Teaching Invariance Using Privileged Mediation Information

Dylan Zapzalka
Division of Computer Science & Engineering
University of Michigan, Ann Arbor
dylanz@umich.edu

Maggie Makar
Division of Computer Science & Engineering
University of Michigan, Ann Arbor
mmakar@umich.edu

Abstract

The performance of deep neural networks often deteriorates in out-of-distribution settings due to relying on easy-to-learn but unreliable spurious associations known as shortcuts. Recent work attempting to mitigate shortcut learning relies on *a priori* knowledge of what the shortcut is and requires a strict overlap assumption with respect to the shortcut and the labels. In this paper, we present a causally-motivated teacher-student framework that encourages invariance to all shortcuts by leveraging *privileged mediation information*. The Teaching Invariance using Privileged Mediation Information (TIPMI) framework distills knowledge from a counterfactually invariant teacher trained using privileged mediation information to a student predictor that uses non-privileged features. We analyze the theoretical properties of our proposed estimator, showing that TIPMI promotes invariance to multiple unknown shortcuts and has better finite-sample efficiency. We empirically verify our theoretical findings by showing that TIPMI outperforms several state-of-the-art methods on one language and two vision datasets.

1 Introduction

Neural networks are often deployed on data that is different from the training data — a phenomenon known as distribution shift [51, 8]. Many predictors have been shown to have brittle performance under distribution shifts [23, 10, 21, 46]. One reason for this is shortcut learning: when a predictor relies on easy-to-learn but inconsistent features in the training data that are spuriously correlated with the target label [14, 32]. If the correlation between the shortcut and the label changes, the model's performance declines significantly. Therefore, for a model to be robust to distribution shifts, it should only rely on features that are predictive of the label and invariant across various distributions [31].

In this paper, we focus on the anti-causal prediction setting, where a target label causally affects the features. As a motivating example, consider a predictor trained using X-ray images from a hospital to predict if a patient has knee osteoarthritis (KOA). An ideal predictor would only use invariant medically relevant features constructed from the X-ray, such as the appearance of the joints or joint space narrowing, to make a diagnosis. However, X-rays often contain inconsistent spurious information that models exploit to make a prediction, such as hospital-specific X-ray artifacts [54].

Previous work has attempted to learn invariant features by leveraging additional information available only at training time known as *privileged information*. Most of these approaches utilize privileged shortcut information, typically in the form of labels representing potential shortcuts or environments that models should be invariant to [1, 32, 48, 44, 15]. Although effective in specific environments, these methods have limitations. (1) They rely on a restrictive overlap assumption. Specifically, they assume that the probability of observing an example from each combination of a target and shortcut label is non-zero. (2) They assume that all potential shortcuts are known at training time. This is particularly difficult because it requires insight into spurious correlations inherent to specific datasets.

Accepted to the NeurIPS 2024 Workshop on Causal Representation Learning.

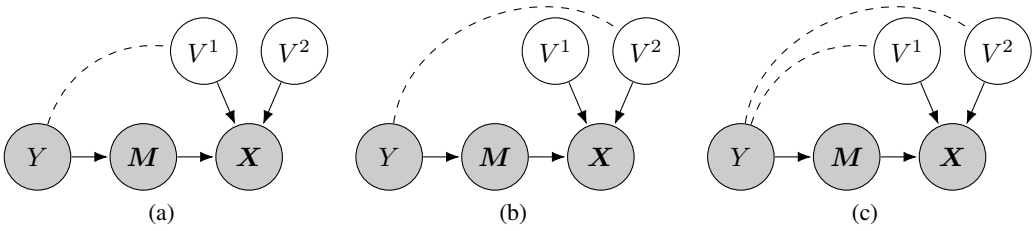


Figure 1: Examples of causal DAGs describing the setting in this paper. Grey nodes are observed at training time, while white nodes are never observed. The dashed lines denote non-causal spurious correlations. Each subfigure represents a possible source dataset distribution with the same mediator. In (a) V^1 is a shortcut, in (b) V^2 is a shortcut, and in (c) V^1 and V^2 are shortcuts.

