ADAPTIVE MODEL INVERSION ATTACKS GENERALIZE A PRIVACY-ROBUSTNESS TRADEOFF

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we show that standard evaluations of high-resolution Model Inversion Attacks (MIAs) significantly underestimate training-data privacy leakage. State-of-the-art privacy defenses, standard training techniques such as MixUp and Adversarial Training, and undefended models all leak training images at rates 1.16 to $6.59\times$ higher on FaceScrub under simple adaptive changes to the attack, with the largest increases among defenses reporting the strongest privacy. We further show that measured leakage depends on the feature basis of the external classifier used to evaluate reconstructions: for the same reconstructed images, an adversarially trained Inception evaluator identifies the targeted identity at different rates than the standard Inception evaluator. Our results suggest that standard MIA evaluation can mistake optimization and measurement failures for privacy.

These underestimated leakage rates also concealed a broader relationship between privacy and adversarial robustness. Once we adapt the attack and vary the evaluator, reconstruction leakage closely tracks adversarial robustness across recent defenses and standard training regimes, suggesting that robustness provides an attack-agnostic proxy for reconstruction vulnerability that applies far more broadly than previously theorized. This raises an open question: can a practical defense reduce training-data reconstruction without paying a corresponding cost in adversarial robustness?

1 INTRODUCTION

Model inversion attacks (MIAs) provide an empirical test of training-data privacy leakage under a powerful adversary. In the high-resolution vision setting, an attacker receives white-box access to a trained model, substantial auxiliary data, and a generative image prior, and attempts to reconstruct recognizable representatives of private training classes (Fredrikson et al., 2015; Zhang et al., 2020; Chen et al., 2021; Struppek et al., 2022; An et al., 2022; Nguyen et al., 2023; Peng et al., 2024; Qiu et al., 2024). Successful reconstruction therefore establishes that recognizable information about the private training data, such as extractable facial geometry, was exposed by the model. MIA evaluation is therefore an empirical worst-case lower bound on training-data exposure: measured leakage is only as strong as the attacker used to measure it. A low training data reconstruction rate may indicate a private model. Alternatively, it may simply reflect a failed attack.

The limited-scope privacy–robustness tradeoff and its outliers. Recent work proposed a surprisingly simple explanation for this leakage. Torp et al. (2026) found that, across three generic MIA defenses (MID, BiDO, and TL-DMI) (Wang et al., 2020; Peng et al., 2022; Ho et al., 2024), reconstruction leakage could be predicted almost perfectly from clean and adversarial accuracy (pseudo-$R^2 = 0.952$). They attribute this relationship to the model's feature basis: robust features are theorized to be semantically meaningful and therefore useful for visual reconstruction (see, e.g., Ilyas et al., 2019; Engstrom et al., 2019; Tsipras et al., 2019), whereas non-robust features remain predictive but are visually imperceptible. Deliberately shifting models toward non-robust features correspondingly reduced reconstruction. The result suggested a striking connection between two normally separate Trustworthy-ML objectives: robustness to adversarial examples and data privacy.

The result is a fascinating theory bridging robustness and privacy, yet it is also theorized *not* to generalize to many of the defenses and training techniques that are relevant in practice. Specifically, the original result explicitly restricted the scope of the theorized privacy–robustness trade-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

off to three generic defenses, MID, BiDO, and TL-DMI, and the theory explicitly excludes recent high-performance defenses such as NegLS, RoLSS, RoLSS-SSF, and Trap-MID (Struppek et al., 2024; Koh et al., 2024; Liu & Chen, 2024) that interfere with the gradients used by contemporary MIAs. Also excluded from the theory are standard data augmentation and training techniques such as MixUp and Adversarial Training (AT) (Zhang et al., 2018; Madry et al., 2017) that practitioners commonly deploy to improve generalization.

Fig. 1 applies the published regression coefficients to this broader set. The fit collapses from pseudo-$R^2 = 0.952$ (RMSE $= 0.050$) on the original defenses to pseudo-$R^2 = 0.317$ (RMSE $= 0.343$) on this out-of-sample set. Most strikingly, AT models leak far less than their extreme robustness predicts, while recent NegLS and MIDRE (Tran et al., 2025) defenses appear to obtain substantial privacy without the corresponding robustness cost. The AT result is particularly surprising because adversarial training produces the robust feature basis implicated by the theory and has previously been shown to increase vulnerability to model inversion (Mejia et al., 2019).

There are two possible explanations. The first is simply that adversarial robustness was a strong correlate of privacy for three particular defense families, but as Torp et al. (2026) theorized, it is not a general property of model inversion.

Are defenses' privacy claims overfit to an evaluation that hides generalizable properties?
In this paper, we consider a second possibility: otherwise generalizable privacy properties that apply across defenses and training techniques in the field are spuriously obscured because contemporary defense evaluations are overfit to a fixed configuration of attacks and evaluations, resulting in falsely inflated claims of privacy improvements among top defenses. The problem we describe is familiar from adversarial robustness. There, defenses against adversarial examples repeatedly appeared robust because they were evaluated against attacks whose optimization failed on the defense, motivating adaptive attacks that account for the defense being evaluated (Tramèr et al., 2020; Athalye et al., 2018). The analogous principle has not become standard in MIA. Recent hi-res MIA defenses are often evaluated using fixed attacks such as PPA (Struppek et al., 2022) with the attack's loss, optimizer, and hyperparameters set to (non-adaptive) published values, even when the defense itself changes the gradients on which that optimization depends (see App. C.1).

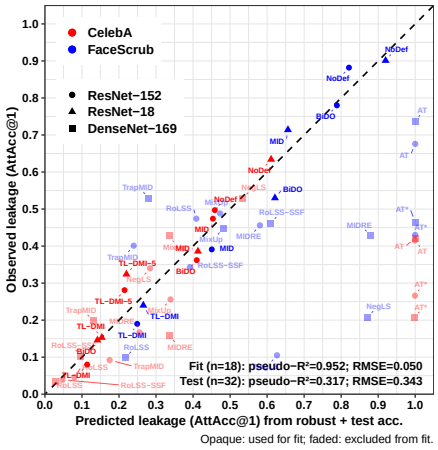


Figure 1: The privacy–robustness relationship in Torp et al. (2026) does not generalize under standard MIA evaluation. Opaque points were included in the original Torp et al. (2026) regression.

A recent geometric account of model inversion suggests a mechanism by which the privacy–robustness relationship could generalize beyond the defenses on which it was originally fit. Peng et al. (2025) observe that generative MIAs do not follow the full input-space gradient: because optimization proceeds through a generator $G(z)$, the attack can move only along directions in the tangent space of the generator manifold. This projection suppresses off-manifold components of the inversion gradient, and Peng et al. show that inversion succeeds more readily when the target model's loss gradients are better aligned with this manifold. Relatedly, Srinivas et al. (2023) tie perceptually aligned input gradients closely to adversarial robustness: for Bayes-optimal classifiers, perceptual alignment with the signal manifold corresponds to off-manifold robustness, while empirically several forms of robust training induce off-manifold robustness, and in turn, perceptually aligned gradients. These are not the same notions of alignment: Peng et al. measure alignment to the generator manifold during inversion, whereas Srinivas et al. study alignment to the signal manifold of the data distribution. However, they suggest a common geometric possibility. Models whose predictive gradients concentrate on semantically meaningful, on-manifold directions may be both more adversarially robust and easier for a generator-constrained attack to reconstruct. If so, defenses that appear to escape the privacy–robustness relationship because a fixed attack fails to optimize may not remove the underlying exposure of their training data; they may instead make training data features harder for a particular attack optimizer configuration to extract.

Our contribution. We test these apparent exceptions under adaptive model inversion across 13 defenses and generic training techniques, spanning nine MIA defenses, an undefended baseline, MixUp, and two Adversarial Training (AT) strengths; two standard high-resolution target architectures (ResNet-152 and DenseNet-169); and the two main high-resolution MIA datasets (FaceScrub and CelebA), under PPA and IF-GMI. We make simple, generic changes to the attack loss function and optimization while holding the attack generator and target model fixed.

We show that under these simple optimization changes, SOTA defenses' measured privacy leakage increases by factors of 1.16–$6.59\times$ in terms of standard AttAcc@1 on FaceScrub, with larger differences between vanilla and adaptive attacks corresponding to several of the recent defenses with the highest claimed performance. On ResNet-152, an adaptive variant increases AttAcc@1 for 25 of 26 model–dataset pairs. For example, NegLS AttAcc@1 on FaceScrub changes from 0.105 to 0.692 and TL-DMI from 0.151 to 0.459; undefended models and models trained with MixUp or AT leak 1.18–$2.61\times$ more with our techniques. We also show that the leakage increases generalize to, e.g., CelebA data, DenseNet-169 architectures, and our techniques also apply to IF-GMI attacks.

Attack optimization is only one source of measurement dependence. High-resolution MIAs conventionally score whether a reconstruction recovered its target identity using an independent Inception-v3 classifier (Struppek et al., 2022; Szegedy et al., 2016). But the feature account of Torp et al. (2026) suggests that this measurement may depend on the evaluator's own feature basis. We therefore score the same reconstructions with both standard and adversarially trained Inception evaluators. Neither evaluator is treated as ground truth; disagreement between them measures how strongly reported leakage depends on the evaluator's feature basis. For example, a generic adversarially-trained RN-152 model's AttAcc@1 leakage changes from 0.430 under the standard evaluator to 0.901 under the robust evaluator, on the same reconstructed images. The results indicate that evaluator models used to compute the main privacy metric in the high-resolution literature underestimate leakage, particularly where the target model relies on robust, visually meaningful features.

Finally, we return to the regression in Fig. 1. Combining adaptive attacks with the robust evaluator yields pseudo-$R^2 = 0.740$ (RMSE $= 0.156$), without adding the held-out defenses or training regimes to the fitting set. This suggests that the apparent failure of the tradeoff is partly a property of the standard evaluation choices rather than evidence that recent defenses break the relationship: once the attack is adapted, the same robustness predictors generalize substantially better to defenses and training regimes that were never used to fit them, including defenses explicitly designated as exceptions in Torp et al. (2026). The fit further improves under a robust evaluator.

Implications for a generic privacy–robustness tradeoff. Our results have two implications for the MIA and Trustworthy-AI fields. First, our results suggest a broader interpretation of the privacy–robustness relationship. Adversarial robustness is not merely a correlate that happened to fit three MIA defenses. In the high-resolution image-classification setting we study, it captures a property of the learned representation that is also implicated by recent geometric accounts of inversion: whether semantically meaningful directions are available to be reconstructed. Gradient manipulation can make those directions harder for a certain optimizer to find, but does not necessarily remove them.

This frames a key question for future research: is it possible to design a defense against training data reconstruction that obtains privacy without a corresponding sacrifice of adversarial robustness?

Implications for MIA privacy defense evaluation. Second, MIA evaluations have substantially underestimated the degree to which state-of-the-art privacy defenses leak information about their training data. By developing defenses that change the gradients on which attack optimization depends, while continuing to evaluate them with fixed attacks, fixed hyperparameters, and a single Inception feature basis, the field has recreated a problem familiar from adversarial examples: apparent privacy can reflect failure of the evaluation procedure rather than removal of the underlying vulnerability. We observe both forms of dependence here; simple adaptive changes recover substantially more leakage from most defenses, while the same reconstructed images can receive sharply different privacy scores under an alternative evaluator feature basis.

This changes the appropriate standard for evaluating whether an empirical MIA defense succeeds. A privacy defense should survive an attacker that adapts to the defense, and its reconstructed images should be evaluated by more than one arbitrarily chosen classifier feature basis. Otherwise, we risk calling an optimization failure 'privacy'.

2 Adaptive Attacks and Robust Leakage Evaluation under MIA

Our goal is to test whether the measured privacy gains of model inversion defenses persist when we change two fixed choices in the standard evaluation: how the attack optimizes reconstructions and how an external classifier scores them. We keep the trained target models fixed throughout.

