieping Ye. Learning adversarial networks for semi-supervised text classification via policy gradient. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1715–1723, 2018.

[3] Yan Li, Ethan Fang, Huan Xu, and Tuo Zhao. Implicit bias of gradient descent based adversarial training on separable data. In *Proceedings of the International Conference on Learning Representations*, 2020.

[4] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the International Conference on Learning Representations*, 2018.

[5] Xuejun Zhao, Wencan Zhang, Xiaokui Xiao, and Brian Lim. Exploiting explanations for model inversion attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 682–692, 2021.