In this paper, we present a causally-motivated teacher-student framework called Teaching Invariance using Privileged Mediation Information (TIPMI) that mitigates shortcut learning. Instead of relying on shortcut information, TIPMI leverages privileged information that mediates the causal effect of the target label on the features, which we refer to as *privileged mediation information*. We consider mediators that are available at training time but expensive to collect at test time, such as clinician annotations of X-rays. Our approach addresses the limitations of work that relies on known shortcuts by relaxing the restrictive overlap assumptions. In addition, we do not assume known shortcuts but rather require knowledge of mediators, which are typically known to domain experts.

Our contributions are as follows: (1) We propose a causally-motivated knowledge distillation framework to discourage shortcut learning. (2) We empirically demonstrate that our approach is better at promoting invariance compared to baselines that utilize privileged shortcut information. (3) We theoretically show that TIPMI leads to better finite-sample efficiency than self-distillation and methods that use privileged shortcut information. (4) We investigate the empirical performance of our approach using two image datasets and one language dataset, showing that our approach leads to more robust and efficient models without restrictive overlap assumptions.

2 Preliminaries

2.1 Problem Setting

We consider a supervised learning setting where we wish to learn some model $f : \boldsymbol{X} \to Y$. We assume that we have access to privileged mediation information $\boldsymbol{M}$, a variable that fully mediates the causal relationship between Y and $\boldsymbol{X}$. In the KOA example, $\boldsymbol{X}$ and Y represent the X-ray and the presence/severity of KOA, while $\boldsymbol{M}$ could represent an expert segmentation of the joint. We assume that $\boldsymbol{M}$ is available only at training time, which happens in settings where data derived from experts may be difficult to acquire at test time due to limited resources. Throughout, we use uppercase letters to denote variables and lowercase to denote their value. Our training data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{m}_i, y_i)\}_{i=1}^n$ is drawn from some source distribution P_s, where the labels Y are spuriously correlated with $\boldsymbol{X}$ through a multiple unknown potential shortcuts $V^1, V^2, ..., V^n$. We denote $\boldsymbol{V}$ as a vector containing all shortcuts. Our goal is to create a model f that is invariant to *any* shortcut represented in $\boldsymbol{V}$.

We assume that the training data is generated in accordance with the causal DAGs in Figure 1, where each DAG represents a possible distribution. For simplicity, we consider a setting where $\boldsymbol{V} = [V^1, V^2]$, but we stress that our results hold for $\boldsymbol{V}$ of arbitrary size. The DAGs represent the anti-causal setting, where $\boldsymbol{X}$ is generated by V^1, V^2 and the mediator $\boldsymbol{M}$. Here, V^1 and V^2 are spuriously correlated but not causally related to Y, as indicated by dashed lines. We also assume the existence of an unknown deterministic function $\pi(\boldsymbol{X})$ that can perfectly recover $\boldsymbol{M}$:

Assumption 1. *$\boldsymbol{M}$ can be fully recovered from $\boldsymbol{X}$ by some deterministic function $\boldsymbol{M} := \pi(\boldsymbol{X})$.*

In the KOA example, spurious correlations include correlations between the KOA status and the X-ray image quality. This happens, for example, if patients with more advanced KOA are more likely to receive care in a skilled nursing facility with lower-quality X-rays, while healthier patients receive care

in large hospitals with high-quality X-rays. However, we assume that medically relevant information $\boldsymbol{M}$ is constant across all hospitals. Following this example, we assume each target distribution, denoted P_t, can be obtained by replacing $P_s(\boldsymbol{V}|Y)$ with the target conditional distribution $P_t(\boldsymbol{V}|Y)$:

$$\mathcal{P} = \{P_s(\boldsymbol{X}|\boldsymbol{M}, \boldsymbol{V})P_s(\boldsymbol{M}|Y)P_s(Y)P_t(\boldsymbol{V}|Y)\}. \tag{1}$$

