PRIVACY RISKS AND MEMORIZATION OF SPURIOUS CORRELATED DATA

Chenxiang Zhang, Jun Pang & Sjouke Mauw
University of Luxembourg
chenxiang.zhang@uni.lu

ABSTRACT

Neural networks are vulnerable to privacy attacks aimed at stealing sensitive data. The risks are amplified in real-world scenario when models are trained on limited and biased data. In this work, we investigate the impact of spurious correlation bias on privacy vulnerability. We introduce *spurious privacy leakage*, a phenomenon where spurious groups are more vulnerable to privacy attacks compared to other groups. Through empirical analysis, we counterintuitively demonstrate that reducing spurious correlation fails to address the privacy disparity between groups. This leads us to introduce a new perspective on privacy disparity based on data memorization. We show that mitigating spurious correlation does not reduce the degree of data memorization, and therefore, neither the privacy risks. Our findings highlight the need to rethink privacy with spurious learning.

1 INTRODUCTION

Neural networks are applied across diverse domains such as face recognition, medical prognosis, or personalized advertisement. All these applications are trained on user-sensitive data that can be of interest to attackers (Shokri et al., 2017; Liu et al., 2021a; Mireshghallah et al., 2020; Yeom et al., 2018). Additionally, real-world collected data are limited and often biased towards specific groups, a subset of the dataset sharing a common characteristic (e.g. gender, ethnicity, or geographic location). On top of the privacy concerns, models trained on real-world data can also inherit biases, causing failures at test time (Sagawa et al., 2019; Geirhos et al., 2020; Shah et al., 2020). Therefore, models deployed in sensitive domains should satisfy multiple constraints, such as ensuring fair performance across groups and protection of sensitive data.

In this work, we focus on the *spurious correlation* bias, a statistical relationship between two variables that appears to be causal but is either caused by a third confounding variable or random chance. Spurious correlation has been widely studied in machine learning (Sagawa et al., 2019; Izmailov et al., 2022; Yang et al., 2023) with the objective to improve the worst-group performance, however, its privacy side-effect has been overlooked. On the other hand, privacy research typically focuses on unbiased datasets such as CIFAR10 or CIFAR100 (Krizhevsky, 2009; Hu et al., 2022), overlooking the privacy risks of sensitive applications that use limited and biased real-world datasets. We address this gap by investigating the privacy of neural networks trained on spurious correlated real-world datasets using *membership inference attacks* (MIA), a family of privacy attacks commonly used for their simplicity and versatility (Murakonda & Shokri, 2020; Carlini et al., 2021).

Contributions. We observe a phenomenon we term *spurious privacy leakage*, where groups with spurious correlation are significantly more vulnerable to MIA than other groups (Section 3.1). This phenomenon adds the fairness requirement to the existing privacy challenges, modeling a realistic scenario for data-sensitive applications. For example, privacy auditing may naively conclude that a model satisfies the privacy requirements by evaluating on an aggregated *average* metric over the entire dataset. However, spurious correlation can cause one group to be significantly more vulnerable than others, violating the requirements for that specific group. Studying the group privacy disparity is important to bridge and advance the privacy and spurious correlation research, to understand the risks of the model, and to improve the auditing process.

We further investigate the consequences of *spurious privacy leakage*. Previous works suggested that improving the generalization across groups can mitigate privacy disparity (Kulynych et al., 2022).

However, to the best of our knowledge, there is no evidence to support the claim. To study this, we use robust training methods (Sagawa et al., 2019; Kirichenko et al., 2022) to mitigate spurious correlations and re-evaluate the privacy vulnerability. Surprisingly, even after the mitigation, we observe no consistent privacy improvement (Section 4). This result leads us to introduce a new perspective on group privacy disparity based on memorization (Zhang et al., 2021): spurious robust training improves wort-class performance but do not reduce the memorization level of data compared to standard training (Figure 3), and therefore neither it can mitigate the privacy risks. We release the code at `https://anonymous.4open.science/r/spurious-mia-6676`.

2 Background & Related work

We provide a concise introduction needed to follow the rest of the work including neural networks, membership inference attacks, and spurious correlation.

Neural networks represent functions $f_{\boldsymbol{\theta}}\colon \mathcal{X} \to \mathcal{Y}$ that map the input data $\boldsymbol{x} \in \mathcal{X}$ to a label $\boldsymbol{y} \in \mathcal{Y}$. The dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$ is a set of labeled pairs used for estimating the model parameters. The neural network is parametrized by $\boldsymbol{\theta} \in \mathbb{R}^n$ and it is updated using a first-order optimizer to minimize a loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. We focus on the classification setting where the cross-entropy loss is commonly used. Formally, the objective is the *empirical risk minimization* (ERM) (Vapnik, 1991):

$$\hat{\boldsymbol{\theta}}_{\text{ERM}} = arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}}(\ell(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x}))$$

Spurious correlation is a statistical relationship between two variables X and Y that first appears to be causal but in reality is either caused by a third confounding (e.g. spurious) variable Z or due to random chance. This relationship is in contrast with causality, where the change of the variable X leads to a direct and predictable outcome of Y while ruling out the presence of any confounding factors Z. For a given dataset with spurious correlation, a feature $\boldsymbol{z}$ is called spurious if it is correlated with the target label $\boldsymbol{y}$ in the training data but not in the test data. For example, in a binary bird classification dataset where waterbirds mainly appear on a water background, a biased model can exploit the background spurious feature instead of the bird invariant feature, leading to a wrong prediction when the input is a waterbird on a land background (Sagawa et al., 2019). Ideally, we would like to suppress the bias coming from the spurious features, which can be expressed as $\Pr(\boldsymbol{y} \mid \boldsymbol{x}) = \Pr(\boldsymbol{y} \mid \boldsymbol{x}_{\text{inv}}, \boldsymbol{z}) = \Pr(\boldsymbol{y} \mid \boldsymbol{x}_{\text{inv}})$ where we decomposed the input $\boldsymbol{x}$ as a combination of invariant features $\boldsymbol{x}_{\text{inv}}$ and spurious features $\boldsymbol{z}$. Sagawa et al. (2019) proposed the group *distributionally robust optimization* (DRO) to mitigate spurious features. DRO minimizes the worst-group loss, while ERM which minimizes the average loss:

$$\hat{\boldsymbol{\theta}}_{\text{DRO}} = arg\min_{\boldsymbol{\theta}} \max_{\boldsymbol{g}\in\mathcal{G}} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y},\boldsymbol{g})\in\mathcal{D}}[\ell(\boldsymbol{y}, f_{\boldsymbol{\theta}}(\boldsymbol{x}))]$$

where the dataset is divided into $\boldsymbol{g}$ groups. The new dataset is $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i, \boldsymbol{g}_i)\}$ where $\boldsymbol{g} \in \mathcal{G}$ is a discrete-valued label (e.g. all the combinations of birds and backgrounds or geographical area information). DRO is considered an oracle method due to its explicit use of the group information for the training (Liu et al., 2021b). Additional methods in the literature suppress the spurious features by learning and assigning a different weight per sample (Liu et al., 2021b; Nam et al., 2020), by retraining the classifier head at the end of the training (Kirichenko et al., 2022; Izmailov et al., 2022; Kang et al., 2019), by group sampling (Yang et al., 2024; Idrissi et al., 2022), or using contrastive methods (Zhang et al., 2022).