2.1 Adaptive attacks

A defense may make reconstruction difficult across attack settings. Alternatively, it may change the attacker's loss landscape in a way that frustrates the particular loss and learning rate used in standard evaluations. We test whether the reported privacy gains survive two ordinary changes to the white-box Plug-and-Play Attack (PPA) of Struppek et al. (2022) used as the primary high-resolution evaluation attack in recent MIA defense evaluations (e.g., Torp et al., 2026; Tran et al., 2025; Liu & Chen, 2024; Struppek et al., 2024). Specifically, we consider replacing its loss and increasing its learning rate by a factor of ten. The generator and target model remain fixed. The same changes can be applied to other attacks that optimize generator latents using target-model gradients.

Attack losses. Let the GAN used by generative model inversion attacks be denoted as G, and let the target classifier T be trained on C classes. PPA optimizes a StyleGAN-2 (Karras et al., 2020) latent vector z for a target identity c using gradients through the target classifier's logits $o = T(G(z)) \in \mathbb{R}^C$. Standard PPA uses a Poincaré loss and an initial Adam learning rate of 0.005, reduced after iterations 30 and 40. Poincaré loss was introduced to address a particular failure mode of cross-entropy: its gradients vanish as target confidence approaches one, which can prevent late-stage refinement of poorly initialized latents (Struppek et al., 2022). The defenses we study can create different optimization failures. Some alter the loss landscape such that PPA encounters local modes before reaching high target confidence; others, such as NegLS, directly suppress the gradients used by the attack. We therefore test whether PPA's fixed objective is itself responsible for part of the claimed privacy gain by replacing Poincaré with either vanilla cross-entropy or cross-entropy with negative label smoothing (an attack adaptation inspired by the NegLS defense). Let $p_j(o) = \exp(o_j)/\sum_k \exp(o_k)$, and denote the one-hot target by e_c.

Then, both alternatives take the form

$$q^{(c)} = (1-\alpha)e_c + \frac{\alpha}{C}\mathbf{1}, \qquad \mathcal{L}_\alpha(o,c) = -\sum_{j=1}^{C} q_j^{(c)} \log p_j(o), \tag{1}$$

with $\alpha = 0$ for ordinary cross-entropy and $\alpha = -0.05$ for negative label smoothing. With $\alpha < 0$, the off-target weights are negative and the gradient with respect to the target logit does not vanish as $p_c(o)$ approaches one. Negative label smoothing therefore retains a class-directed gradient even in the high-confidence regime that originally motivated PPA's substitution of Poincaré loss. We apply both losses to all target models, including those trained without negative label smoothing.

Tenfold learning rate. A defense that introduces local modes into PPA's optimization landscape may impede the default optimizer's progress without preventing reconstruction at some other step size. We test this possibility in the simplest conceivable way by raising the initial learning rate tenfold from 0.005 to 0.05 and reducing it after iterations 20, 30, and 40. We refer to our cross-entropy and negatively smoothed attacks at this tenfold-increased learning rate as *C-LR* and *N-LR*.

2.2 Robust evaluation

Separately, a defense's measured privacy gain may reflect how its target model's feature basis aligns with the evaluator's, rather than how much recognizable facial information the attack recovers. The standard Attack Accuracy (AttAcc@1) metric uses an independent Inception-v3 (Szegedy et al., 2016) classifier as a proxy for whether a reconstructed face is recognizable as the targeted identity (Struppek et al., 2022). That proxy depends on the features Inception itself uses to recognize a face. As Torp et al. (2026) show, defenses can shift the target model toward non-robust identity features. A generator may encode these predictive but semantically meaningless features in its reconstructions, and a conventionally trained Inception model may classify from them, even if the reconstructed face is not recognizable to a person. The reverse is also possible: Inception may miss visible facial structure recovered from a target that relies more heavily on robust features.

Accordingly, we test whether defenses' privacy gains persist under an evaluator trained to place more weight on robust features associated with visible facial structure. For each dataset, we score the *same saved reconstructions* using the standard Inception-v3 evaluator and an independently adversarially trained Inception-v3 identity classifier. AttAcc@1 and Robust-AttAcc@1 are the respective fractions for which each evaluator's top prediction matches the attacked identity. We report both scores and each evaluator's accuracy on held-out natural images. The attack, target model, and reconstruction remain fixed in each comparison. Thus, disagreement between AttAcc@1 and Robust-AttAcc@1 empirically lower-bounds the degree to which measured defense performance depends on a single arbitrarily chosen classifier feature basis.

3 EXPERIMENTS

Our goal in this section is to test whether the reported privacy gains of current model inversion defenses survive our adaptive attack techniques across the standard high-resolution evaluation settings in the defense literature. To accomplish this, we run unmodified SOTA PPA attacks alongside PPA attacks adapted with our two natural substitute loss functions with a tenfold increased learning rate, C-LR and N-LR. App. F also applies our techniques to alternate IF-GMI attacks (Qiu et al., 2024).

Experimental scope. We consider 13 target-model conditions: 9 SOTA MIA baselines, 3 models trained using standard techniques for improving generalization (MixUp and AT under two perturbation magnitudes ϵ), and an undefended model (NoDef). The 9 baselines include the generic information-dependency defenses used to estimate the original privacy–robustness relationship (MID (Wang et al., 2020), BiDO (Peng et al., 2022), and TL-DMI (Ho et al., 2024)), recent high-performance and anti-gradient defenses (NegLS (Struppek et al., 2024), Trap-MID (Liu & Chen, 2024), RoLSS (Koh et al., 2024), RoLSS-SSF, MIDRE (Tran et al., 2025)), and a causal instrument (AT-AT (Torp et al., 2026)). We evaluate defenses on the two main hi-res MIA datasets (FaceScrub (Ng & Winkler, 2014) and CelebA (Liu et al., 2015)), across the two main hi-res architectures (ResNet-152 (He et al., 2015) and DenseNet-169 (Huang et al., 2017)), under PPA and our C-LR, and N-LR techniques. This yields 13 target-models $\times$ 3 PPA variants $\times$ 2 datasets $\times$ 2 architectures $=$ 156 conditions. App. F shows our techniques applied to IF-GMI attacks (Qiu et al., 2024).

Standard & robust privacy metrics. For each defense, we report standard test accuracy and 5 reconstruction metrics. We report AttAcc@1 and AttAcc@5, the fractions for which an independent Inception-v3 evaluator ranks the target class first or among its top five predictions, respectively. We also report Robust-AttAcc@1 and Robust-AttAcc@5, which are the same metrics but computed with an adversarially-trained Inception-v3 model. The gap between AttAcc@1 and Robust-AttAcc@1 lower-bounds the sensitivity of measured defense performance to the evaluator's feature basis. We also report the L2-Face metric, which captures reconstructions' distance from private target images in FaceNet representation space.

Results. Fig. 2 plots the performance of each defense under both unmodified SOTA PPA attacks and PPA attacks adapted with our techniques. Table 1 summarizes each defense's worst-case increase in leakage under our techniques as compared to vanilla PPA. On the defenses widely regarded as SOTA in terms of performance, AttAcc@1 leakage increases from, e.g., 0.105 to 0.692 (NegLS on FaceScrub under RN-152), or from 0.401 to 0.873 (Trap-MID on FaceScrub under RN-152). In terms of generic techniques, MixUp increases from, e.g., 0.256 to 0.668 (on CelebA under RN-152). For RN-152, AttAcc@1 exceeds standard PPA on 25 of 26 model–dataset pairs. App. F shows full results, plus results applying our techniques to IF-GMI attacks. **Code is on Anonymous Github.**

We note that these increases in leakage are substantial. For comparison, on FaceScrub under our techniques, SOTA defense Trap-MID leaks more in terms of AttAcc@1 than an undefended model leaks under vanilla PPA. Our results therefore indicate that a substantial share of reported defense performance was overfit against one fixed PPA configuration. Interestingly, the lone 'defense' little-affected by our techniques is AT-AT, which is not a practical defense, but a tool to show privacy gain obtainable by deliberately sacrificing the model's adversarial robustness (Torp et al., 2026).

Robust evaluation results. Table 2 and App. F recompute the 2 metrics that depend on an external Inception-v3 model (AttAcc@1, AttAcc@5) using a robust AT-trained version of this model, holding the attacker reconstructions of each class fixed. For the higher-ϵ FaceScrub AT model, standard PPA obtains AttAcc@1 of 0.430 under the conventional evaluator and 0.901 under the robust evalu-

Table 1: ResNet-152 adaptive attack results. Values show standard PPA → worst-case over {PPA, PPA N-LR, and PPA C-LR}, computed independently. Markers in (·) denote the strongest attack.

	FaceScrub			CelebA		
Model	**TestAcc ↑**	**AttAcc@1 ↓**	**L2-Face ↑**	**TestAcc ↑**	**AttAcc@1 ↓**	**L2-Face ↑**
NoDef	0.943	0.823 → 0.969 (★)	0.806 → 0.712 (★)	0.845	0.459 → 0.695 (★)	0.852 → 0.838 (★)
MID	0.950	0.391 → 0.590 (▲)	0.950 → 0.867 (▲)	0.802	0.474 → 0.781 (★)	0.821 → 0.709 (★)
BiDO	0.920	0.754 → 0.876 (★)	0.832 → 0.797 (★)	0.747	0.362 → 0.521 (★)	0.923 → 0.876 (★)
TL-DMI	0.891	0.151 → 0.459 (★)	1.088 → 0.954 (★)	0.814	0.281 → 0.513 (★)	0.929 → 0.863 (★)
RoLSS	0.886	0.474 → 0.631 (★)	0.858 → 0.832 (★)	0.537	0.046 → 0.072 (★)	1.147 → 1.117 (▲)
RoLSS-SSF	0.869	0.343 → 0.492 (★)	0.918 → 0.894 (★)	0.428	0.039 → 0.045 (▲)	1.183 → 1.183 (▬)
MIDRE	0.941	0.456 → 0.704 (★)	0.891 → 0.815 (★)	0.756	0.167 → 0.258 (★)	0.987 → 0.951 (★)
Trap-MID	0.926	0.401 → 0.873 (★)	0.980 → 0.777 (★)	0.806	0.092 → 0.302 (★)	1.358 → 1.150 (★)
NegLS	0.911	0.105 → 0.692 (★)	1.217 → 0.796 (★)	0.805	0.340 → 0.359 (★)	0.908 → 0.897 (★)
AT-AT	0.941	0.058 → 0.058 (▬)	1.225 → 1.225 (▬)	0.819	0.049 → 0.055 (★)	1.245 → 1.245 (▬)
MixUp	0.971	0.488 → 0.906 (▲)	0.932 → 0.725 (▲)	0.899	0.256 → 0.668 (▲)	0.965 → 0.792 (▲)
AT	0.928	0.676 → 0.842 (★)	0.758 → 0.758 (▬)	0.818	0.415 → 0.569 (★)	0.822 → 0.822 (▬)
AT*	0.891	0.430 → 0.632 (★)	0.849 → 0.829 (★)	0.748	0.266 → 0.364 (★)	0.887 → 0.887 (▬)

PPA (baseline): ▬ PPA N-LR: ★ PPA C-LR: ▲

Table 2: ResNet-152 with standard and robust Inception evaluation. Each pair shows AttAcc@k → Robust-AttAcc@k for $k \in \{1, 5\}$ from the same vanilla PPA run. Bold values mark changes >0.1.