Let $P^\circ \in \mathcal{P}$ where $P^\circ = P_s(\boldsymbol{X}|\boldsymbol{M}, \boldsymbol{V})P_s(\boldsymbol{M}|Y)P_s(Y)P^\circ(\boldsymbol{V})$ denotes the unconfounded distribution [32, 11]. Note that unless $P_s = P^\circ$, the Bayes optimal predictor will not be invariant to spurious correlations between $\boldsymbol{X}$ and Y [32, 48]. This is because there are two open pathways between $\boldsymbol{X}$ and Y: the front door path, $\boldsymbol{X} \to \boldsymbol{M} \to Y$, and the spurious back door path, $\boldsymbol{X} \to \boldsymbol{V} \to Y$. Our approach hinges on using $\boldsymbol{M}$ at training time to discourage the use of the spurious back door path.

2.2 Counterfactual Invariance

Using the potential outcomes notation [42], we denote $\boldsymbol{X}(v)$ as the counterfactual value of $\boldsymbol{X}$ where some V is set to v, and all other variables remain the same. Since we assume that the association between V and Y is purely spurious, the predictions made by a robust model should not change given $\boldsymbol{X}(v)$ or $\boldsymbol{X}(v')$, where $v \neq v'$. Ideally, this property should hold for all possible shortcuts in $\boldsymbol{V}$. We say that models that satisfy this robustness property are counterfactually invariant to $\boldsymbol{V}$. This definition extends the definition from [48] to account for all shortcuts.

Definition 1. *Let $\boldsymbol{V} = [V^1, V^2, ..., V^n]$ contain all possible shortcuts under P_s. A model $f : \boldsymbol{X} \to Y$ is counterfactually invariant to $\boldsymbol{V}$ if $f(\boldsymbol{X}(v^i)) = f(\boldsymbol{X}(v^{i'}))$ almost everywhere for all v^i, $v^{i'}$ in the sample space of V^i where $i = 1, 2, ..., n$.*

While counterfactual invariance is a natural property to strive for in a robust predictor, we cannot verify or directly enforce that a predictor $f : \boldsymbol{X} \to Y$ is counterfactually invariant without counterfactual examples, which are never observed. Instead, previous work has promoted counterfactual invariance by enforcing conditional independences, or "signatures," that are held by the unknown counterfactually invariant predictor. For instance, [48, 32, 55] enforce $f(\boldsymbol{X}) \perp\!\!\!\perp V|Y$, where V is an observed shortcut. These approaches are limited as (i) they only promote invariance to known shortcuts, (ii) they assume access V at training time, and (iii) they assume strict overlap assumptions, e.g., $0 < P(Y|V = v) < 1$.

3 Teaching Invariance Using Privileged Mediation Information

In this section, we present TIPMI, a novel approach that uses privileged mediation information to enforce invariance to all unknown shortcuts.

3.1 Counterfactually Invariant Signatures With Mediation Information

Following previous work that constructs signatures of counterfactual invariance through conditional independence relationships [48], a valid but naïve signature to enforce with mediation information would be $f(\boldsymbol{X}) \perp\!\!\!\perp Y|\boldsymbol{M}$. This approach is limited as when $\boldsymbol{M}$ is high dimensional, it's difficult to enforce with conditional independence tests due to the curse of dimensionality [40, 41]. A more reliable signature to enforce is $f(\boldsymbol{X}) = g_0(\boldsymbol{M})$, where g_0 is the optimal counterfactually invariant predictor $\mathbb{E}_{P_s}[Y|\boldsymbol{M}]$ and $\boldsymbol{M}, \boldsymbol{X} \sim P_s$. Importantly, the Bayes optimal function under any $P_s \in \mathcal{P}$ that satisfies this signature must be counterfactually invariant. We state this in Proposition 1.