Membership inference attacks (MIA) aim to determine whether a specific input data was used during the model training. MIA is usually used to audit a model's privacy level thanks to its simplicity (Murakonda & Shokri, 2020) and versatility for creating a more complex attack (Carlini et al., 2021). The membership inference problem can be defined as learning a function $\mathcal{A}\colon \mathcal{X} \to [0, 1]$, where $\mathcal{A}$ is the attacker model that takes input $\boldsymbol{x} \in \mathcal{X}$ and outputs 1 if $\boldsymbol{x}$ was used during the model training. We assume the black-box (Shokri et al., 2017) access to the target model, where the only target information accessible is the output probability vector $\boldsymbol{p}$. Shokri et al. (2017) introduced the first MIA for neural networks with black-box access, where several *shadow* models are trained to mimic the behavior of the *target* model. More advanced attacks have been developed based on the idea of shadow models (Yeom et al., 2018; Liu et al., 2022; Carlini et al., 2022; Ye et al., 2022; Sablayrolles et al., 2019; Watson et al., 2021; Long et al., 2020). In this work, we focus on the state-of-the-art LiRA

method (Carlini et al., 2022). Given an input $\boldsymbol{x}$, LiRA predicts its membership by training N shadow models, each on a different subset of the dataset. Half of the models are named INs and contain $\boldsymbol{x}$ and the other half named OUTs do not. Each shadow model IN outputs a confidence score $\phi(\boldsymbol{p}_{\text{shadow}})$ which is used to estimate the parameters of a Gaussian $\mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}})$, and in the same way, OUTs are used to estimate $\mathcal{N}(\mu_{\text{out}}, \sigma_{\text{out}})$. Finally, the result of the attack is defined as a likelihood-ratio test:

$$\Lambda = \frac{\Pr(\phi(\boldsymbol{p}_{\text{target}}) \mid \mathcal{N}(\mu_{\text{in}}, \sigma_{\text{in}}))}{\Pr(\phi(\boldsymbol{p}_{\text{target}}) \mid \mathcal{N}(\mu_{\text{out}}, \sigma_{\text{out}}))}$$

where $\phi(\boldsymbol{p}_{\text{target}}) = log(\boldsymbol{p}/(1-\boldsymbol{p}))$ is the confidence score obtained by querying the target model with $\boldsymbol{x}$. The score Λ is used by the attacker to determine how likely it is that the given $\boldsymbol{x}$ is a member.

Privacy and safety. The intersection of privacy and ML safety topics has been extensively studied. Wang et al. (2020) focused on how pruning can mitigate privacy attacks and Shokri et al. (2021) explored the connection between privacy and explainability. Song et al. (2019) found that adversarial training can increase privacy leakage, but Li et al. (2024) reported contradictory findings when using a better evaluation (Carlini et al., 2022). In our work, we investigate the privacy risk of real-world spurious correlated datasets, which is related to ensuring fairness across groups. Prior works on privacy-fairness reported that subpopulations can exhibit varying levels of privacy risk (Truex et al., 2019; Tian et al., 2024; Zhong et al., 2022; Kulynych et al., 2022). Tian et al. (2024) showed that applying fairness methods can mildly mitigate MIA risks under an *average* metrics. Instead, we use a *per-group* analysis, revealing the privacy disparity between different groups. Kulynych et al. (2022) and Zhong et al. (2022) also investigated privacy disparity on synthetic and tabular datasets. They hypothesize that group fairness improvements can be an effective mitigation methods. In Section 4, we show that this approach does not address privacy disparities in real-world spurious datasets. Similar to our setting, Yang et al. (2022) found that synthetic spurious datasets MNIST/CIFAR have privacy disparity measured with an average privacy metric, which is suboptimal (Carlini et al., 2022). We extend their results on real-world spurious datasets, showing that for certain datasets, only state-of-the-art evalution with TPR at low FPR can reveal the privacy disparity, highlighting the importance of re-evaluating prior works. Lastly, while it is known that out-of-distribution samples have higher vulnerability (Carlini et al., 2022), real-world spurious correlated samples are by definition in-distribution and have been overlooked. We provide a comprehensive set of results to resolve the conflicts in the literature.

3 Spurious Correlation and Privacy Risks

We demonstrate the differences in privacy leakage between spurious and non-spurious correlated groups. Our results show that auditing the privacy level on the whole dataset is misleading in the presence of spurious correlations (Carlini et al., 2022; Feldman & Zhang, 2020) where the spurious groups can have significantly higher privacy leakage.

3.1 Spurious Privacy Leakage

Spurious correlations are characterized by the presence of spurious features. Assuming we have the labels of the spurious features, learning with spurious correlation is equivalent to learning with an imbalanced dataset. We refer to spurious groups as the minority groups with the worst performance (e.g. worst-group accuracy) compared to the majority groups.

Experiment setup. We select the datasets that are used by the spurious correlation community (Yang et al., 2023): Waterbirds (Sagawa et al., 2019), CelebA (Liu et al., 2014), FMoW (Koh et al., 2021), MultiNLI (Williams et al., 2017), and CivilComments (Koh et al., 2021). These datasets contain real-world spurious correlations, diverse modalities, and different target complexity (see Appendix A for details). Moreover, to the best of our knowledge, we are the first to study MIA attacks on subgroups of these realistic datasets. We use the pretrained ResNet50 (He et al., 2016) on ImageNet1k and finetune using random crop and horizontal flip. For text datasets, BERT's bert-base-uncased model (Devlin et al., 2019) is used. We perform hyperparameter optimization for each dataset using a grid search over learning rate (lr), weight decay (wd), and epochs. The grid search and its best hyperparameters are in Appendix B. We report the training and test accuracy to evaluate the performance and their difference to quantify the overfitting level. We do the same for the worst-group accuracy (WGA),