	FaceScrub			CelebA		
Model	**TestAcc ↑**	**AttAcc@1 ↓**	**AttAcc@5 ↓**	**TestAcc ↑**	**AttAcc@1 ↓**	**AttAcc@5 ↓**
NoDef	0.943	**0.823 → 0.463**	**0.961 → 0.761**	0.845	0.459 → 0.518	0.750 → 0.786
MID	0.950	**0.391 → 0.219**	**0.699 → 0.489**	0.802	0.474 → 0.441	0.725 → 0.688
BiDO	0.920	**0.754 → 0.494**	**0.937 → 0.781**	0.747	0.362 → 0.387	0.629 → 0.653
TL-DMI	0.891	0.151 → 0.131	0.384 → 0.318	0.814	0.281 → 0.311	0.550 → 0.591
RoLSS	0.886	0.474 → 0.470	0.760 → 0.777	0.537	0.046 → 0.066	0.145 → 0.182
RoLSS-SSF	0.869	0.343 → 0.350	0.635 → 0.653	0.428	0.039 → 0.030	0.118 → 0.097
MIDRE	0.941	**0.456 → 0.343**	**0.761 → 0.624**	0.756	0.167 → 0.223	0.377 → 0.461
Trap-MID	0.926	**0.401 → 0.232**	**0.676 → 0.481**	0.806	0.092 → 0.097	0.193 → 0.195
NegLS	0.911	0.105 → 0.069	0.277 → 0.200	0.805	0.340 → 0.333	0.623 → 0.639
AT-AT	0.941	0.058 → 0.042	0.190 → 0.145	0.819	0.049 → 0.059	0.138 → 0.151
MixUp	0.971	**0.488 → 0.331**	**0.727 → 0.588**	0.899	0.256 → 0.294	0.483 → 0.511
AT	0.928	**0.676 → 0.934**	0.896 → 0.995	0.818	**0.415 → 0.713**	**0.693 → 0.912**
AT*	0.891	**0.430 → 0.901**	**0.715 → 0.982**	0.748	**0.266 → 0.734**	**0.527 → 0.929**

ator. C-LR rises from 0.545 to 0.960 under the same evaluator substitution, and N-LR from 0.632 to 0.942. These results indicate that standard AttAcc@1 used as the main privacy metric in the high-resolution literature significantly understates leakage precisely where the target model relies most on robust, visually meaningful features.

4 Generalizing the Privacy–Robustness Tradeoff

We now show that the privacy–robustness tradeoff (Torp et al., 2026) applies generally across defenses and generalization training techniques in the field rather than to just a small set of 3 generic defenses, and importantly, the tradeoff does not exclude high-performance anti-gradient defenses as theorized in Torp et al. (2026), or generalization training techniques widely used by practitioners. Fig. 3 plots the fit of the same leakage-vs-robustness regression as Torp et al. (2026), run on the same in-model defenses (opaque points: NoDef, BiDO, MID, TL-DMI). Unlike Torp et al. (2026), we also plot high-performance and anti-gradient defenses (NegLS, Trap-MID, RoLSS, RoLSS-SSF, MIDRE) and generalization techniques (MixUp, AT, and AT*) as an out-of-sample test-set (semi-transparent points). We exclude AT-AT (Torp et al., 2026) here, as it is a causal instrument and

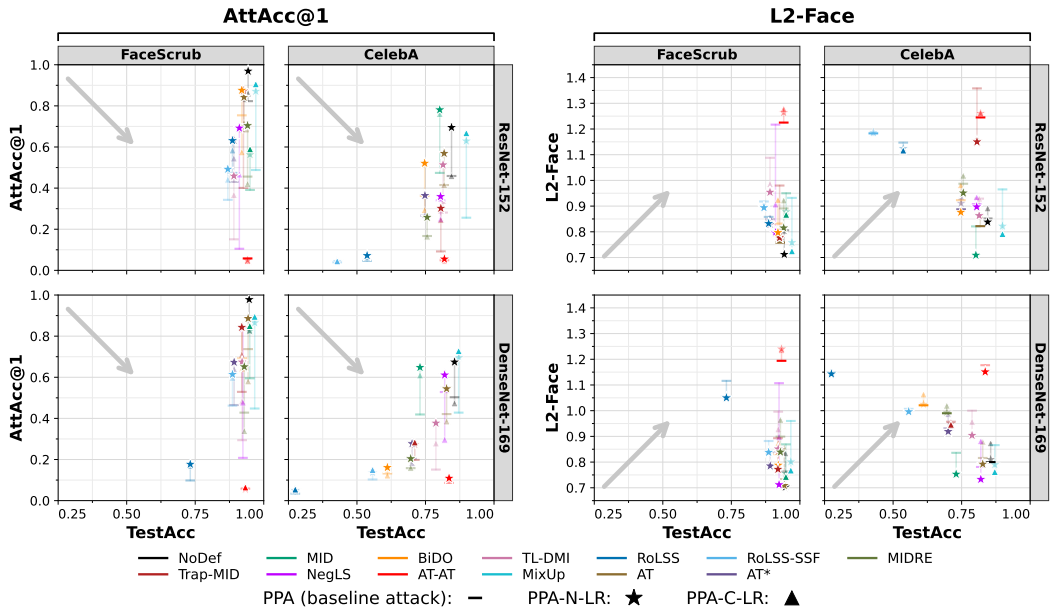


Figure 2: Leakage (AttAcc@1 & L2-Face) vs. test accuracy. Shapes denote attacks. Vertical lines measure vanilla PPA vs. the adaptive attacks, and opaque markers highlight the worst-case leakage.

not a practical defense, and it deliberately optimizes the non-robustness predictor in the regressions (Fig. 6 in the Appendix refits on all 52 models including AT-AT (pseudo-R^2 0.697 and 0.773)). Top-row regressions show the regression model using vanilla PPA as in Torp et al. (2026), and the bottom-row recomputes the regressions using the worst-case over the adaptive attacks and PPA. The left column uses vanilla AttAcc@1, and the right column recomputes with Robust AttAcc@1 under the AT-trained Inception v3 variant. Thus, the bottom-right regression panel combines both changes.

Each change improves privacy–robustness leakage prediction on held-out defenses. Under vanilla PPA and vanilla evaluation (top left), the regression obtains test pseudo-R^2=0.598, RMSE =0.281. Adaptive attacks raise pseudo-R^2 to 0.655 and lower RMSE to 0.205. Robust evaluation raises pseudo-R^2 to 0.701 and lowers RMSE to 0.164. Combining adaptive attacks (bottom) and robust evaluation (right) produces the strongest held-out fit: R^2=0.740, RMSE =0.156 (bottom-right).

The original privacy–robustness tradeoff was demonstrated only on 3 generic defenses and theorized not to apply to recent gradient-blocking ones (Torp et al., 2026). Here under adaptive attack and robust evaluation, it generalizes to the tested recent defenses and to both generic training techniques, without adding any of them to the fitting set. Thus, the privacy–robustness tradeoff generalizes when measured by attacks and evaluators to which defenses are not overfit. We also emphasize that these substantial improvements to the fit were made with a small set of simple attack adaptations and a single more robust evaluation variant. We hypothesize that as adaptive attacks improve to better approximate the optimal reconstructions, the privacy–robustness fit will improve accordingly.

5 Discussion

The improved out-of-sample fit suggests a generalizable empirical relationship between adversarial robustness and privacy leakage that applies more broadly than previously theorized. This poses a key question: why should robustness predict privacy?

From Robustness to Gradient Alignment. Recent work on input gradient geometry suggests a possible explanation. Srinivas et al. (2023) argue that input gradients $\nabla_x T(x)$ are *perceptually aligned* when they lie on the signal manifold containing the task-relevant components of the data distribution. They further show that for Bayes-optimal classifiers, such on-manifold alignment is equivalent to off-manifold robustness, and show empirically that off-manifold robustness is prevalent across several forms of robust training. In this way, prior work connects robustness and perceptual

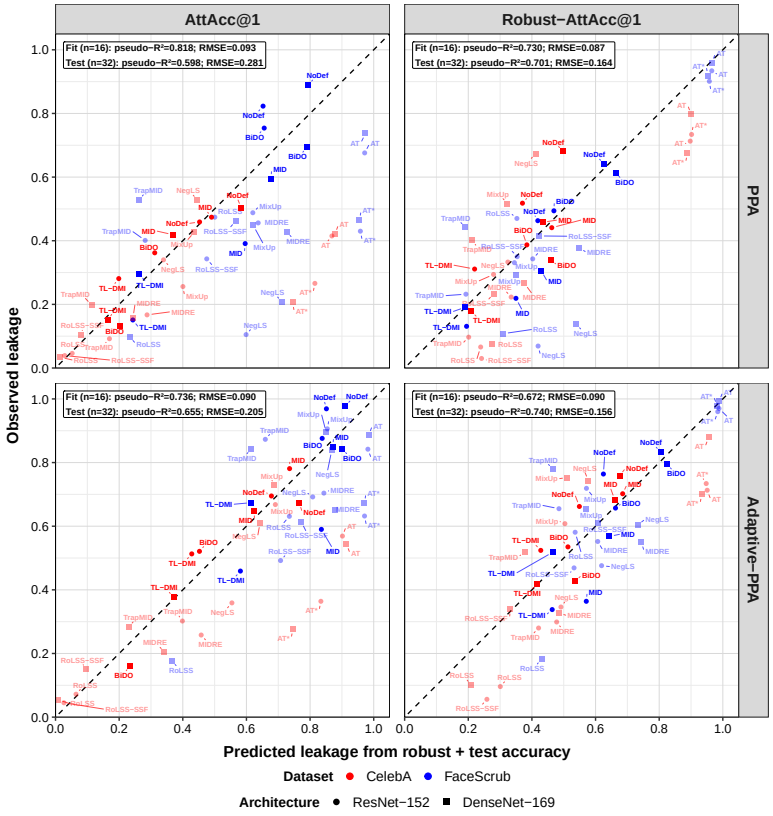


Figure 3: The privacy–robustness relationship generalizes substantially better under adaptive attacks (bottom row) and a Robust Inception-v3 evaluator (right column). Opaque points were used for fitting (the same set of defenses that were used to fit in the Torp et al. (2026) regression).

alignment in both directions: robust models tend to exhibit perceptually aligned gradients (Srinivas et al., 2023; Kaur et al., 2019), while directly promoting perceptual alignment can itself increase robustness (Ganz et al., 2023). Separately, Peng et al. (2025) show that generative MIAs implicitly project inversion-time loss gradients onto the tangent space of the generator manifold, filtering out components that the generator cannot follow. They empirically show that models become more vulnerable to inversion when their loss gradients are better aligned with that manifold.

These are different notions of alignment. The signal manifold is defined by the task and data distribution; the generator manifold is defined by the attacker's auxiliary generator. However, when the generator captures the relevant semantic structure of the private distribution, the two results suggest a natural connection. Robust models concentrate more of their predictive gradients on semantically meaningful directions. If those directions are also represented by the generator, then it follows that more of the target model's gradient is available to a generator-constrained attack. In contrast, non-robust features remain generally predictive, but are visually imperceptible and need not provide equally useful directions for reconstructing recognizable identity. This provides one geometric explanation for why adversarial robustness predicts reconstruction leakage.

Peng et al. (2025) quantify their mechanism using an inversion-time alignment score; Fig. 5 regresses this alignment score with AttAcc@1 across the same 48 target models shown in Fig. 3. We find the expected positive relationship with leakage, but considerably weaker predictive performance than adversarial robustness: the alignment regression obtains pseudo-$R^2 = 0.244$ on the fitting set and 0.085 across the test set. This reflects an important distinction: Alignment is measured along a particular attack trajectory and therefore depends on the generator, loss, current reconstruction, and attack optimization. In contrast, robustness is measured independently of the inversion attack. In our experiments, it provides a stronger attack-agnostic proxy for reconstruction vulnerability, whereas alignment helps explain the mechanism by which inversion reconstructs data.

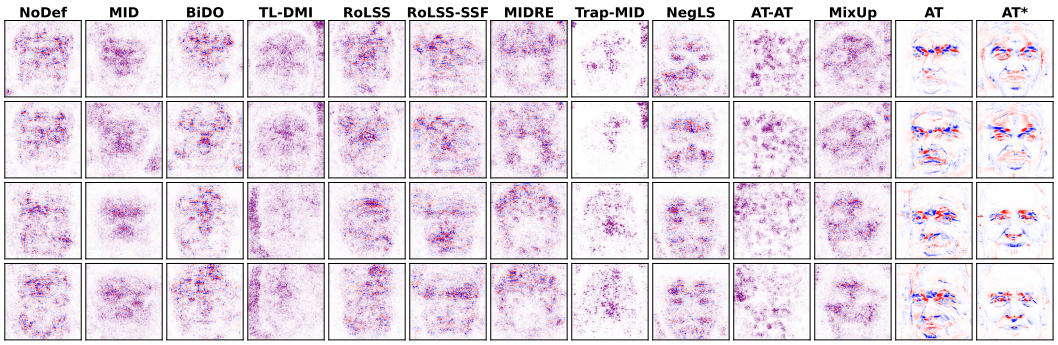


Figure 4: Heatmaps of inversion-time loss gradients $\nabla_x \mathcal{L}_{\text{cls}}(x)$, where $x = G(z)$ is a GAN-generated reconstruction at various PPA iterations on ResNet-152. Gradient magnitudes are *normalized*, i.e., incomparable across target models. Rows 1 & 2 correspond to FaceScrub class 12 at attack iterations 9 and 49, respectively. Rows 3 & 4 correspond to the same iterations of class 16.