Proposition 1. *Let P_s be any source distribution consistent with the causal DAGs in Figure 1. Also, let $g_0(\boldsymbol{M}) = \mathbb{E}_{P_s}[Y|\boldsymbol{M}]$ and $f_0(\boldsymbol{X}) = \mathbb{E}_{P_s}[g_0(\boldsymbol{M})|\boldsymbol{X}]$. Then $g_0(\boldsymbol{M})$ is an optimal counterfactually invariant predictor and $f_0(\boldsymbol{X})$ is counterfactually invariant under any distribution in $\mathcal{P}$.*

The signature $f(\boldsymbol{X}) = g_0(\boldsymbol{M})$ has several advantages compared to signatures constructed from conditional independences. First, it can be viewed as a more powerful signature than any conditional independence relationship, as it enforces all the conditional independencies that are necessary for the predictor to be counterfactually invariant to $\boldsymbol{V}$. This is because if $f(\boldsymbol{X}) = g_0(\boldsymbol{M})$ for $\boldsymbol{M}, \boldsymbol{X} \sim P_s$, then all of the conditional independences of $g_0(\boldsymbol{M})$ must also hold for $f(\boldsymbol{X})$ under P_s. Critically, this implies that $f(\boldsymbol{X}) = g_0(\boldsymbol{M})$ enforces conditional independences $f(\boldsymbol{X}) \perp\!\!\!\perp V|Y$ for all shortcuts in $\boldsymbol{V}$, even if they are unobserved. This means that the signature promotes invariance to all shortcuts. Furthermore, the mediator signature does not require the overlap assumption $0 < P(Y|V = v) < 1$.

3.2 TIPMI Algorithm

TIPMI trains a teacher to predict Y from $\boldsymbol{M}$ and trains a student to match the teacher's predictions from $\boldsymbol{X}$. The TIPMI signature can be viewed as a semiparametric inference problem, where g_0 is some unknown nuisance parameter, and the teacher $\hat{g}$ is a plug-in estimate of g_0 [12, 9]. This interpretation is useful as it emphasizes the importance of cross-fitting to avoid teacher overfitting [9].

Teacher (Step 1) Train multiple teachers using cross-fitting:

1. Generate K-fold partitions I_1, I_2, ..., I_k of the training data $|\mathcal{D}| = N$, where $|I_k| = \frac{N}{K}$.
2. For each fold $k \in \{1, ..., K\}$ and a suitable loss function ℓ_g, train a teacher model on the combined folds $\mathcal{D}_k^c = \{(\boldsymbol{m}_i, y_i)\}_{i \in I_k^c}$ where $I_k^c := \{1, 2, ..., N\} \backslash I_k$ as follows
$$\hat{g}^k = \arg\min_g \frac{1}{|\mathcal{D}_k^c|} \sum_{i=1}^{|\mathcal{D}_k^c|} \ell_g(g(\boldsymbol{m}_i), y_i)$$
3. For $k \in \{1, 2, ..., K\}$, use the teacher models to generate new datasets where $\hat{y}_i = \hat{g}^k(\boldsymbol{m}_i)$:
$$\mathcal{D}_k = \{(\boldsymbol{x}_i, \hat{y}_i)\}_{i \in I_k}$$
4. Combine the datasets generated by each teacher to create the final dataset for the student:
$$\mathcal{D}_{cf} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup ... \cup \mathcal{D}_k$$

Student (Step 2) Using the cross-fitting dataset $\mathcal{D}_{cf}$, we train the student using the mean squared error (MSE) so that the students predictions match the teachers:
$$\hat{f} = \arg\min_f \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - \hat{y}_i)^2.$$

4 Efficient Learning Using Privileged Mediation Information

In this section, we analyze the finite-sample properties of TIPMI and show that it leads to better finite-sample efficiency by comparing it directly to self-distillation (SD). We restrict the analysis to settings where the training and testing distributions are the unconfounded distribution P°, as we are only interested in how fast each method can learn invariant features rather than the robustness of the predictors. To simplify our exposition, we consider the special case of linear binary classifiers, where the teacher models take the form of $g(\boldsymbol{x}) = \sigma(\boldsymbol{w}^T \boldsymbol{x})$, ℓ_g is the squared loss, and $\sigma(x)$ denotes the sigmoid function. Extensions of our analysis to DNNs can be done using tools presented in [16].