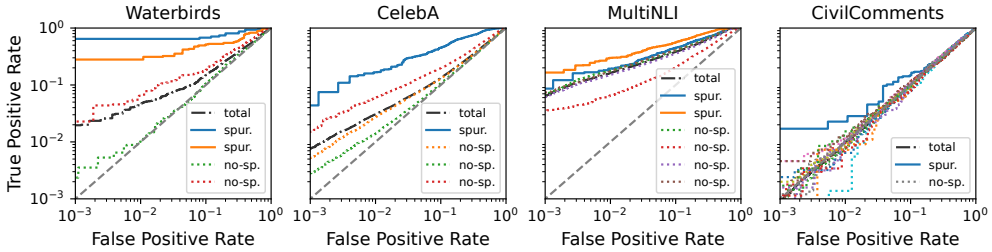


Figure 1: Attack success rate divided per group on Waterbirds, CelebA, MultiNLI, and CivilComments respectively. Across the datasets, there is a spurious group (solid lines) with consistent higher privacy leakage compared to non-spurious groups under the LiRA attack.

which is a commonly used proxy metric to measure the mitigation success of spurious features (Sagawa et al., 2019). For privacy evaluation, we follow the guidelines from Carlini et al. (2022). We train non-overfit models and report the full log-scale ROC curves, the true positive rate (TPR) at a low false positive rate (FPR) region, and also the AUROC curve for completeness. We train 32 shadow models for Waterbirds/CelebA and 16 for FMoW/MultiNLI/CivilComments.

Across all the spurious correlated datasets, the group performance disparity is consistently present, with the spurious groups having the lowest accuracy among all the groups (see Table 1). For example, in Waterbirds, ERM has a test *average* accuracy of 81.08% while one of the spurious group only 34.41%. Beyond these performance disparity, we show that spurious correlations also cause privacy issues. Using the state-of-the-art MIA method LiRA (Carlini et al., 2022), we analyze the privacy leakage of each group of the five spurious correlated datasets. For each dataset, we train the shadow models using 50% of the sampled training data as in the LiRA algorithm. We ensure that the sampled subset maintains a similar group proportion as the original dataset by first sampling per group, and then combining all the sampled groups together. Figure 1 shows that across the datasets, there exists a spurious group that exhibits higher privacy leakage than non-spurious groups. The largest disparity is observed at $\approx 3\%$ FPR area of Waterbirds, where the samples in the most *spurious group are* ≈ 10 *times more vulnerable than samples in the non-spurious group.* At $\approx 0.1\%$ FPR of CelebA, we continue to observe a significant disparity, with the most spurious group being ≈ 10 times more vulnerable than the least spurious group. In both the text datasets MultiNLI and CivilComments, the disparity is persists with ≈ 4 times difference between the most and least vulnerable groups (see Table 2 for the exact TPR at low FPR). The existence of spurious privacy leakage unfairly exposes some data groups, allowing an attacker to craft better targeted attacks. Prior research focused on privacy and fairness have observed the disparity between different subpopulations (Zhong et al., 2022; Kulynych et al., 2022; Tian et al., 2024). Our results complement their findings by analyzing real-world spurious correlated data, exposing the vulnerability of spurious groups. Surprisingly, we do not observe spurious privacy leakage in FMoW, which we investigate in Appendix B.

Finding I. *Spurious privacy leakage is present in real-world datasets, where spurious groups can have disproportionately higher vulnerability to privacy attacks than other groups.*

4 Privacy Risks of Spurious Robust Methods

Counterintuitively, we demonstrate that reducing the impact of spurious features does not mitigate *spurious privacy leakage*. We train models using spurious robust methods and observe that the group privacy disparity persists due to data *memorization* (Zhang et al., 2021).

Spurious correlations can be suppressed using robust training methods such as group *distributional robust optimization* (DRO) (Sagawa et al., 2019) or *deep feature reweighting* (DFR) (Kirichenko et al., 2022). Extensive benchmarks (Izmailov et al., 2022; Yang et al., 2023) report DRO and DFR as the most effective methods for mitigating spurious correlations. DRO is referred as an oracle method because it requires a group label to minimize the worst-group error in its objective function (Liu et al., 2021b), while DFR achieves the highest average worst-group accuracy compared to 17

Table 1: Robust methods DRO and DFR mitigate spurious features and improve WGA. The train-test difference on the whole dataset is misleading to detect overfitting. One should monitor the train-test difference for each group. *DRO fails to improve over ERM on FMoW and we omit it.

Data	Model	Train Acc. (↑)	Test Acc. (↑)	Diff. Acc. (↓)	Train WGA (↑)	Test WGA (↑)	Diff. WGA (↓)
Waterb.	ERM	**97.16 ± 0.11**	81.12 ± 0.35	16.0	50.18 ± 2.70	34.30 ± 1.27	15.8
	DRO	96.16 ± 0.23	**86.42 ± 0.38**	9.7	**93.73 ± 0.44**	**78.12 ± 0.84**	15.6
	DFR	92.63 ± 1.13	85.98 ± 0.60	**6.7**	85.81 ± 1.95	77.67 ± 2.13	**8.2**
CelebA	ERM	**97.12 ± 0.03**	**95.82 ± 0.06**	1.3	62.81 ± 1.82	42.67 ± 0.62	20.2
	DRO	94.47 ± 0.05	93.23 ± 0.21	**1.2**	**91.84 ± 0.36**	**86.11 ± 0.89**	5.7
	DFR	95.43 ± 0.14	90.52 ± 0.22	4.9	89.46 ± 0.36	84.00 ± 0.60	**5.4**
MultiNLI	ERM	**97.26 ± 0.04**	**80.74 ± 0.04**	16.5	**91.43 ± 0.79**	61.76 ± 0.28	29.7
	DRO	89.69 ± 0.09	78.76 ± 0.07	**10.9**	85.34 ± 0.23	**72.96 ± 0.66**	**12.4**
	DFR	96.36 ± 0.14	79.17 ± 0.06	17.2	90.84 ± 0.11	71.33 ± 0.13	19.5
CivilCom.	ERM	**97.45 ± 0.04**	**88.03 ± 0.06**	9.4	**90.41 ± 0.32**	53.26 ± 0.59	37.1
	DRO	90.00 ± 0.29	81.11 ± 0.31	8.9	81.84 ± 0.85	68.60 ± 0.47	13.2
	DFR	86.88 ± 0.17	79.38 ± 0.05	**7.5**	77.49 ± 0.69	**69.47 ± 0.26**	**8.0**
FMoW*	ERM	**91.58 ± 0.04**	**50.85 ± 0.08**	**40.7**	**90.84 ± 0.06**	31.04 ± 0.20	59.8
	DRO	-	-	-	-	-	-
	DFR	91.20 ± 0.38	48.62 ± 0.09	42.6	88.57 ± 0.55	**32.44 ± 0.34**	**56.1**

Table 2: Comparing the privacy of spurious robust methods. Despite improving the WGA, DRO and DFR do not consistently mitigate the privacy attack across datasets. *Waterbirds is evaluated at ≈3% FPR due to the limited samples. The spurious groups are highlighted.