Semantic Structure in Inversion Gradients. Interestingly, the same relationship is visible in the inversion gradients themselves. Fig. 4 plots normalized vanilla-PPA gradients across the target models in our experiments, as done in Peng et al. (2025), holding the attack, generator, architecture, and dataset fixed. The plots therefore visualize the gradient structure exposed by each target model under the same inversion attack. Across the target models, greater adversarial robustness generally corresponds to more recognizable structure in the inversion gradients. The adversarially trained models provide the clearest examples: across both identities and attack iterations, their gradients trace the eyes, nose, mouth, and contours of the face. This illustration is consistent with prior work showing that robust classifiers produce perceptually aligned gradients (e.g., Tsipras et al., 2019; Kaur et al., 2019; Ganz et al., 2023; Srinivas et al., 2023).

More surprisingly, low leakage under vanilla PPA does not imply the absence of semantically structured gradients. NegLS provides the clearest example: vanilla PPA obtains AttAcc@1 of 0.105, yet its gradients visibly trace the eyes, nose, and mouth across all four panels; our adaptive attack raises leakage to 0.692. TL-DMI and RoLSS-SSF exhibit weaker but still recognizable facial structure, and their leakage likewise increases under adaptive PPA. Trap-MID illustrates the complementary case. Its vanilla gradients are sparse and substantially less facial, yet adaptive PPA raises AttAcc@1 from 0.401 to 0.873. The heatmaps therefore cannot by themselves measure reconstruction vulnerability. They describe the gradients encountered by one attack trajectory. A fixed attack may fail despite receiving semantically useful gradients, or may follow a trajectory that never exposes the gradients available under another loss. In either case, low measured leakage under one attack does not establish that the recoverable information has been removed.

6 Conclusion

In this paper, we show that SOTA privacy defenses, generic training techniques intended to improve generalization, and undefended models all expose their training data more, and in some cases several fold more, than reported in standard MIA evaluations. Simple changes to the attack loss and learning rate recover substantially more leakage, while scoring the same reconstructions under an alternative evaluator feature basis can materially change measured leakage. This is a familiar failure mode from adversarial robustness: defenses can appear effective because they overfit to the standard evaluation configuration, rather than because they mitigate the underlying vulnerability. These underestimated leakage rates also concealed a more general relationship; once we adapt the attack and vary the evaluator, the privacy–robustness relationship generalizes substantially better to held-out defenses and standard training regimes that were excluded from the original fit. Adversarial robustness therefore provides an attack-agnostic proxy for reconstruction vulnerability in this setting. It remains an open question whether any practical defense can beat this privacy–robustness curve by reducing training-data exposure without paying the corresponding cost in adversarial robustness.

AI USE STATEMENT

We used generative AI tools to assist with proofreading and polishing both writeup and code. However, we did not use generative AI for research ideation or to originate the hypotheses studied in this paper. All AI-assisted writing was reviewed by the authors, and all AI-assisted code changes were manually reviewed and tested. The authors take responsibility for the final text, code, results, and claims in this work.

ETHICS STATEMENT

This work develops adaptive model inversion attacks. The corresponding ethical risk is that stronger attacks can also improve an adversary's ability to extract information about private training data. However, the primary purpose of our work is defensive evaluation: we show that current best-practice MIA evaluations can substantially underestimate the degree to which trained models expose recognizable information about their training data, including biometric information such as facial geometry. We also characterize a tradeoff between this privacy leakage and another Trustworthy-ML objective, adversarial robustness. Better measurement of both vulnerabilities and their tradeoffs provides a stronger basis for designing and evaluating privacy defenses. It also provides practitioners with a better understanding of the risks to which they expose the humans represented in their datasets.

Nonetheless, we deliberately restrict our experiments to widely studied public datasets, public model architectures, and experimental settings already common in recent MIA research. Our experiments reproduce standard high-resolution MIA settings and modify the attack and evaluation procedures used to measure leakage.

REPRODUCIBILITY STATEMENT

The experiments in this paper were designed from the outset to be replicated. We provide full (anonymized) replication code for the attacks, evaluations, regressions, and experiments reported in this paper: Anonymous Github

Wherever possible, we use the original authors' implementations and parameters for MIAs and MIA defenses; any modifications we make, including our adaptive attack losses, learning-rate schedules, and alternative evaluation procedure, are specified explicitly in Sec. 2. App. B provides the regression specifications and coefficients, and App. F reports the full experimental results across attacks, defenses, datasets, architectures, and evaluation metrics. Our code is also organized so that the adaptive attacks and evaluation methods introduced here can be applied directly to additional MIA defenses.

REFERENCES

Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. MIRROR: Model Inversion for Deep Learning Network with High Fidelity. In *Proceedings 2022 Network and Distributed System Security Symposium*, San Diego, CA, USA, 2022. Internet Society. doi: 10.14722/ndss.2022.24335. URL https://www.ndss-symposium.org/wp-content/uploads/2022-335-paper.pdf.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square Attack: a query-efficient black-box adversarial attack via random search, July 2020. URL http://arxiv.org/abs/1912.00049. arXiv:1912.00049 [cs].

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Sebastian Bordt, Uddeshya Upadhyay, Zeynep Akata, and Ulrike von Luxburg. The manifold hypothesis for gradient-based explanations. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3697–3702. IEEE, 2023.

Si Chen, Mostafa Kahla, Ruoxi Jia, and Guo-Jun Qi. Knowledge-Enriched Distributional Model Inversion Attacks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, October 2021. IEEE. doi: 10.1109/iccv48922.2021.01587. URL https://ieeexplore.ieee.org/document/9711082/.

Francesco Croce and Matthias Hein. Minimally distorted Adversarial Examples with a Fast Adaptive Boundary Attack. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2196–2205. PMLR, November 2020a. URL https://proceedings.mlr.press/v119/croce20a.html. ISSN: 2640-3498.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2206–2216. PMLR, November 2020b. URL https://proceedings.mlr.press/v119/croce20b.html. ISSN: 2640-3498.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial Robustness as a Prior for Learned Representations, September 2019. URL http://arxiv.org/abs/1906.00945. arXiv:1906.00945 [stat].

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Silvia Ferrari and Francisco Cribari-Neto. Beta regression for modelling rates and proportions. *Journal of applied statistics*, 31(7):799–815, 2004.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, Denver Colorado USA, October 2015. ACM. ISBN 978-1-4503-3832-5. doi: 10.1145/2810103.2813677. URL https://dl.acm.org/doi/10.1145/2810103.2813677.

Roy Ganz, Bahjat Kawar, and Michael Elad. Do perceptually aligned gradients imply robustness? In *International conference on machine learning*, pp. 10628–10648. PMLR, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

Sy-Tuyen Ho, Koh Jun Hao, Keshigeyan Chandrasegaran, Ngoc-Bao Nguyen, and Ngai-Man Cheung. Model Inversion Robustness: Can Transfer Learning Help?, May 2024. URL http://arxiv.org/abs/2405.05588. arXiv:2405.05588 [cs].

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial Examples Are Not Bugs, They Are Features, August 2019. URL http://arxiv.org/abs/1905.02175. arXiv:1905.02175 [stat].

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019. URL https://arxiv.org/abs/1812.04948.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, Seattle, WA, USA, June 2020. IEEE. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00813. URL https://ieeexplore.ieee.org/document/9156570/.

Simran Kaur, Jeremy Cohen, and Zachary C Lipton. Are perceptually-aligned gradients a general property of robust classifiers? *arXiv preprint arXiv:1910.08640*, 2019.

Jun Hao Koh, Sy-Tuyen Ho, Ngoc-Bao Nguyen, and Ngai-man Cheung. On the Vulnerability of Skip Connections to Model Inversion Attacks, September 2024. URL http://arxiv.org/abs/2409.01696. arXiv:2409.01696 [cs].

Zhen-Ting Liu and Shang-Tse Chen. Trap-MID: Trapdoor-based defense against model inversion attacks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=GNhrGRCerd.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738, Santiago, Chile, December 2015. IEEE. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.425. URL http://ieeexplore.ieee.org/document/7410782/.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks, September 2017. URL http://arxiv.org/abs/1706.06083. arXiv:1706.06083 [stat].

Felipe A. Mejia, Paul Gamble, Zigfried Hampel-Arias, Michael Lomnitz, Nina Lopatina, Lucas Tindall, and Maria Alejandra Barrios. Robust or Private? Adversarial Training Makes Models More Vulnerable to Privacy Attacks, June 2019. URL http://arxiv.org/abs/1906.06449. arXiv:1906.06449 [cs].

Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *2014 IEEE International Conference on Image Processing (ICIP)*, pp. 343–347, October 2014. doi: 10.1109/ICIP.2014.7025068. URL https://ieeexplore.ieee.org/document/7025068/. ISSN: 2381-8549.

Ngoc-Bao Nguyen, Keshigeyan Chandrasegaran, Milad Abdollahzadeh, and Ngai-Man Cheung. Re-Thinking Model Inversion Attacks Against Deep Neural Networks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16384–16393, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01572. URL https://ieeexplore.ieee.org/document/10203343/.

Xiong Peng, Feng Liu, Jingfeng Zhang, Long Lan, Junjie Ye, Tongliang Liu, and Bo Han. Bilateral Dependency Optimization: Defending Against Model-inversion Attacks. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1358–1367, Washington DC USA, August 2022. ACM. ISBN 978-1-4503-9385-0. doi: 10.1145/3534678.3539376. URL https://dl.acm.org/doi/10.1145/3534678.3539376.

Xiong Peng, Bo Han, Feng Liu, Tongliang Liu, and Mingyuan Zhou. Pseudo-private data guided model inversion attacks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pyqPUf36D2.

Xiong Peng, Bo Han, Fengfei Yu, Tongliang Liu, Feng Liu, and Mingyuan Zhou. Generative model inversion through the lens of the manifold hypothesis. *Advances in Neural Information Processing Systems*, 38:70612–70651, 2025.

Yixiang Qiu, Hao Fang, Hongyao Yu, Bin Chen, MeiKang Qiu, and Shu-Tao Xia. A Closer Look at GAN Priors: Exploiting Intermediate Features for Enhanced Model Inversion Attacks, September 2024. URL http://arxiv.org/abs/2407.13863. arXiv:2407.13863.

Suraj Srinivas, Sebastian Bordt, and Himabindu Lakkaraju. Which models have perceptually-aligned gradients? an explanation via off-manifold robustness. *Advances in neural information processing systems*, 36:21172–21195, 2023.

Lukas Struppek, Dominik Hintersdorf, Antonio De Almeida Correia, Antonia Adler, and Kristian Kersting. Plug & Play Attacks: Towards Robust and Flexible Model Inversion Attacks, June 2022. URL http://arxiv.org/abs/2201.12179. arXiv:2201.12179 [cs].

Lukas Struppek, Dominik Hintersdorf, and Kristian Kersting. Be Careful What You Smooth For: Label Smoothing Can Be a Privacy Shield but Also a Catalyst for Model Inversion Attacks, July 2024. URL http://arxiv.org/abs/2310.06549. arXiv:2310.06549 [cs].

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.308. URL http://ieeexplore.ieee.org/document/7780677/.

Rasmus Torp, Shailen Smith, and Adam Breuer. Reducing information dependency does not cause training data privacy. Adversarially non-robust features do. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=BnEG8pn3pK.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in neural information processing systems*, 33:1633–1645, 2020.