$$\mathcal{G}_{\text{SD}} := \{g : \boldsymbol{x} \to \sigma(\boldsymbol{w}^T \boldsymbol{x}), \|\boldsymbol{w}\|_2 \leq A\}$$
$$\mathcal{G}_{\text{TIPMI}} := \{g : \boldsymbol{m} \to \sigma(\boldsymbol{w}^T \boldsymbol{m}), \|\boldsymbol{w}\|_2 \leq A\}$$

Our results are a direct corollary of Lemma 4 in [9], which shows that the students models performance increases as the teachers predictions get closer to the Bayes optimal probabilities. Following this insight, we show that, in general, the teacher estimation error under TIPMI is much lower than the error under SD. To prove this, we present an upper bound on the teacher estimation error using the Rademacher complexity. A formal definition of the Rademacher complexity, the full generalization bounds, and additional analysis comparing TIPMI to other methods are in the Appendix.

We start by defining $\mathcal{M}$ as an $d \times n$ matrix where the columns are mediators $\boldsymbol{m}$ that span the sample space of $\boldsymbol{M}$. We define the projection matrix $\Pi := \mathcal{M}(\mathcal{M}^T \mathcal{M})^{-1} \mathcal{M}^T$, which projects any vector onto $\mathcal{M}$ where $\boldsymbol{m}_\parallel := \Pi \boldsymbol{x}$ and $\boldsymbol{m}_\perp := (I - \Pi)\boldsymbol{x}$. Here, we can think of $\boldsymbol{m}_\perp$ as the part of $\boldsymbol{x}$ that contains all irrelevant or spurious information and $\boldsymbol{m}_\parallel$ as the part of $\boldsymbol{x}$ that contains mediator information. Similarly, we define $\boldsymbol{w}_\parallel := \Pi \boldsymbol{w}$ and $\boldsymbol{w}_\perp := (I - \Pi)\boldsymbol{w}$.

Proposition 2. *Let $\boldsymbol{m}_\parallel := \Pi \boldsymbol{x}$ and $\boldsymbol{m}_\perp := (I - \Pi)\boldsymbol{x}$, and $\mathcal{R}(\mathcal{G})$ be the Rademacher complexity of some function space $\mathcal{G}$. For training data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{m}_i, y_i)\}_{i=1}^n$ where $\mathcal{D} \sim P^\circ$, also we have that $\sup_{\boldsymbol{m}_\perp} \|\boldsymbol{m}_\perp\|_2 \leq B_\perp$, and $\sup_{\boldsymbol{m}_\parallel} \left\|\boldsymbol{m}_\parallel\right\|_2 \leq B_\parallel$. Then*

$$\mathcal{R}(\mathcal{G}_{SD}) \leq \frac{A\sqrt{B_\parallel^2 + B_\perp^2}}{\sqrt{n}} \quad \textit{and} \quad \mathcal{R}(\mathcal{G}_{TIPMI}) \leq \frac{A \cdot B_\parallel}{\sqrt{n}}$$

Proposition 2 shows the upper bound for the Rademacher complexity of the TIPMI teacher is always smaller than or equal to the bound of the SD teacher. These results are similar to Proposition 5 in [32], which shows that leveraging privileged shortcut information can reduce Rademacher complexity. However, unlike [32], our results reduce the Rademacher complexity with respect to all shortcuts.

5 Experiments

In this section, we empirically demonstrate that TIPMI (1) promotes invariance to multiple unknown shortcuts, (2) works when the shortcut and target labels have no overlap, and (3) has better finite-sample efficiency than the baselines. Additional experiments and information are in the Appendix.

Datasets. *Waterbirds* is a binary image classification dataset comprised of waterbirds and landbirds over a water or land background [44]. The objective is to determine the type of bird (Y) appearing in an image ($\boldsymbol{X}$) where the bird type is spuriously correlated with the background ($\boldsymbol{V}$) and image artifacts represented by black patches. For TIPMI, the teacher uses the segmentation of the bird ($\boldsymbol{M}$).