	TPR @ 0.1% FPR (↓)		
Data	ERM	DRO	DFR
Waterb.*	3.12 ± 0.10	3.12 ± 0.11	3.13 ± 0.10
	30.91 ± 2.81	31.06 ± 2.76	33.20 ± 2.83
CelebA	0.27 ± 0.01	**0.26 ± 0.01**	**0.26 ± 0.01**
	4.61 ± 0.50	**4.56 ± 0.48**	4.77 ± 0.46
MultiNLI	5.88 ± 0.42	**5.73 ± 0.45**	5.86 ± 0.40
	8.26 ± 0.55	9.08 ± 1.37	**7.78 ± 0.57**
CivilCom.	**0.10 ± 0.02**	0.14 ± 0.03	0.12 ± 0.03
	0.44 ± 0.10	0.43 ± 0.17	**0.32 ± 0.12**
FMoW	**7.45 ± 0.27**	-	7.61 ± 0.28
	6.30 ± 1.80	-	6.42 ± 1.80

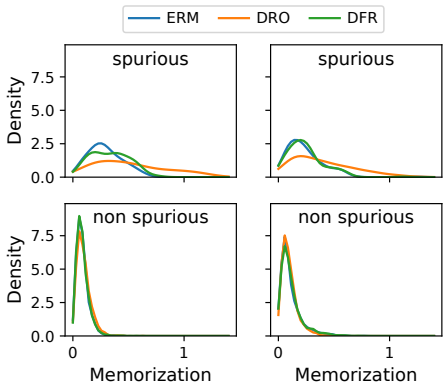


Figure 2: Memorization score per group for each method. Robust methods DRO and DFR do not mitigate the data memorization and therefore neither the privacy leakage.

spurious robust methods across 12 datasets (Yang et al., 2023). Therefore, we choose these methods for our analysis, comparing the privacy leakage of ERM, DRO, and DFR.

Experiment setup. For each dataset and training method, we train the shadow models by following the same LiRA setup as in Section 3. We ensure that models across different training methods use the same subset of data by fixing the random seeds. For privacy evaluation, we train non-overfit models by monitoring the difference between the train-val losses (Yeom et al., 2018).

Per-group overfitting. The performance results in Table 1 show the average and worst-group accuracy of robust training methods for all five datasets. We highlight that relying only on the *average* train-test accuracy difference can be misleading in detecting overfitting. When comparing two models, the first can have a lower average train-test difference but a higher train-test difference in one of the groups. For example in CelebA, the ERM method has a lower average difference than DFR (1.3% vs 4.9%) but a higher WGA difference (20.2% vs 5.4%). The same pattern can be observed for MultiNLI and FMoW. To truly avoid overfitting, we recommend assessing the performance disparity of all the groups.

Robust methods do not mitigate privacy leakage. Yeom et al. (2018) demonstrated that overfitting is a sufficient condition for MIA to succeed. Moreover, Kulynych et al. (2022) suggested that improving group performance fairness can mitigate the privacy disparity. However, to the best of our knowledge,

there is no evidence to support the claim. Therefore, *does mitigating performance disparity really mitigate spurious privacy leakage?* Firstly, we confirmed that spurious robust methods DRO and DFR significantly improve the WGA compared to ERM by mitigating the spurious correlation (see Table 1). Then, we run LiRA using ERM trained shadow models and ERM, DRO, and DFR as targets. Table 2 report the privacy attack success rate for each dataset, group, and training method. The *average* privacy at low FPR ("T" rows) is mitigated across datasets (see Appendix, Table 5). Tian et al. (2024) observed similar results using fairness methods. However, our per-group analysis reveal additional insight over the *average* analysis. The privacy leakage for spurious groups stay consistent for ERM, DRO, and DFR across datasets, indicating that the *spurious privacy leakage* issue persists despite successfully mitigating spurious correlations. Our results may be surprising, but overfitting is only a sufficient and not a necessary condition for MIA to succeed (Yeom et al., 2018). We provide another perspective on privacy disparity based on memorization.

Finding II. *Spurious robust training reduce group performance disparity but fail to address spurious privacy leakage on real-world datasets.*

Spurious privacy leakage and memorization. We provide an alternative view of the *spurious privacy leakage* phenomenon using the memorization score of data, which is also responsible for the success of MIA (Feldman, 2020; Feldman & Zhang, 2020; Carlini et al., 2022). Firstly, we demonstrate that spurious groups are more vulnerable to LiRA due to a higher memorization score compared to other groups (see Appendix B.2). Additionally, our results in Figure 2 confirm that spurious robust methods DRO and DFR do not reduce the memorization score. In particular, ERM and DFR share a similar distribution of memorization scores for both spurious and non-spurious groups. This is because DFR only retrains the last layer but the sample memorization is distributed across different layers (Feldman & Zhang, 2020; Maini et al., 2023). Therefore, DFR can hardly affect memorization and privacy despite mitigating the group fairness. DRO has a similar privacy leakage (Table 2) and memorization to ERM and DFR for non spurious groups, but even higher memorization for spurious groups. Izmailov et al. (2022) showed that DRO acts as DFR by learning not better features, but a better reweighting of a similar set of features, which can explain the similar privacy leakage of DRO, ERM, and DFR.

Finding III. *Spurious correlated data have higher memorization score than non-spurious data even after mitigating spurious correlation using state-of-the-art robust methods.*

5 Conclusion

Our findings confirm critical privacy concerns when training neural networks on real-world datasets with spurious correlations. The existence of *spurious privacy leakage* makes spurious data more vulnerable to privacy attacks than non-spurious data. Leveraging this information, an attacker can craft more powerful attacks targeting specific demographic groups. We emphasize the need to avoid aggregate metrics over the entire dataset, instead, privacy audits must include fine-grained group-level analyses to ensure performance and privacy fairness. Moreover, we show that state-of-the-art *spurious robust training mitigate spurious correlations but do not affect spurious privacy leakage*. Our results present opportunities for future research on privacy and spurious correlations focused on mitigating data memorization as a potential solution.

Impact. Our work impacts the machine learning communities concerned with bias, fairness, and security. Understanding the connection between spurious correlations and privacy is important for assessing the risks for data-sensitive domains. In particular, we suggest practitioners working in auditing to carefully assess privacy disparities.

Limitations. Our results are based on a state-of-the-art attack-based evaluation rather than analytical guarantees. While our approach is more practical and provides realistic empirical evidence, it is limited to the choice of our experiment settings.

Reproducibility. We have made our code publicly available through an anonymized repository (Section 1). Details for each experiment setup are presented in the respective sections, along with the corresponding grid search and the best hyperparameters in the appendix (see Table 3).