Viet-Hung Tran, Ngoc-Bao Nguyen, Son T. Mai, Hans Vandierendonck, Ira Assent, Alex Kot, and Ngai-Man Cheung. Random erasing vs. model inversion: A promising defense or a false hope? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=S9CwKnPHaO. Featured Certification.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Tianhao Wang, Yuheng Zhang, and Ruoxi Jia. Improving Robustness to Model Inversion Attacks via Mutual Information Regularization, September 2020. URL http://arxiv.org/abs/2009.05241. arXiv:2009.05241 [cs].

Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=r1Ddp1-Rb.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks, April 2020. URL http://arxiv.org/abs/1911.07135. arXiv:1911.07135 [cs].

A Deferred Details of Experiments

A.1 Datasets

- **FaceScrub (Ng & Winkler, 2014).** FaceScrub contains 530 celebrity identities and is widely used in model inversion research (Struppek et al., 2024; Koh et al., 2024; Ho et al., 2024; Qiu et al., 2024; Struppek et al., 2022). Following prior work, we use a commonly used Kaggle-hosted version of the dataset[1].
- **CelebA (Liu et al., 2015).** CelebA contains 202,599 images spanning 10,177 celebrity identities. Following prior model inversion work (Struppek et al., 2024; Liu & Chen, 2024; Qiu et al., 2024; Struppek et al., 2022), we retain the 1,000 identities with the most images, yielding 30,038 images. We additionally process the images using the CelebA HD Cropper[2].

A.2 Model Architectures

We evaluate two architectures commonly used in recent high-resolution model inversion work:

- **ResNet-152 (He et al., 2015).** ResNet-152 is widely used in prior MIA studies (Struppek et al., 2022; Qiu et al., 2024; Koh et al., 2024; Struppek et al., 2024) and serves as our primary architecture.
- **DenseNet-169 (Huang et al., 2017).** We additionally evaluate DenseNet-169, following prior MIA work (Struppek et al., 2022; Qiu et al., 2024; Koh et al., 2024; Liu & Chen, 2024).

A.3 Target Models

To reduce the computational and environmental impact of this paper, we reached out to authors of Torp et al. (2026) for access to their target model weights. We note here which ones we were able to use and which we have trained from scratch in this paper:

- **No Defense (NoDef).** We use a vanilla ResNet-152 and DenseNet-169 trained without any model inversion defense as the baseline.
- **Mutual Information Defense (MID) (Wang et al., 2020).** The MID defense utilizes a regularizer that has the goal of reducing training-data-to-model information dependency. It does so by penalizing an approximation of the mutual information of training data and intermediate feature representations. We get the trained weights from Torp et al. (2026).
- **Bilateral Dependency Optimization (BiDO) (Peng et al., 2022).** BiDO uses two regularizers: one reduces a kernel-based dependency measure between inputs and intermediate representations to limit information flow from the training data, while the other increases dependency between intermediate representations and outputs to preserve predictive performance. We get the trained weights from Torp et al. (2026).
- **Transfer Learning Defense against Model Inversion (TL-DMI) (Ho et al., 2024).** Most defenses train the target model by fully fine-tuning a publicly pre-trained model on the private train set. In contrast, TL-DMI fine-tunes only the deeper layers, with the goal of reducing the number of parameters that encode sensitive information from the private dataset. We train it for ResNet-152 FaceScrub, freezing the first 6 layers. For all other model architecture/defense combinations we get the weights from Torp et al. (2026).
- **Negative Label Smoothing (NegLS) (Struppek et al., 2024).** NegLS applies negative label smoothing to make the target model systematically overconfident in its predictions. Rather than removing training-data information from the model, this deliberate overconfidence disrupts the confidence-score landscape used by PPA and related gradient-based GAN attacks, such as IF-GMI, making the gradients used for reconstruction less effective.

[1] https://www.kaggle.com/datasets/rajnishe/facescrub-full
[2] https://github.com/LynnHo/HD-CelebA-Cropper

- **Removal of Last Stage Skip-Connection (RoLSS) (Koh et al., 2024).** RoLSS removes the skip connection in the final stage of the target model to disrupt gradient flow during the optimization of PPA and related attacks. We get the trained weights from Torp et al. (2026).
- **Skip-Connection Scaling Factor (RoLSS-SSF) (Koh et al., 2024).** RoLSS-SSF reduces, rather than removes, the contribution of the final-stage skip connection, aiming to impede gradient flow for inversion attacks while retaining more of the original model architecture and predictive performance. We get the trained weights from Torp et al. (2026).
- **Trapdoor-based Defense (Trap-MID) (Liu & Chen, 2024).** Trap-MID embeds a label-specific trapdoor into the target model, causing inputs containing the corresponding trigger to be classified as that label. This trapdoor acts as a shortcut for model inversion attacks, steering reconstructions toward the trigger pattern rather than private training data. We use the models from Torp et al. (2026), which reports two hyperparameter variants for each dataset–architecture pair, and retain only the more private variant for each dataset: $(\alpha = 0.007, \beta = 0.5)$ for FaceScrub and $(\alpha = 0.02, \beta = 0.2)$ for CelebA.
- **Model Inversion Defense via Random Erasing (MIDRE) (Tran et al., 2025).** For the MIDRE defense, the authors argue for a target model trained on randomly partially erased images, limiting its exposure to complete objects. This aims to create a mismatch between private-data features and MI reconstructions. We train the MIDRE defense with $\alpha = 0.4$, which denotes the range of random erasing. The implementation uses `Torchvision RandomErasing` in the same manner as the original authors.

Although originally presented by Torp et al. (2026) as a causal mechanism to demonstrate a privacy-robustness tradeoff, and not as a model inversion defense, we also evaluate AT-AT in our model inversion attack experiments.

- **Anti-Adversarial Training (AT-AT) (Torp et al., 2026).** The AT-AT training regime has the opposite goal of ordinary adversarial training, that is to be as non-robust as possible under even very small magnitude perturbations, while still retaining good generalization on clean data. This is achieved by balancing a loss of vanilla cross-entropy on the label and cross-entropy of a targeted PGD attack with respect to a random target class.

We additionally consider these commonly used training setups that we find to greatly affect the standard model inversion attack setups:

- **Adversarial Training (AT) (Madry et al., 2017).** AT trains the target model on adversarially perturbed inputs, encouraging robustness within a bounded perturbation region and reducing reliance on non-robust features. AT and AT* use $\epsilon = 0.01$ and 0.031 on FaceScrub, and $\epsilon = 0.0025$ and 0.01 on CelebA, respectively; AT* always denotes the higher ϵ. All AT models are trained with 10-step PGD under an ℓ_∞ perturbation constraint.
- **MixUp (Zhang et al., 2018).** Mixup trains the target model on convex combinations of pairs of training examples and their labels, encouraging smoother behavior between samples and reducing memorization. It is a widely used data augmentation technique for its generalization improvements. We use Mixup with $\alpha = 0.4$.

A.4 Target Model Training

Unless otherwise specified, we follow the target-model training setup of Struppek et al. (2024; 2022). Models are initialized from ImageNet-pretrained Torchvision weights and trained for 100 epochs using Adam with $\beta = (0.9, 0.999)$, batch size 128, and an initial learning rate of 0.001. We use a MultiStepLR scheduler with $\gamma = 0.1$ and retain the checkpoint with the lowest validation loss. Final prediction accuracy is reported on the test split.

A.5 Evaluation Model Training

- **Vanilla evaluation model.** Following Struppek et al. (2022), we use independently trained Inception-V3 models for evaluation. The CelebA Inception model is hot-started from the FaceScrub Inception model to improve accuracy, as is standard (Struppek et al., 2022).

The FaceScrub and CelebA vanilla Inception models achieve test accuracies of 96.9% and 91.2%, respectively.

- **Robust evaluation model.** We additionally train an adversarially robust evaluation model using 10-step PGD adversarial training with an ℓ_∞ perturbation budget of $\epsilon = 0.031$ for both FaceScrub and CelebA. The FaceScrub and CelebA robust Inception models achieve test accuracies of 93.3% and 83.4%, respectively.

A.6 Model Inversion Attacks

- **Plug-and-Play Attack (PPA) (Struppek et al., 2022).** PPA performs inversion in three stages: latent sampling, latent optimization, and final sample selection. It uses a StyleGAN-2 (Karras et al., 2020) generator pre-trained on FFHQ (Karras et al., 2019), which has no identity overlap with FaceScrub or CelebA. We evaluate PPA on 100 target classes with a batch size of 40. Latent codes are optimized for 50 steps, and otherwise we follow the standard hyperparameter setup of Struppek et al. (2022; 2024).
- **Intermediate Features enhanced Generative Model Inversion (IF-GMI) (Qiu et al., 2024).** IF-GMI builds on GAN-based inversion by additionally optimizing intermediate StyleGAN-2 features, increasing the flexibility of the reconstructed images. We use the same FFHQ-pretrained StyleGAN-2 prior and evaluate the attack on 100 target classes. We otherwise follow the hyperparameter settings of the original implementation.

A.6.1 Adversarial Attacks

- **AutoAttack (Croce & Hein, 2020b).** Following (Torp et al., 2026), we measure adversarial robustness of target models using the AutoAttack ensemble of 4 attacks (APGD-CE, APGD-T, FAB-T, and Square) meant to provide a thorough, diverse signal of robustness with minimal hyperparameter tuning (Croce & Hein, 2020b;a; Andriushchenko et al., 2020). We use an l_∞ norm and measure across various ϵ levels $\{0.031, 0.0025, 0.0005, \text{1e-5}\}$. We exclude $\epsilon = 0.031$ from the regression as it is less informative in these experimental settings.

A.7 Alignment Score

In order to quantify the notion of inversion-time loss gradient alignment, Peng et al. (2025) propose an *alignment score* measure, defined as the cosine of the angle ϕ between a loss gradient $\nabla_x \mathcal{L}_{\text{cls}}$ and its projection onto the tangent space $T_x\mathcal{M}_{\text{aux}}$ of the GAN manifold at point x: $\cos(\phi) = \|\mathbf{P}_\mathbf{x}\nabla_\mathbf{x}\mathcal{L}_{\text{cls}}\| \;/\; \|\nabla_\mathbf{x}\mathcal{L}_{\text{cls}}\|$, where P_x is an orthogonal projection matrix onto the tangent space $T_x\mathcal{M}_{\text{aux}}$. Higher alignment scores are hypothesized to correspond to increased MIA vulnerability.

One alignment score can be calculated per sample per attack iteration. Since alignment scores are very computationally costly to calculate, we must subsample. We compute on every tenth iteration (6 iterations total), a subset of 15 classes, and just 1 sample per 40-size batch. With 5 batches per class, this leads to 450 alignment scores per attack; we report the mean of these scores in the tables and regressions.

A.8 Compute Setup

We run most target model training and attacks on NVIDIA Blackwell Pro 6000 Max-Q, NVIDIA A100, and NVIDIA H100 GPUs. Default target model training and attack hyperparameters are set up to fit on a single 80 GB GPU.

B Deferred Regression Details

We use the same Beta-regression specification as Torp et al. (2026), using the mean–precision parameterization of Ferrari & Cribari-Neto (2004). For target-model condition j, let $Y_j \in (0, 1)$ denote

Table 3: Beta regression (logit link) coefficients predicting leakage from test and AutoAttack robust accuracy, with pseudo-R^2 and RMSE on the fitted and held-out models.