Knee osteoarthritis is a binary image classification dataset comprised of knee X-rays ($\boldsymbol{X}$) where the goal is to predict if the knee has osteoarthritis (Y). The diagnosis of osteoarthritis is spuriously correlated with boxes ($\boldsymbol{V}$) in the X-ray, which are supposed to mimic metallic tokens [54]. The teacher is provided joint space width measurements ($\boldsymbol{M}$), which we assume fully mediate the diagnosis [20].

Food Review dataset is derived from the Amazon Food Review [33] dataset that contains reviews ($\boldsymbol{X}$), review scores (Y), and summaries of the reviews ($\boldsymbol{M}$). To create shortcuts, we add perturbation ($\boldsymbol{V}$) to the words "a" and "the" such that "a" becomes "axxxxx" and "the" becomes "thexxxxx" [48]. An additional shortcut is added similarly. We use this dataset for sensitivity analysis as we expect the mediator does not fully mediate the causal effect of Y on $\boldsymbol{X}$, which is a violation of our assumptions.

Baseline Algorithms Our experiments analyze **TIPMI**, which is our proposed method. We also conduct an ablation study to highlight the importance of cross-fitting with **TIPMI-NC**, which doesn't use any form of sample splitting. For baselines, we use **MMD**, which enforces the signature $f(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{V}|Y$ [48], **GDRO** [44], **IRM** [1], **L2** regularization, and **SD**.

Experiments At training time, for some shortcut(s) $\boldsymbol{V}$, we fix $P_s(Y|\boldsymbol{V})$ to a distribution such that Y and $\boldsymbol{V}$ are spuriously correlated. At test time, we sample the data from different test distributions where $P_t(Y|\boldsymbol{V})$ varies. We measure the area under receiver operating curve (AUROC) on each of the test distributions. A counterfactually invariant model's performance will remain the same across all distributions, whereas the performance of a model that relies on shortcuts will vary significantly.

5.1 Invariance Without Overlap

We analyze how well each method promotes invariance to a single shortcut when overlap with respect to the shortcut and label is violated. Both IRM and MMD were not evaluated as they are not well defined in this setting. In Figure 2, the bottom subplots report the results from where the training data was sampled from a spuriously correlated distribution with $P(Y = 1|V^1 = 1) = P(Y = 0|V^1 = 0) = 1.0$. For both the waterbirds and KOA datasets, TIPMI has a similar performance to L2 in distribution and better performance in all other distributions. We also note that in the waterbirds and food review datasets, TIPMI outperforms TIPMI-NC by a significant margin, suggesting that cross-fitting can help enforce our signature. However, there is no difference between TIPMI and TIPMI-NC under the KOA dataset. This is expected as the teacher used in these experiments is a small, single-layer neural network, which is less prone to overfitting than the large networks used in the other experiments. GDRO does no better than L2 in all three experiments, highlighting a significant limitation. This happens as GDRO minimizes the prediction error for the worst-performing group, where the group is defined with respect to the shortcut. In the case with no overlap, some of these groups are not observed, which means GDRO does not control their error. From our analysis of the food review experiments, we showed that TIPMI promotes invariance even when our assumptions of full mediation are violated. Here, TIPMI performs better for most distributions but performs worse for distributions close to the training distribution.