ACKNOWLEDGMENTS

CZ is funded by the research project R-STR-8019-00-B from the University of Luxembourg. The experiments are supported by the HPC facilities of the University of Luxembourg.

REFERENCES

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *USENIX Security Symposium*, 2021.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *Symposium on Security and Privacy*, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Symposium on Theory of Computing*, 2020.

Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems*, 2020.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys*, 2022.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In *Conference on Causal Learning and Reasoning*, 2022.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew G Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, 2022.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, 2021.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. Disparate vulnerability to membership inference attacks. *Proceedings on Privacy Enhancing Technologies*, 2022.

Xiao Li, Qiongxiu Li, Zhan Hu, and Xiaolin Hu. On the privacy effect of data enhancement via the lens of memorization. *Transactions on Information Forensics and Security*, 2024.

Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys*, 2021a.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, 2021b.

Yiyong Liu, Zhengyu Zhao, Michael Backes, and Yang Zhang. Membership inference attacks by exploiting loss trajectory. In *Conference on Computer and Communications Security*, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2014.

Yunhui Long, Lei Wang, Diyue Bu, Vincent Bindschaedler, Xiaofeng Wang, Haixu Tang, Carl A Gunter, and Kai Chen. A pragmatic approach to membership inferences on machine learning models. In *European Symposium on Security and Privacy*, 2020.

Pratyush Maini, Michael C Mozer, Hanie Sedghi, Zachary C Lipton, J Zico Kolter, and Chiyuan Zhang. Can neural network memorization be localized? In *International Conference on Machine Learning*, 2023.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Praneeth Vepakomma, Abhishek Singh, Ramesh Raskar, and Hadi Esmaeilzadeh. Privacy in deep learning: A survey. *arXiv preprint arXiv:2004.12254*, 2020.

Sasi Kumar Murakonda and Reza Shokri. Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. *arXiv preprint arXiv:2007.09339*, 2020.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *Advances in Neural Information Processing Systems*, 2020.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, 2019.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2019.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Symposium on Security and Privacy*, 2017.

Reza Shokri, Martin Strobel, and Yair Zick. On the privacy risks of model explanations. In *Conference on AI, Ethics, and Society*, 2021.

Liwei Song, Reza Shokri, and Prateek Mittal. Privacy risks of securing machine learning models against adversarial examples. In *Conference on Computer and Communications Security*, 2019.

Huan Tian, Guangsheng Zhang, Bo Liu, Tianqing Zhu, Ming Ding, and Wanlei Zhou. When fairness meets privacy: Exploring privacy threats in fair binary classifiers via membership inference attacks. In *International Joint Conference on Artificial Intelligence*, 2024.

Stacey Truex, Ling Liu, Mehmet Emre Gursoy, Wenqi Wei, and Lei Yu. Effects of differential privacy and data skewness on membership inference vulnerability. In *International Conference on Trust, Privacy and Security in Intelligent Systems and Applications*, 2019.

Vladimir Vapnik. Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems*, 1991.

Yijue Wang, Chenghong Wang, Zigeng Wang, Shanglin Zhou, Hang Liu, Jinbo Bi, Caiwen Ding, and Sanguthevar Rajasekaran. Against membership inference attack: Pruning is all you need. *arXiv preprint arXiv:2008.13578*, 2020.

Lauren Watson, Chuan Guo, Graham Cormode, and Alex Sablayrolles. On the importance of difficulty calibration in membership inference attacks. *arXiv preprint arXiv:2111.08440*, 2021.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics*, 2017.

Yao-Yuan Yang, Chi-Ning Chou, and Kamalika Chaudhuri. Understanding rare spurious correlations in neural networks. *arXiv preprint arXiv:2202.05189*, 2022.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In *International Conference on Artificial Intelligence and Statistics*, 2024.

Yuzhe Yang, Haoran Zhang, Dina Katabi, and Marzyeh Ghassemi. Change is hard: A closer look at subpopulation shift. In *International Conference on Machine Learning*, 2023.

Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *Computer and Communications Security*, 2022.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Computer Security Foundations Symposium*, 2018.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 2021.

Michael Zhang, Nimit S Sohoni, Hongyang R Zhang, Chelsea Finn, and Christopher Ré. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. In *International Conference on Machine Learning*, 2022.

Da Zhong, Haipei Sun, Jun Xu, Neil Gong, and Wendy Hui Wang. Understanding disparate effects of membership inference attacks and their countermeasures. In *Asia Conference on Computer and Communications Security*, 2022.

APPENDIX

We report the dataset details, additional results on group privacy disparity, a comparison of different membership inference attack methods, define and show the memorization score for each dataset, and more results on differential privacy and model architectures.

A DATASET

Waterbirds Sagawa et al. (2019). Vision dataset where the task is to classify whether landbird or waterbird. The background is the spurious feature represented as water or land background. The presence of the spurious features induces four data groups: landbird on land background, landbird on water background, waterbird on water background, and waterbird on land background. The groups have respectively 3498, 184, 1057, and 56 samples. Therefore, the type of bird is spurious correlated with the same type of background.

CelebA Liu et al. (2014). Vision dataset where the task is to classify whether a celebrity is a male or female. The hair color is the spurious features represented as dark or blonde hair. The presence of spurious features induces four data groups: female with blonde hair, female with dark hair, male with dark hair, and male with blonde hair. The groups have respectively 71629, 66874, 22880, and 1387 samples. Therefore, blonde hair is spurious correlated with female celebrities.

FMoW Koh et al. (2021). Vision dataset where the task is to identify between 62 classes the type of land usage, e.g. hospital, airport, single or multi-use residential area. The geographical location is the spurious feature representing the continents: Asia, Europe, Africa, Americas, and Oceania. The groups have respectively 17809, 34816, 1582, 20973, and 1641 samples whereas the African countries have the majority of samples as single-use residential areas (36%). Therefore, samples collected from Africa are spurious correlated with the single-unit residential areas. Moreover, the test set presents a distribution shift with samples collected from different years.

MultiNLI Williams et al. (2017). Text dataset where the task is to identify the relationship between two pairs of text as a contradiction, entailment, or neither. The negation is the spurious feature usually found in the contradiction class. The presence of the spurious feature induces six data groups: contradiction without negation, contradiction with negation, entailment without negation, entailment with negation, neutral without negation, and neutral with negation. The groups have respectively 57498, 11158, 67376, 1521, 66630, and 1991 samples. Therefore, samples with the spurious feature negation are correlated with the contradiction class.

CivilComments Koh et al. (2021). Text dataset with the task of detecting toxic comments of online articles. The demographic identities (male, female, LGBTQ, Christian, Muslim, other religions, Black, and White) combined with the target (toxic or not) divides the dataset into 16 groups groups. The group "other religions" is spurious correlated with the target. The groups have respectively 16568, 26846, 5638, 27824, 11064, 4402, 4727, 9812, 2435, 3928, 1865, 1867, 2964, 608, 2076, and 3462 samples.