	PPA		Adaptive PPA	
Term	**AttAcc@1**	**Robust-AttAcc@1**	**AttAcc@1**	**Robust-AttAcc@1**
Intercept	-5.836	-0.970	-7.303	-1.917
TestAcc	6.422	-0.208	11.005	3.142
AutoAcc.0025	3.434	4.271	2.736	4.662
AutoAcc.0005	1.608	1.060	1.132	0.429
AutoAcc1e-5	-1.222	-0.316	-2.602	-1.229
Fit ($n = 16$) pseudo-R^2	0.818	0.730	0.736	0.672
Fit ($n = 16$) RMSE	0.093	0.087	0.090	0.090
Test ($n = 32$) pseudo-R^2	0.598	0.701	0.655	0.740
Test ($n = 32$) RMSE	0.281	0.164	0.205	0.156

measured reconstruction leakage, $\mathrm{TestAcc}_j$ denote clean test accuracy, and $\mathrm{AutoAcc}_{\epsilon,j}$ denote AutoAttack robust accuracy at perturbation magnitude ϵ. We estimate

$$Y_j \sim \mathrm{Beta}(\mu_j\phi, (1-\mu_j)\phi),$$
$$\mathrm{logit}(\mu_j) = \beta_0 + \beta_1 \mathrm{TestAcc}_j + \beta_2 \mathrm{AutoAcc}_{0.0025,j} + \beta_3 \mathrm{AutoAcc}_{0.0005,j} + \beta_4 \mathrm{AutoAcc}_{10^{-5},j}. \quad (2)$$

As in Torp et al. (2026), clean test accuracy controls for the known relationship between model accuracy and MIA leakage, while the three AutoAttack terms measure robustness to small adversarial perturbations. We omit $\epsilon = 0.031$, at which robust accuracy is near zero for nearly all non-adversarially-trained target models.

For the main regressions in Fig. 3, we fit Eq. 2 only on NoDef, MID, BiDO, and TL-DMI across FaceScrub and CelebA and ResNet-152 and DenseNet-169 ($n = 16$). NegLS, Trap-MID, MIDRE, RoLSS, RoLSS-SSF, MixUp, AT, and AT* are held out entirely ($n = 32$). AT-AT is excluded because it is a causal instrument deliberately constructed to manipulate the non-robust feature basis captured by the robustness predictors (we also re-run the regression with AT-AT included, as described below).

We estimate four regressions, changing only the leakage response Y_j: AttAcc@1 under vanilla PPA; Robust-AttAcc@1 on the same vanilla-PPA reconstructions; the maximum AttAcc@1 across PPA, N-LR, and C-LR; and the maximum Robust-AttAcc@1 across those attacks. The two maxima are computed independently. Predictors, fitting observations, and held-out observations are otherwise identical across all four regressions. Table 3 reports the fitted coefficients and fit statistics.

Three additional regression figures use related specifications. Fig. 1 applies the published coefficients of Torp et al. (2026), fit on their $n = 18$ observations, to the $n = 32$ held-out observations in this paper without refitting. Fig. 5 replaces the robustness predictors with the vanilla-PPA alignment score as the predictor of AttAcc@1, while retaining the same $n = 16$ fitting and $n = 32$ held-out observations. Finally, Fig. 6 refits Eq. 2 on all 52 target-model conditions, including AT-AT, with no held-out set.

Table 3 reports β coefficients for the four main body regressions (Fig. 3).

C Additional Related Work

C.1 Adaptive Attacks under MIA

While white-box model inversion has lacked a consistent, rigorous approach to adaptive attacks, some recent defense papers do make an effort to handle them.

Struppek et al. (2024) reports how NegLS performs against various changes in MIA loss function, learning rate, and initial latent vector sampling. Interestingly, while a raised learning rate and cross-entropy loss are components of our *C-LR* adaptive attack, Struppek et al. (2024) did not test these two changes in combination, and so was unable to spot the increased MIA vulnerability we report for NegLS. This is an indication that even seemingly-thorough efforts to search the space of adaptive attacks can occasionally fall short.

Liu & Chen (2024) explore specific adaptive attacks where the adversary has different levels of familiarity with (1) the existence of trapdoors and (2) the specific trapdoor signatures.

Tran et al. (2025) explore one type of adaptive attack where adversaries are aware of the sections of images that are randomly being erased during optimization, which proved not to be effective against MIDRE as a defense.

Overall, adaptive attack coverage in MIA defense work remains limited, raising the risk that defense success is over-estimated.

C.2 The Manifold Hypothesis and MIA

The manifold hypothesis (Fefferman et al., 2016; Bordt et al., 2023) argues that real-world high-dimensional data distributions p tend to lie on a relatively low-dimensional manifold $\mathcal{M}$. In the setting of model inversion, we can say that the private dataset $\mathcal{D}_{\text{pri}}$ has a corresponding low-dimensional manifold $\mathcal{M}_{\text{pri}}$, and the attack GAN G ranges over a manifold $\mathcal{M}_{\text{aux}}$, hopefully similar to $\mathcal{M}_{\text{pri}}$ (Peng et al., 2025). $\mathcal{M}_{\text{pri}}$ can be called the *data manifold* and $\mathcal{M}_{\text{aux}}$ the *generator manifold* (Peng et al., 2025).

D Deferred Alignment Results

Figure 5 reports 15-class alignment scores regressed against MIA AttAcc@1.

E Deferred Regression Results

Figure 6 showcases AutoAttack regressions against AttAcc@1 and Robust AttAcc@1 fit over all 52 training regimes, including AT-AT (Torp et al., 2026).

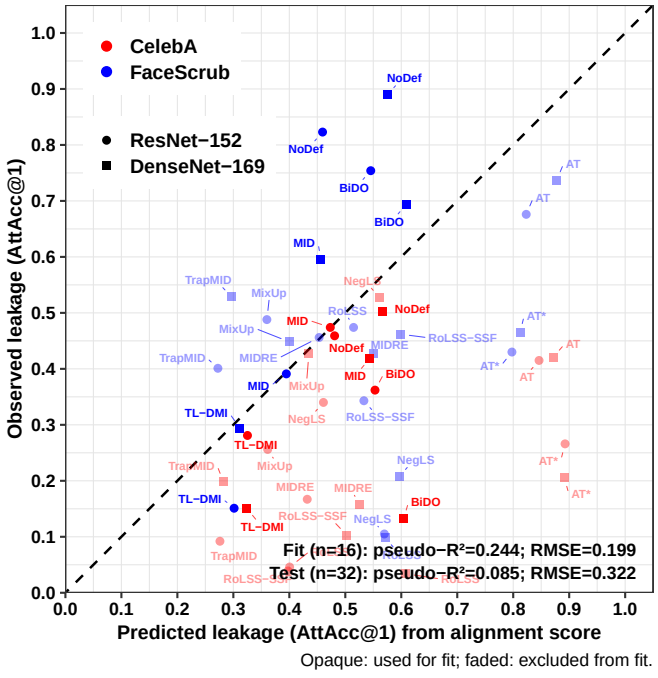
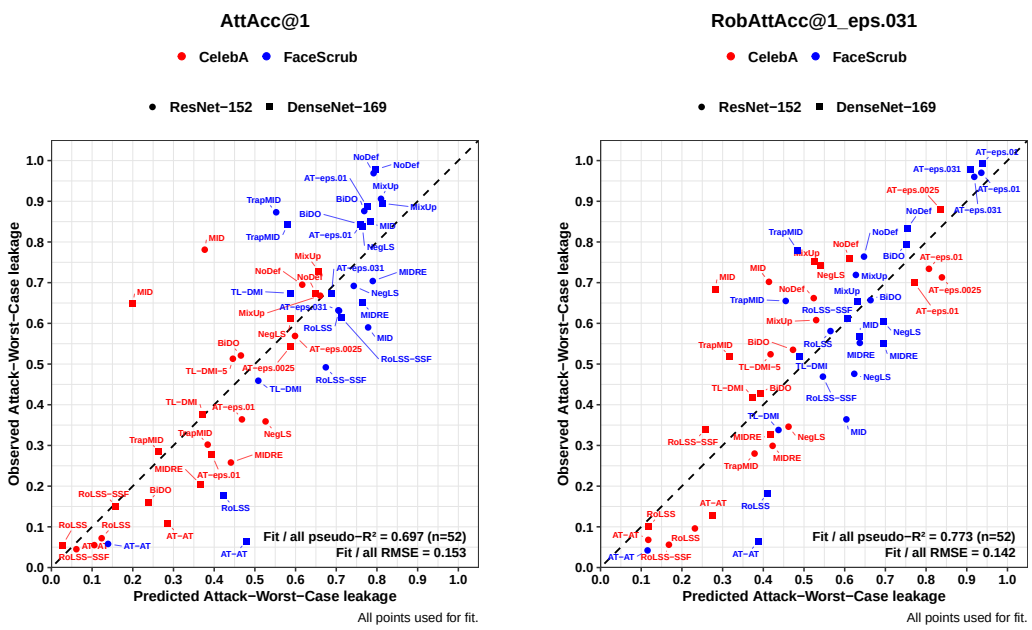


Figure 5: Regression predicting leakage from the alignment score. The fitted alignment coefficient is positive (6.16), consistent with greater alignment being associated with greater leakage, although held-out predictive fit is weak.



(a) AttAcc@1

(b) Robust-AttAcc@1

Figure 6: Regressions fit on all data points using the worst-case adaptive attack.

F Deferred Model Inversion Results

Table 4 reports mean 15-class alignment scores across (non-adaptive) PPA runs.

Table 5 reports an initial attempt applying *N-LR* and *C-LR* adaptive attacks to the IF-GMI attack (Qiu et al., 2024) on ResNet-152 models trained on FaceScrub. While these specific adaptive attacks are less effective overall, we still see a significant leakage increase on the NegLS defense. We hypothesize that further improvement may be obtained by (for example) trying multiple learning rate increases, rather than fixing the increase at an arbitrary 10x, though we leave such optimizations to future work.

Table 6 shows that robust evaluation models still reveal significant additional leakage for AT models under IF-GMI. **Table 7** details full IF-GMI adaptive attack results.

Table 8 reports AutoAttack worst-case robustness measurements across four epsilon levels; three epsilon values were used in all of the robustness regressions (0.0025, 0.0005, 1e-5), where $\epsilon = 0.031$ is shown to indicate its lack of descriptiveness in the MIA setting.

Tables 9 and 10 showcase complete *N-LR* and *C-LR* adaptive attack results on PPA.

Table 4: Alignment scores of the target models calculated with vanilla PPA.

	FaceScrub		CelebA	
Model	ResNet-152	DenseNet-169	ResNet-152	DenseNet-169
NoDef	0.144	0.220	0.158	0.214
MID	0.101	0.142	0.153	0.199
BiDO	0.200	0.242	0.205	0.239
TL-DMI	0.034	0.042	0.052	0.051
RoLSS	0.180	0.217	0.105	0.241
RoLSS-SSF	0.192	0.235	0.103	0.172
MIDRE	0.140	0.203	0.126	0.187
Trap-MID	0.011	0.030	0.014	0.019
NegLS	0.216	0.234	0.145	0.210
AT-AT	0.021	0.031	0.018	0.019
MixUp	0.077	0.105	0.078	0.128
AT	0.420	0.490	0.447	0.481
AT*	0.393	0.408	0.514	0.512

Table 5: ResNet-152 adaptive IF-GMI attack results on FaceScrub. Each attack metric shows standard IF-GMI $\rightarrow$ worst-case adaptive IF-GMI over IF-GMI, IF-GMI N-LR, and IF-GMI C-LR: maximum AttAcc@1 and minimum L2-Face, computed independently. Small parenthesized symbols identify the attack attaining each worst-case value (all symbols are shown for ties). TestAcc is from standard IF-GMI. AT and AT* use $\epsilon = 0.01$ and 0.031, respectively.