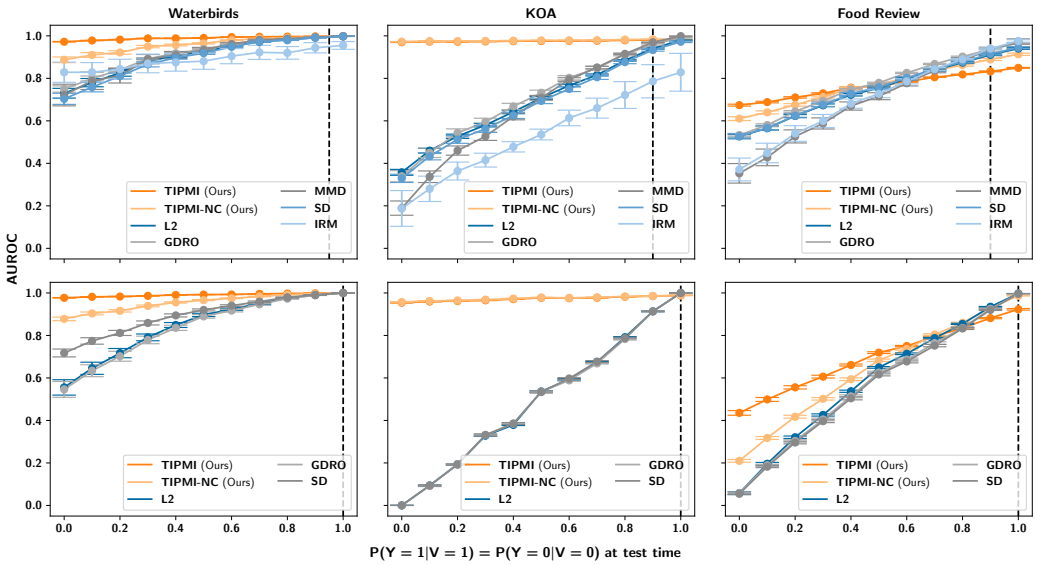


Figure 2: In the top three figures, the x-axis shows $P(Y|V^2)$ at test time under different shifted distributions for one of the two shortcuts, whereas for the bottom three figures, the x-axis shows $P(Y|V^1)$ under different shifted distributions for a single shortcut. For all figures, the y-axis shows the AUROC over the test data, and the dashed vertical line shows $P(Y|\boldsymbol{V})$ at training time. **(Left Figures)** Waterbirds dataset: TIPMI outperforms all baselines. **(Middle Figures)** KOA dataset: TIPMI outperforms all baselines. **(Right Figures)** Food review dataset: TIPMI is more robust than others but performs worse in distribution due to a violation of assumptions.

5.2 Invariance to Multiple Shortcuts

We analyze how well each method promotes invariance to multiple shortcuts. In Figure 2, the top subplots report the results from where V^1 and V^2 were spuriously associated with the label. At evaluation time, $P(V^1|Y=1) = P(V^1|Y=0) = 0.50$, and the distribution of the second shortcut $P(Y|V^2)$ varies. MMD, GDRO, and IRM only have access to labels for the first shortcut. For the waterbirds dataset, TIPMI outperforms all baselines over *all* distributions. For KOA, TIPMI outperforms all other models except for the distribution where the shortcut is strongest. TIPMI also promotes invariance in the food review setting, however, it is at the cost of in-distribution performance as the full mediation assumption is violated. In contrast to TIPMI, GDRO and MMD fail to promote invariance when there are multiple shortcuts. We note that MMD performs significantly worse over the food review and KOA datasets than a normal L2 model, suggesting that models that only promote invariance to some of the shortcuts can make models *more* prone to learning shortcuts. Since the MMD models are unable to leverage V_1, they instead rely more heavily on V_2 to make predictions, resulting in worse performance.

5.3 Finite-Sample Efficiency Improvements

To isolate and highlight the improvements in finite-sample efficiency that TIPMI introduces, we conduct a set of experiments where the training data is sampled from the ideal distribution, P°. Figure 3 shows the results for the KOA dataset. The results show that both TIPMI and TIPMI-NC outperform SD, L2, and MMD models across all test distributions. In this experiment, TIPMI-NC outperforms TIPMI, as (1) the TIPMI-NC teacher isn't susceptible to overfitting, and (2) the TIPMI teachers are trained on less data due to cross-fitting. Overall, these results support our theory that suggests when data is sampled from P°, TIPMI leads to better finite-sample efficiency than normal models and models that leverage privileged shortcut information. Additional experiments for waterbirds are in the Appendix.