B SPURIOUS PRIVACY LEAKAGE

We report additional technical details related to Section 3 and include additional results: comparing different membership inference attacks on spurious data, demonstrating how memorization of spurious data causes higher privacy leakage.

Hyperparameters. For Section 3, we apply grid search to find the best hyperparameters for each dataset. For Waterbirds and CelebA we search the learning rate between [1e-3, 1e-4] and weight decay [1e-1, 1e-2, 1e-3]. For FMoW the learning rate [1e-3, 3e-3, 1e-4, 3e-4], weight decay [1e-1, 1e-2, 1e-3], and epochs [20, 30, 40]. For MultiNLI the learning rate [1e-5, 3e-5], weight decay [1e-5, 1e-4]. For CivilComments the learning rate [1e-5, 1e-6], weight decay [1e-3, 1e-4]. The best hyperparameters are reported at Table 3.

Table 3: Hyperparameters used to train shadow models for each dataset. Adapted from the hyperparameters of Izmailov et al. (2022). Since we trained the models using LiRA algorithm with 50% of the total dataset, we had to grid search and validate on the validation set.

Data	Optim	Batch size	LR	WD	Epochs	C
Waterbirds	SGD	32	1e-3	1e-2	100	1
CelebA	SGD	32	1e-3	1e-2	20	5
FMoW	SGD	32	3e-3	1e-2	20	1
MultiNLI	AdamW	16	1e-5	1e-4	5	8
CivilComments	AdamW	32	1e-5	1e-4	5	8

B.1 Membership inference attacks comparison

Most of the previous MIAs are limited by the assumption that all the samples have the same level of importance (or hardness) (Yeom et al., 2018; Shokri et al., 2017), which is incorrect since natural data follow a long-tail distribution (Feldman, 2020). We compare three different state-of-the-art MIAs and show that the phenomenon of *spurious privacy leakage* exists regardless of the attack used. We use two different versions of LiRA (Carlini et al., 2022), online and offline, and TrajMIA (Liu et al., 2022). The results in Table 4 show that all the methods successfully reveal the disparity on Waterbirds, and LiRA online is the strongest attack on vulnerable groups.

Table 4: Comparing the attack success rate of different membership inference attacks on ERM models trained with Waterbirds. All the methods can be used to identify the privacy disparity, but LiRA poses a greater risk for more vulnerable spurious groups. *TPRs are reported at ~1% and ~3% for groups 1 and 2 respectively due to their limited sample size. The spurious groups are highlighted.

	TPR @ 0.1% FPR (↑)			AUROC (↑)		
Group	LiRA	LiRA (offline)	TrajMIA	LiRA	LiRA (offline)	TrajMIA
1	0.22 ± 0.03	0.14 ± 0.02	$\mathbf{1.67 \pm 3.27}$	51.78 ± 0.15	49.97 ± 0.22	$\mathbf{58.20 \pm 3.42}$
2*	$\mathbf{10.87 \pm 1.18}$	5.39 ± 0.78	3.18 ± 0.47	$\mathbf{75.07 \pm 0.54}$	61.32 ± 1.01	70.28 ± 1.22
3*	$\mathbf{30.91 \pm 2.81}$	18.98 ± 2.13	14.60 ± 1.69	85.83 ± 0.76	69.50 ± 1.67	$\mathbf{86.16 \pm 2.55}$
4	1.73 ± 0.19	0.83 ± 0.11	$\mathbf{6.57 \pm 0.59}$	60.52 ± 0.34	53.63 ± 0.42	$\mathbf{72.40 \pm 2.31}$
T	1.16 ± 0.07	0.44 ± 0.04	$\mathbf{1.68 \pm 0.00}$	55.44 ± 0.14	51.43 ± 0.16	$\mathbf{74.74 \pm 0.00}$

B.2 Memorization score of spurious groups

Feldman (2020) introduced the notion of label memorization (Definition B.1) as the difference in the label of a model trained with or without $\boldsymbol{x}$. We use the models from the LiRA algorithm from Section 3.1 to approximate the memorization score. Carlini et al. (2022) proposed the privacy score $d = |\mu_{\text{in}} - \mu_{\text{out}}|/(\sigma_{\text{in}} + \sigma_{\text{out}})$ to measure the difference between the loss distributions coming from IN and OUT shadow models of LiRA. Note that both mem(.) and d measure the difference between two probability distributions conditioned on D and $D \setminus \{i\}$ but with a different level of granularity; label memorization is coarser than d and collapses the whole distributions to a single scalar, the probability of outputting the correct label.

Definition B.1 (Label memorization)**.** Label memorization is the difference in the output label of a model $f \sim \mathcal{A}(\mathcal{D})$ fit on the dataset D with or without a specific data point $(\boldsymbol{x}_i, \boldsymbol{y}_i) \sim D$. Formally, $\text{mem}(\mathcal{A}, D, i) = \left|\Pr_{f\sim\mathcal{A}(D)}\left(f(\boldsymbol{x}_i) = \boldsymbol{y}_i\right) - \Pr_{f\sim\mathcal{A}(D\setminus\{i\})}\left(f(\boldsymbol{x}_i) = \boldsymbol{y}_i\right)\right|$

We compute d for each data point and use a Gaussian kernel density estimator to fit each group. The results in Fig. 3 show the estimated frequency of the memorization score for the whole dataset divided per group. We observe that the spurious groups have, on average, higher memorization scores compared to non-spurious groups (except for FMoW as in Fig. 1). The increase can be attributed to the presence of spurious features, which turn typical examples into atypical ones that the model has to memorize. A higher memorization score is known to be linked to a higher vulnerability under privacy attacks (Feldman, 2020), which matches what we observed previously.

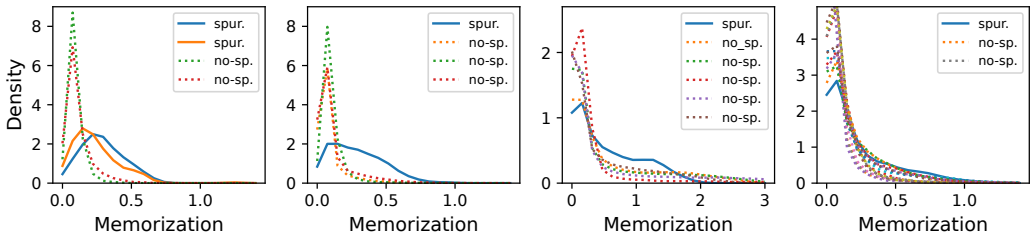


Figure 3: Memorization score divided per group on Waterbirds, CelebA, MultiNLI, and CivilComments respectively. Spurious correlated groups (solid lines) have on average a higher memorization score than non-spurious groups, which indicates that models treat spurious groups as atypical examples. In the FMoW dataset, all the groups have similar levels of memorization.