	FaceScrub		
Model	**TestAcc** $\uparrow$	**AttAcc@1** $\downarrow$	**L2-Face** $\uparrow$
NoDef	0.943	$0.976 \rightarrow 0.976$ (▬)	$0.724 \rightarrow 0.724$ (▬)
MID	0.950	$0.380 \rightarrow 0.380$ (▬)	$1.013 \rightarrow 1.013$ (▬)
BiDO	0.920	$0.935 \rightarrow 0.935$ (▬)	$0.797 \rightarrow 0.797$ (▬)
TL-DMI	0.891	$0.383 \rightarrow 0.383$ (▬)	$0.962 \rightarrow 0.962$ (▬)
RoLSS	0.886	$0.512 \rightarrow 0.546$ (★)	$0.908 \rightarrow 0.908$ (▬)
RoLSS-SSF	0.869	$0.340 \rightarrow 0.399$ (★)	$0.979 \rightarrow 0.953$ (★)
MIDRE	0.941	$0.774 \rightarrow 0.774$ (▬)	$0.780 \rightarrow 0.780$ (▬)
Trap-MID	0.926	$0.633 \rightarrow 0.633$ (▬)	$0.982 \rightarrow 0.982$ (▬)
NegLS	0.911	$0.087 \rightarrow 0.532$ (★)	$1.269 \rightarrow 0.884$ (★)
AT-AT	0.941	$0.071 \rightarrow 0.071$ (▬)	$1.220 \rightarrow 1.220$ (▬)
MixUp	0.971	$0.761 \rightarrow 0.761$ (▬)	$0.832 \rightarrow 0.832$ (▬)
AT	0.928	$0.899 \rightarrow 0.899$ (▬)	$0.730 \rightarrow 0.730$ (▬)
AT*	0.891	$0.677 \rightarrow 0.677$ (▬)	$0.788 \rightarrow 0.788$ (▬)

IF-GMI (baseline): ▬ IF-GMI N-LR: ★ IF-GMI C-LR: ▲

Table 6: ResNet-152 standard and robust IF-GMI attack accuracy on FaceScrub. Each pair shows AttAcc@k $\rightarrow$ Robust-AttAcc@k for $k \in \{1, 5\}$ from the same standard IF-GMI run. Bold values mark changes >0.1.

	FaceScrub		
Model	**TestAcc** $\uparrow$	**AttAcc@1** $\downarrow$	**AttAcc@5** $\downarrow$
NoDef	0.943	**0.976** $\rightarrow$ **0.754**	0.998 $\rightarrow$ 0.922
MID	0.950	**0.380** $\rightarrow$ **0.192**	**0.654** $\rightarrow$ **0.437**
BiDO	0.920	**0.935** $\rightarrow$ **0.696**	0.988 $\rightarrow$ 0.889
TL-DMI	0.891	**0.383** $\rightarrow$ **0.264**	**0.667** $\rightarrow$ **0.522**
RoLSS	0.886	0.512 $\rightarrow$ 0.492	0.770 $\rightarrow$ 0.784
RoLSS-SSF	0.869	0.340 $\rightarrow$ 0.339	0.628 $\rightarrow$ 0.644
MIDRE	0.941	**0.774** $\rightarrow$ **0.574**	**0.949** $\rightarrow$ **0.820**
Trap-MID	0.926	**0.633** $\rightarrow$ **0.427**	**0.762** $\rightarrow$ **0.643**
NegLS	0.911	0.087 $\rightarrow$ 0.053	0.248 $\rightarrow$ 0.177
AT-AT	0.941	0.071 $\rightarrow$ 0.049	0.215 $\rightarrow$ 0.156
MixUp	0.971	**0.761** $\rightarrow$ **0.545**	**0.870** $\rightarrow$ **0.733**
AT	0.928	0.899 $\rightarrow$ 0.986	0.985 $\rightarrow$ 0.999
AT*	0.891	**0.677** $\rightarrow$ **0.975**	**0.880** $\rightarrow$ **0.996**

Table 7: Full adaptive IF-GMI attack results on **ResNet-152** trained on FaceScrub.

Data	**Model**	**Attack**	**TestAcc** $\uparrow$	**AttAcc@1** $\downarrow$	**Robust-AttAcc@1** $\downarrow$	**AttAcc@5** $\downarrow$	**Robust-AttAcc@5** $\downarrow$	**L2-Face** $\uparrow$
FaceScrub	NoDef	IF-GMI	0.943	0.976	0.754	0.998	0.922	0.724
		IF-GMI N-LR	0.943	0.908	0.665	0.988	0.895	0.799
		IF-GMI C-LR	0.943	0.781	0.450	0.944	0.737	0.888
	MID	IF-GMI	0.950	0.380	0.192	0.654	0.437	1.013
		IF-GMI N-LR	0.950	0.294	0.172	0.519	0.373	1.105
		IF-GMI C-LR	0.950	0.299	0.180	0.509	0.360	1.107
	BiDO	IF-GMI	0.920	0.935	0.696	0.988	0.889	0.797
		IF-GMI N-LR	0.920	0.797	0.568	0.955	0.844	0.892
		IF-GMI C-LR	0.920	0.457	0.252	0.758	0.534	1.010
	TL-DMI	IF-GMI	0.891	0.383	0.264	0.667	0.522	0.962
		IF-GMI N-LR	0.891	0.290	0.214	0.552	0.450	1.024
		IF-GMI C-LR	0.891	0.231	0.169	0.474	0.391	1.051
	RoLSS	IF-GMI	0.886	0.512	0.492	0.770	0.784	0.908
		IF-GMI N-LR	0.886	0.546	0.486	0.831	0.807	0.920
		IF-GMI C-LR	0.886	0.494	0.446	0.744	0.734	0.935
	RoLSS-SSF	IF-GMI	0.869	0.340	0.339	0.628	0.644	0.979
		IF-GMI N-LR	0.869	0.399	0.366	0.696	0.697	0.953
		IF-GMI C-LR	0.869	0.350	0.358	0.619	0.644	0.988
	MIDRE	IF-GMI	0.941	0.774	0.574	0.949	0.820	0.780
		IF-GMI N-LR	0.941	0.600	0.438	0.868	0.737	0.870
		IF-GMI C-LR	0.941	0.308	0.239	0.614	0.490	0.973
	Trap-MID	IF-GMI	0.926	0.633	0.427	0.762	0.643	0.982
		IF-GMI N-LR	0.926	0.594	0.408	0.746	0.617	1.002
		IF-GMI C-LR	0.926	0.568	0.390	0.723	0.599	1.021
	NegLS	IF-GMI	0.911	0.087	0.053	0.248	0.177	1.269
		IF-GMI N-LR	0.911	0.532	0.342	0.840	0.663	0.884
		IF-GMI C-LR	0.911	0.297	0.177	0.606	0.427	1.006
	AT-AT	IF-GMI	0.941	0.071	0.049	0.215	0.156	1.220
		IF-GMI N-LR	0.941	0.025	0.019	0.088	0.070	1.367
		IF-GMI C-LR	0.941	0.023	0.017	0.083	0.068	1.369
	MixUp	IF-GMI	0.971	0.761	0.545	0.870	0.733	0.832
		IF-GMI N-LR	0.971	0.720	0.525	0.849	0.718	0.887
		IF-GMI C-LR	0.971	0.736	0.547	0.858	0.730	0.874
	AT	IF-GMI	0.928	0.899	0.986	0.985	0.999	0.730
		IF-GMI N-LR	0.928	0.779	0.923	0.950	0.996	0.858
		IF-GMI C-LR	0.928	0.658	0.944	0.872	0.991	0.874
	AT*	IF-GMI	0.891	0.677	0.975	0.880	0.996	0.788
		IF-GMI N-LR	0.891	0.563	0.910	0.830	0.987	0.915
		IF-GMI C-LR	0.891	0.496	0.932	0.740	0.987	0.934

Table 8: Test accuracy and AutoAttack robust accuracy of the target models.

Data	Arch.	Model	TestAcc ↑	AutoAcc.031	AutoAcc.0025	AutoAcc.0005	AutoAcc1e-5
FaceScrub	ResNet-152	NoDef	0.943	0.000	0.054	0.853	0.939
		MID	0.950	0.000	0.006	0.770	0.928
		BiDO	0.920	0.000	0.103	0.841	0.927
		TL-DMI	0.891	0.000	0.000	0.000	0.836
		RoLSS	0.886	0.000	0.009	0.744	0.882
		RoLSS-SSF	0.869	0.000	0.013	0.721	0.857
		MIDRE	0.941	0.000	0.039	0.850	0.937
		Trap-MID	0.926	0.000	0.000	0.000	0.856
		NegLS	0.911	0.000	0.068	0.784	0.906
		AT-AT	0.941	0.000	0.000	0.000	0.119
		MixUp	0.971	0.000	0.000	0.789	0.964
		AT	0.928	0.002	0.891	0.920	0.926
		AT*	0.891	0.128	0.848	0.877	0.884
	DenseNet-169	NoDef	0.947	0.000	0.244	0.888	0.947
		MID	0.949	0.000	0.059	0.867	0.912
		BiDO	0.921	0.000	0.284	0.864	0.913
		TL-DMI	0.921	0.000	0.000	0.005	0.910
		RoLSS	0.732	0.000	0.002	0.514	0.727
		RoLSS-SSF	0.886	0.000	0.070	0.780	0.884
		MIDRE	0.929	0.000	0.177	0.842	0.912
		Trap-MID	0.920	0.000	0.000	0.000	0.904
		NegLS	0.924	0.000	0.170	0.836	0.919
		AT-AT	0.933	0.000	0.000	0.000	0.744
		MixUp	0.967	0.000	0.000	0.812	0.968
		AT	0.943	0.000	0.891	0.928	0.937
		AT*	0.892	0.036	0.827	0.855	0.861
CelebA	ResNet-152	NoDef	0.845	0.000	0.036	0.678	0.813
		MID	0.802	0.004	0.139	0.542	0.574
		BiDO	0.747	0.000	0.062	0.590	0.748
		TL-DMI	0.814	0.000	0.000	0.119	0.799
		RoLSS	0.537	0.000	0.000	0.086	0.530
		RoLSS-SSF	0.428	0.000	0.000	0.042	0.406
		MIDRE	0.756	0.000	0.021	0.545	0.715
		Trap-MID	0.806	0.000	0.000	0.000	0.760
		NegLS	0.805	0.000	0.013	0.559	0.771
		AT-AT	0.819	0.000	0.000	0.000	0.155
		MixUp	0.899	0.000	0.000	0.463	0.888
		AT	0.818	0.000	0.638	0.797	0.819
		AT*	0.748	0.001	0.658	0.725	0.744
	DenseNet-169	NoDef	0.856	0.000	0.142	0.750	0.833
		MID	0.730	0.011	0.129	0.405	0.396
		BiDO	0.611	0.000	0.143	0.484	0.594
		TL-DMI	0.788	0.000	0.000	0.044	0.748
		RoLSS	0.275	0.000	0.001	0.122	0.267
		RoLSS-SSF	0.557	0.000	0.001	0.289	0.538
		MIDRE	0.696	0.000	0.061	0.535	0.685
		Trap-MID	0.711	0.000	0.000	0.021	0.673
		NegLS	0.820	0.000	0.077	0.670	0.808
		AT-AT	0.836	0.000	0.000	0.000	0.567
		MixUp	0.871	0.000	0.000	0.642	0.854
		AT	0.827	0.000	0.651	0.774	0.802
		AT*	0.701	0.001	0.629	0.689	0.705

Table 9: Full adaptive PPA attack results on **ResNet-152** trained on FaceScrub and CelebA.

Data	Model	Attack	TestAcc ↑	AttAcc@1 ↓	Robust-AttAcc@1 ↓	AttAcc@5 ↓	Robust-AttAcc@5 ↓	L2-Face ↑
FaceScrub	NoDef	PPA	0.943	0.823	0.463	0.961	0.761	0.806
		PPA N-LR	0.943	0.969	0.764	0.999	0.945	0.712
		PPA C-LR	0.943	0.861	0.578	0.979	0.829	0.805
	MID	PPA	0.950	0.391	0.219	0.699	0.489	0.950
		PPA N-LR	0.950	0.563	0.358	0.828	0.657	0.878
		PPA C-LR	0.950	0.590	0.364	0.844	0.657	0.867
	BiDO	PPA	0.920	0.754	0.494	0.937	0.781	0.832
		PPA N-LR	0.920	0.876	0.657	0.989	0.913	0.797
		PPA C-LR	0.920	0.575	0.366	0.848	0.668	0.925
	TL-DMI	PPA	0.891	0.151	0.131	0.384	0.318	1.088
		PPA N-LR	0.891	0.459	0.338	0.750	0.630	0.954
		PPA C-LR	0.891	0.367	0.289	0.665	0.574	0.986
	RoLSS	PPA	0.886	0.474	0.470	0.760	0.777	0.858
		PPA N-LR	0.886	0.631	0.581	0.894	0.880	0.832
		PPA C-LR	0.886	0.584	0.560	0.836	0.850	0.850
	RoLSS-SSF	PPA	0.869	0.343	0.350	0.635	0.653	0.918
		PPA N-LR	0.869	0.492	0.469	0.786	0.788	0.894
		PPA C-LR	0.869	0.442	0.439	0.704	0.737	0.910
	MIDRE	PPA	0.941	0.456	0.343	0.761	0.624	0.891
		PPA N-LR	0.941	0.704	0.552	0.928	0.829	0.815
		PPA C-LR	0.941	0.421	0.333	0.725	0.621	0.923
	Trap-MID	PPA	0.926	0.401	0.232	0.676	0.481	0.980
		PPA N-LR	0.926	0.873	0.655	0.970	0.891	0.777
		PPA C-LR	0.926	0.851	0.626	0.967	0.869	0.800
	NegLS	PPA	0.911	0.105	0.069	0.277	0.200	1.217
		PPA N-LR	0.911	0.692	0.476	0.945	0.806	0.796
		PPA C-LR	0.911	0.467	0.291	0.797	0.608	0.907
	AT-AT	PPA	0.941	0.058	0.042	0.190	0.145	1.225
		PPA N-LR	0.941	0.049	0.037	0.163	0.134	1.264
		PPA C-LR	0.941	0.046	0.036	0.168	0.128	1.279
	MixUp	PPA	0.971	0.488	0.331	0.727	0.588	0.932
		PPA N-LR	0.971	0.871	0.681	0.957	0.857	0.758
		PPA C-LR	0.971	0.906	0.719	0.970	0.883	0.725
	AT	PPA	0.928	0.676	0.934	0.896	0.995	0.758
		PPA N-LR	0.928	0.842	0.965	0.972	0.998	0.766
		PPA C-LR	0.928	0.717	0.970	0.908	0.998	0.789
	AT*	PPA	0.891	0.430	0.901	0.715	0.982	0.849
		PPA N-LR	0.891	0.632	0.942	0.879	0.997	0.829
		PPA C-LR	0.891	0.545	0.960	0.797	0.994	0.858
CelebA	NoDef	PPA	0.845	0.459	0.518	0.750	0.786	0.852
		PPA N-LR	0.845	0.695	0.662	0.912	0.897	0.838
		PPA C-LR	0.845	0.459	0.483	0.718	0.746	0.892
	MID	PPA	0.802	0.474	0.441	0.725	0.688	0.821
		PPA N-LR	0.802	0.781	0.702	0.939	0.891	0.709
		PPA C-LR	0.802	0.761	0.670	0.935	0.880	0.714
	BiDO	PPA	0.747	0.362	0.387	0.629	0.653	0.923
		PPA N-LR	0.747	0.521	0.535	0.784	0.796	0.876
		PPA C-LR	0.747	0.294	0.335	0.565	0.613	0.982
	TL-DMI	PPA	0.814	0.281	0.311	0.550	0.591	0.929
		PPA N-LR	0.814	0.513	0.524	0.786	0.815	0.863
		PPA C-LR	0.814	0.338	0.402	0.617	0.667	0.930
	RoLSS	PPA	0.537	0.046	0.066	0.145	0.182	1.147
		PPA N-LR	0.537	0.072	0.095	0.192	0.231	1.124
		PPA C-LR	0.537	0.066	0.096	0.190	0.242	1.117
	RoLSS-SSF	PPA	0.428	0.039	0.030	0.118	0.097	1.183
		PPA N-LR	0.428	0.043	0.054	0.136	0.155	1.187
		PPA C-LR	0.428	0.045	0.056	0.147	0.154	1.186
	MIDRE	PPA	0.756	0.167	0.223	0.377	0.461	0.987
		PPA N-LR	0.756	0.258	0.299	0.514	0.581	0.951
		PPA C-LR	0.756	0.165	0.215	0.383	0.462	1.019
	Trap-MID	PPA	0.806	0.092	0.097	0.193	0.195	1.358
		PPA N-LR	0.806	0.302	0.280	0.482	0.465	1.150
		PPA C-LR	0.806	0.245	0.246	0.418	0.415	1.165
	NegLS	PPA	0.805	0.340	0.333	0.623	0.639	0.908
		PPA N-LR	0.805	0.359	0.346	0.686	0.675	0.897
		PPA C-LR	0.805	0.250	0.262	0.538	0.553	0.935
	AT-AT	PPA	0.819	0.049	0.059	0.138	0.151	1.245
		PPA N-LR	0.819	0.055	0.068	0.149	0.177	1.261
		PPA C-LR	0.819	0.048	0.064	0.142	0.178	1.263
	MixUp	PPA	0.899	0.256	0.294	0.483	0.511	0.965
		PPA N-LR	0.899	0.629	0.550	0.834	0.779	0.822
		PPA C-LR	0.899	0.668	0.608	0.850	0.817	0.792
	AT	PPA	0.818	0.415	0.713	0.693	0.912	0.822
		PPA N-LR	0.818	0.569	0.671	0.795	0.903	0.863
		PPA C-LR	0.818	0.416	0.700	0.685	0.886	0.866
	AT*	PPA	0.748	0.266	0.734	0.527	0.929	0.887
		PPA N-LR	0.748	0.364	0.703	0.626	0.927	0.911
		PPA C-LR	0.748	0.271	0.721	0.507	0.907	0.937

Table 10: Full adaptive PPA attack results on **DenseNet-169** trained on FaceScrub and CelebA.

Data	Model	Attack	TestAcc ↑	AttAcc@1 ↓	Robust-AttAcc@1 ↓	AttAcc@5 ↓	Robust-AttAcc@5 ↓	L2-Face ↑
FaceScrub	NoDef	PPA	0.947	0.890	0.641	0.980	0.889	0.763
		PPA N-LR	0.947	0.978	0.833	0.999	0.974	0.703
		PPA C-LR	0.947	0.828	0.589	0.964	0.860	0.835
	MID	PPA	0.949	0.595	0.306	0.835	0.575	0.869
		PPA N-LR	0.949	0.826	0.552	0.967	0.825	0.755
		PPA C-LR	0.949	0.850	0.568	0.970	0.837	0.743
	BiDO	PPA	0.921	0.693	0.614	0.891	0.866	0.831
		PPA N-LR	0.921	0.843	0.795	0.967	0.949	0.788
		PPA C-LR	0.921	0.643	0.614	0.874	0.861	0.874
	TL-DMI	PPA	0.921	0.294	0.191	0.575	0.443	0.996
		PPA N-LR	0.921	0.674	0.518	0.911	0.806	0.851
		PPA C-LR	0.921	0.484	0.359	0.764	0.651	0.928
	RoLSS	PPA	0.732	0.098	0.108	0.267	0.267	1.116
		PPA N-LR	0.732	0.177	0.182	0.424	0.423	1.050
		PPA C-LR	0.732	0.169	0.180	0.412	0.408	1.058
	RoLSS-SSF	PPA	0.886	0.462	0.415	0.761	0.738	0.882
		PPA N-LR	0.886	0.614	0.611	0.851	0.883	0.838
		PPA C-LR	0.886	0.598	0.606	0.840	0.862	0.851
	MIDRE	PPA	0.929	0.428	0.377	0.734	0.674	0.899
		PPA N-LR	0.929	0.651	0.550	0.906	0.836	0.839
		PPA C-LR	0.929	0.339	0.330	0.642	0.621	0.965
	Trap-MID	PPA	0.920	0.529	0.442	0.787	0.720	0.893
		PPA N-LR	0.920	0.843	0.779	0.959	0.949	0.772
		PPA C-LR	0.920	0.711	0.678	0.910	0.888	0.842
	NegLS	PPA	0.924	0.208	0.138	0.450	0.323	1.107
		PPA N-LR	0.924	0.839	0.604	0.971	0.895	0.712
		PPA C-LR	0.924	0.478	0.323	0.798	0.652	0.899
	AT-AT	PPA	0.933	0.058	0.063	0.190	0.164	1.194
		PPA N-LR	0.933	0.056	0.057	0.203	0.176	1.240
		PPA C-LR	0.933	0.065	0.060	0.205	0.173	1.236
	MixUp	PPA	0.967	0.448	0.293	0.704	0.525	0.960
		PPA N-LR	0.967	0.865	0.601	0.958	0.790	0.801
		PPA C-LR	0.967	0.895	0.654	0.965	0.824	0.768
	AT	PPA	0.943	0.737	0.959	0.920	0.996	0.734
		PPA N-LR	0.943	0.886	0.992	0.983	0.999	0.709
		PPA C-LR	0.943	0.581	0.907	0.818	0.985	0.857
	AT*	PPA	0.892	0.465	0.919	0.734	0.986	0.839
		PPA N-LR	0.892	0.673	0.954	0.897	0.997	0.785
		PPA C-LR	0.892	0.640	0.978	0.861	0.998	0.803
CelebA	NoDef	PPA	0.856	0.503	0.681	0.788	0.897	0.800
		PPA N-LR	0.856	0.674	0.759	0.886	0.936	0.810
		PPA C-LR	0.856	0.474	0.616	0.741	0.840	0.874
	MID	PPA	0.730	0.419	0.458	0.675	0.718	0.836
		PPA N-LR	0.730	0.648	0.683	0.853	0.872	0.753
		PPA C-LR	0.730	0.610	0.636	0.833	0.854	0.764
	BiDO	PPA	0.611	0.132	0.339	0.313	0.607	1.021
		PPA N-LR	0.611	0.161	0.427	0.374	0.674	1.025
		PPA C-LR	0.611	0.123	0.356	0.300	0.598	1.064
	TL-DMI	PPA	0.788	0.151	0.178	0.361	0.402	1.000
		PPA N-LR	0.788	0.377	0.419	0.654	0.703	0.904
		PPA C-LR	0.788	0.279	0.329	0.524	0.591	0.957
	RoLSS	PPA	0.275	0.034	0.077	0.126	0.195	1.149
		PPA N-LR	0.275	0.048	0.101	0.166	0.242	1.143
		PPA C-LR	0.275	0.055	0.100	0.155	0.234	1.150
	RoLSS-SSF	PPA	0.557	0.103	0.234	0.285	0.482	1.007
		PPA N-LR	0.557	0.137	0.311	0.352	0.591	0.996
		PPA C-LR	0.557	0.150	0.340	0.358	0.590	1.008
	MIDRE	PPA	0.696	0.158	0.266	0.351	0.510	0.990
		PPA N-LR	0.696	0.204	0.327	0.429	0.602	0.994
		PPA C-LR	0.696	0.160	0.273	0.352	0.508	1.020
	Trap-MID	PPA	0.711	0.198	0.401	0.443	0.661	0.957
		PPA N-LR	0.711	0.273	0.518	0.545	0.769	0.949
		PPA C-LR	0.711	0.284	0.514	0.532	0.757	0.946
	NegLS	PPA	0.820	0.528	0.674	0.800	0.916	0.781
		PPA N-LR	0.820	0.611	0.743	0.877	0.933	0.733
		PPA C-LR	0.820	0.297	0.396	0.604	0.703	0.882
	AT-AT	PPA	0.836	0.086	0.091	0.205	0.220	1.177
		PPA N-LR	0.836	0.109	0.128	0.255	0.275	1.151
		PPA C-LR	0.836	0.108	0.128	0.264	0.285	1.152
	MixUp	PPA	0.871	0.428	0.517	0.690	0.748	0.866
		PPA N-LR	0.871	0.695	0.694	0.873	0.870	0.790
		PPA C-LR	0.871	0.728	0.753	0.883	0.895	0.762
	AT	PPA	0.827	0.421	0.797	0.690	0.949	0.816
		PPA N-LR	0.827	0.544	0.879	0.809	0.975	0.792
		PPA C-LR	0.827	0.386	0.749	0.647	0.915	0.879
	AT*	PPA	0.701	0.206	0.675	0.425	0.879	0.932
		PPA N-LR	0.701	0.277	0.700	0.508	0.887	0.919
		PPA C-LR	0.701	0.183	0.639	0.389	0.847	0.987