B.3 Task complexity and privacy leakage

The FMoW dataset has similar privacy vulnerabilities across the groups. FMoW is a more challenging task with 62 classes compared to 2 of Waterbirds. Given a fixed dataset, we show that the number of classes is related to the feature complexity and *spurious privacy leakage*.

Experiment setup. We create two new datasets with 16 (FMoW16) and 4 (FMoW4) classes by sequentially clustering the 62 classes of the original FMoW. We train 16 shadow models for each dataset as in Section 3.1 and use LiRA for privacy analysis. The results are averaged over 5 different target models.

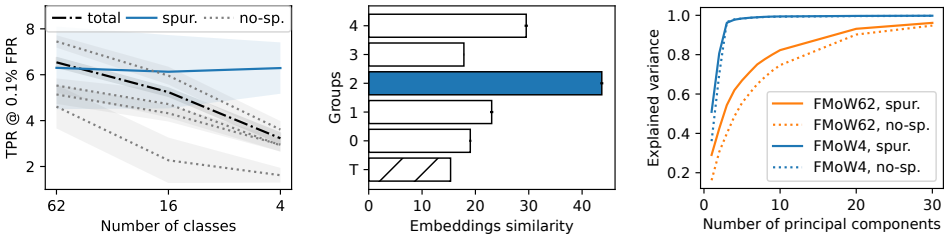
Figure 4a shows the decrease of average privacy risk over the total dataset as the task simplifies from 62 to 4 classes (black dot-dashed line). This observation is consistent with prior works on balanced datasets, such as the higher vulnerability in CIFAR100 compared to CIFAR10 (Shokri et al., 2017; Carlini et al., 2022). Interestingly, when zooming in on the dataset using our per-group analysis, we observe that *the group privacy disparity emerges between the spurious and non-spurious groups as the task simplifies*. While the leakage for most of the groups drops, the spurious group 2 remains consistently vulnerable at 6% TPR at 0.1% FPR across the dataset with different classes.

We hypothesize that spurious groups are characterized by *similar* and *fewer* discriminative features across tasks (FMoW62 and FMoW4), causing spurious privacy leakage. Firstly, we show that the embeddings of the spurious group between FMoW62 and FMoW4 are more *similar* than the other groups. We use the linear centered kernel alignment (CKA, Kornblith et al. (2019)) to quantify the similarity between the pre-layer layer embeddings. Figure 4b shows that the spurious group (blue-colored bar) has the highest CKA embedding similarity among all the groups, confirming a higher number of discriminative features in common. Secondly, we show that as the task simplifies, *fewer* discriminative features are learned. We use the PCA on the same embeddings and compute its explained variance. The results in Figure 4c show that the FMoW4's feature embeddings variance can be explained with *fewer* features than FMoW62, indicating simpler learned features. Additionally, for each dataset, spurious groups require *fewer* features than non-spurious groups to explain the variance. These analysis support our hypothesis, showing that spurious privacy leakage depends on the task complexity.

C Robust training

We report additional technical details related to Section 4 and include an additional result analyzing the privacy side effect of choosing L2 vs L1 regularization in DFR.

Hyperparameters. We use the same hyperparameters as in Table 3. Robust training DRO requires an extra hyperparameter C. For Waterbirds and CelebA we tune C within [0, 1, 2, 3, 4], for FMoW [0, 1, 2, 4, 8, 16], and for MultiNLI and CivilComments [0, 1, 2, 4, 8, 16]. For DFR, we do not use the validation set for retraining but use a group-balanced subset sampled from the training set. This allows a fairer comparison with other methods by not exploiting additional data, and it is also necessary for a fair privacy analysis since adding extra data invalidates the membership inference comparison.



(a) Privacy disparity (b) Group similarity (c) Feature variance

Figure 4: (a) Group privacy disparity increases as the target complexity reduces from FMoW62 to FMoW4. The solid line, representing the spurious group 2, remains constant while the other groups become less vulnerable. (b) Embeddings similarity of each group between FMoW62 and FMoW4 using linear CKA. The most similar group is the spurious group 2 colored in blue. (c) Feature complexity using explainable variance for embeddings of models trained on FMoW62 and FMoW4. FMoW4 requires only 3 principal components compared to 25 of FMoW62 to explain $\approx 95\%$ of the variance. Additionally, spurious groups need fewer components than non-spurious groups.

Experiment setup. For the LiRA attack, we train 32 ERM shadow models for Waterbirds and CelebA and 16 ERM shadow models for FMoW, MultiNLI, and CivilComments. We also train 5 DRO and DFR models for all the datasets. We use the online version of LiRA with a fixed variance for all the attacks to audit the privacy level. Table 2 reports the mean and standard error of using the ERM trained shadow models to attack 32 target models for each training type of Waterbirds and CelebA, and 5 for FMoW and MultiNLI.

D Compute resources

All the experiments are run on our internal cluster with the GPU Tesla V100 16GB/32GB of memory. We give an estimate of the amount of compute required for each experiment. For Section 3, we trained 96 shadow models for Waterbirds and CelebA, and 48 for FMoW, MultiNLI, and CivilComments which took ~600 hours of computing. We trained 16 shadow models for FMoW4 and FMoW16 which took another ~100 hours. The full research required additional computing for hyperparameter grid searches.

Table 5: Comparing the attack success rate of different training methods for spurious and non-spurious groups. This it the full version of the Table 2 in the main text. The spurious groups are highlighted. *The spurious groups 1 and 2 of Waterbirds are evaluated at 1% and 3% respectively due to the limited samples. DRO fails to improve the accuracy on FMoW after an extensive grid search, therefore we omit it.

		TPR @ 0.1% FPR (↓)			AUROC (↓)		
Data	**Group (n)**	**ERM**	**DRO**	**DFR**	**ERM**	**DRO**	**DFR**
Waterb.	0 (1749)	0.22 ± 0.03	0.22 ± 0.03	0.22 ± 0.03	51.78 ± 0.15	**51.59 ± 0.16**	51.64 ± 0.16
	1 (92)*	**10.87 ± 1.18**	10.91 ± 1.08	11.16 ± 1.20	75.07 ± 0.54	**74.69 ± 0.58**	75.15 ± 0.52
	2 (28)*	**30.91 ± 2.81**	31.06 ± 2.76	33.20 ± 2.83	85.83 ± 0.76	**85.54 ± 0.77**	86.17 ± 0.79
	3 (528)	**1.73 ± 0.19**	**1.73 ± 0.19**	1.91 ± 0.20	60.52 ± 0.34	**60.33 ± 0.42**	60.66 ± 0.33
	T (2397)	1.16 ± 0.07	**1.13 ± 0.06**	1.19 ± 0.06	55.44 ± 0.14	**55.23 ± 0.17**	55.39 ± 0.15
CelebA	0 (35814)	0.53 ± 0.01	**0.51 ± 0.02**	0.52 ± 0.02	53.12 ± 0.05	**52.89 ± 0.15**	53.04 ± 0.11
	1 (33437)	0.27 ± 0.01	**0.26 ± 0.01**	**0.26 ± 0.01**	50.58 ± 0.05	**50.48 ± 0.10**	50.56 ± 0.06
	2 (11440)	1.64 ± 0.05	**1.58 ± 0.06**	1.62 ± 0.05	59.77 ± 0.08	59.44 ± 0.26	**59.36 ± 0.26**
	3 (693)	4.61 ± 0.50	**4.56 ± 0.48**	4.77 ± 0.46	80.51 ± 0.21	**79.95 ± 0.52**	80.00 ± 0.48
	T (81384)	0.76 ± 0.01	**0.73 ± 0.02**	0.74 ± 0.01	53.43 ± 0.04	**53.22 ± 0.14**	53.30 ± 0.11
MultiNLI	0 (14374)	6.95 ± 0.66	**6.65 ± 0.61**	6.78 ± 0.62	74.36 ± 0.36	**74.23 ± 0.40**	74.26 ± 0.34
	1 (2789)	**2.03 ± 0.21**	2.12 ± 0.22	2.13 ± 0.21	**56.81 ± 1.21**	56.95 ± 1.37	56.98 ± 1.36
	2 (16844)	5.88 ± 0.42	**5.73 ± 0.45**	5.86 ± 0.40	72.04 ± 0.31	**71.93 ± 0.30**	72.01 ± 0.30
	3 (380)	6.14 ± 1.84	**5.66 ± 1.73**	6.22 ± 1.84	77.41 ± 0.33	77.28 ± 0.27	**77.25 ± 0.30**
	4 (16657)	5.81 ± 0.25	**5.67 ± 0.25**	5.85 ± 0.28	75.83 ± 0.15	**75.64 ± 0.20**	75.67 ± 0.13
	5 (498)	8.26 ± 0.55	9.08 ± 1.37	**7.78 ± 0.57**	83.70 ± 0.53	**83.49 ± 0.61**	83.59 ± 0.57
	T (51542)	5.95 ± 0.42	**5.83 ± 0.44**	5.93 ± 0.42	73.44 ± 0.16	**73.31 ± 0.19**	73.33 ± 0.13
CivilCom.	0 (8284)	0.13 ± 0.02	0.12 ± 0.02	**0.10 ± 0.02**	**50.46 ± 0.14**	50.59 ± 0.36	50.55 ± 0.36
	1 (13423)	0.12 ± 0.03	**0.12 ± 0.02**	0.13 ± 0.02	**50.38 ± 0.22**	50.68 ± 0.47	50.68 ± 0.46
	2 (2819)	0.16 ± 0.04	**0.11 ± 0.03**	0.13 ± 0.03	**49.79 ± 0.38**	50.19 ± 0.26	49.97 ± 0.24
	3 (13912)	**0.10 ± 0.02**	0.14 ± 0.03	0.12 ± 0.03	50.60 ± 0.25	50.44 ± 0.24	**50.40 ± 0.24**
	4 (5532)	**0.11 ± 0.01**	0.12 ± 0.02	0.13 ± 0.01	**50.09 ± 0.18**	50.55 ± 0.22	50.56 ± 0.22
	5 (2201)	0.23 ± 0.07	0.30 ± 0.09	**0.18 ± 0.05**	**50.59 ± 0.44**	50.82 ± 0.42	50.84 ± 0.38
	6 (2363)	0.08 ± 0.06	**0.07 ± 0.05**	0.08 ± 0.06	**49.60 ± 0.25**	50.03 ± 0.22	50.18 ± 0.21
	7 (4906)	0.11 ± 0.02	0.13 ± 0.04	**0.09 ± 0.02**	50.07 ± 0.36	**50.00 ± 0.28**	50.08 ± 0.22
	8 (1217)	0.20 ± 0.07	**0.11 ± 0.05**	0.14 ± 0.07	50.03 ± 0.34	**49.55 ± 0.59**	50.22 ± 0.49
	9 (1964)	0.21 ± 0.06	**0.16 ± 0.04**	0.18 ± 0.04	49.99 ± 0.44	49.91 ± 0.22	**49.81 ± 0.39**
	10 (932)	**0.11 ± 0.04**	**0.11 ± 0.04**	0.11 ± 0.04	**50.13 ± 0.86**	50.51 ± 1.01	50.64 ± 0.96
	11 (933)	**0.07 ± 0.04**	0.21 ± 0.12	0.24 ± 0.11	49.31 ± 0.69	48.87 ± 0.61	**48.75 ± 0.60**
	12 (1482)	0.02 ± 0.02	**0.00 ± 0.00**	**0.00 ± 0.00**	49.66 ± 0.24	50.23 ± 0.44	**49.84 ± 0.33**
	13 (304)	0.44 ± 0.10	0.43 ± 0.17	**0.32 ± 0.12**	50.72 ± 1.70	50.91 ± 1.86	**50.53 ± 1.71**
	14 (1038)	0.22 ± 0.11	**0.19 ± 0.11**	0.26 ± 0.12	51.84 ± 0.59	51.84 ± 0.63	**51.80 ± 0.56**
	15 (1731)	0.38 ± 0.10	**0.26 ± 0.07**	0.30 ± 0.09	49.78 ± 0.48	50.25 ± 0.62	**49.76 ± 0.47**
	T (38018)	**0.09 ± 0.01**	**0.09 ± 0.01**	**0.09 ± 0.01**	**50.24 ± 0.07**	50.41 ± 0.11	50.38 ± 0.09
FMoW	0 (8904)	**5.14 ± 0.41**	-	5.22 ± 0.37	83.70 ± 0.05	-	**83.60 ± 0.05**
	1 (17408)	**7.45 ± 0.27**	-	7.61 ± 0.28	85.12 ± 0.07	-	**84.96 ± 0.08**
	2 (791)	**6.30 ± 1.80**	-	6.42 ± 1.80	**81.54 ± 0.22**	-	81.62 ± 0.24
	3 (10486)	**5.53 ± 0.31**	-	5.69 ± 0.32	82.85 ± 0.13	-	**82.74 ± 0.13**
	4 (820)	**4.61 ± 0.95**	-	5.34 ± 0.88	80.37 ± 0.45	-	**80.26 ± 0.52**
	T (38409)	6.54 ± 0.21	-	**6.47 ± 0.17**	**84.02 ± 0.05**	-	83.90 ± 0.03