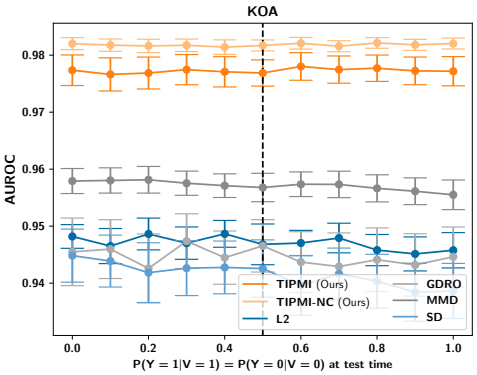


Figure 3: The x-axis shows $P(Y|\boldsymbol{V})$ at test time under different shifted distributions for one shortcut and the y-axis shows the AUROC over the test data. TIPMI-NC and TIPMI both perform much better than baselines when trained under P°.

6 Conclusion

In this work, we presented TIPMI, a framework that uses privileged mediation information to promote invariance to shortcuts and improve finite-sample efficiency. Our framework consists of a teacher trained using privileged mediation information and a student that learns from the teacher through a form of knowledge distillation. We showed theoretically and empirically that TIPMI promotes invariance to shortcuts better than methods that use privileged shortcut information and that it increases finite-sample efficiency.

7 Acknowledgments and Disclosure of Funding

We are thankful for the anonymous reviewers and for feedback from Alex D'Amour. This material is based upon work supported by the National Science Foundation under Grant No. 2337529. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[2] P. Bhargava, A. Drozd, and A. Rogers. Generalization in nli: Ways (not) to go beyond simple heuristics, 2021.

[3] C. Buciluǎ, R. Caruana, and A. Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 535–541, 2006.

[4] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl. Learning by cheating. In *Conference on Robot Learning*, pages 66–75. PMLR, 2020.

[5] Y. Chen and L. Bazzani. Learning joint visual semantic matching embeddings for language-guided retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 136–152. Springer, 2020.

[6] M. Craven and J. Shavlik. Extracting tree-structured representations of trained networks. *Advances in neural information processing systems*, 8, 1995.

[7] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[8] P. Cui and S. Athey. Stable learning establishes some common ground between causal inference and machine learning. *Nature Machine Intelligence*, 4(2):110–115, 2022.

[9] T. Dao, G. M. Kamath, V. Syrgkanis, and L. Mackey. Knowledge distillation as semiparametric inference. In *International Conference on Learning Representations*, 2020.

[10] L. Engstrom, B. Tran, D. Tsipras, L. Schmidt, and A. Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR, 2019.

[11] A. Feder, Y. Wald, C. Shi, S. Saria, and D. Blei. Causal-structure driven augmentations for text ood generalization. *Advances in Neural Information Processing Systems*, 36, 2024.

[12] D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

[13] Z. Gao, J. Chung, M. Abdelrazek, S. Leung, W. K. Hau, Z. Xian, H. Zhang, and S. Li. Privileged modality distillation for vessel border detection in intracoronary imaging. *IEEE transactions on medical imaging*, 39(5):1524–1534, 2019.

[14] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[15] K. Goel, A. Gu, Y. Li, and C. Ré. Model patching: Closing the subgroup performance gap with data augmentation. *arXiv preprint arXiv:2008.06775*, 2020.

[16] N. Golowich, A. Rakhlin, and O. Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[18] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[19] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[20] D. J. Hunter, M.-P. H. Le Graverand, and F. Eckstein. Radiologic markers of osteoarthritis progression. *Current opinion in rheumatology*, 21(2):110–117, 2009.

[21] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, and A. Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[23] P. Kirichenko, P. Izmailov, and A. G. Wilson. Last layer re-training is sufficient for robustness to spurious correlations, 2023. URL `https://arxiv.org/abs/2204.02937`.

[24] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. Concept bottleneck models. In *International conference on machine learning*, pages 5338–5348. PMLR, 2020.

[25] M. D. Kohn, A. A. Sassoon, and N. D. Fernando. Classifications in brief: Kellgren-lawrence classification of osteoarthritis. *Clinical Orthopaedics and Related Research®*, 474:1886–1893, 2016.

[26] D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] W. Lee, J. Lee, D. Kim, and B. Ham. Learning with privileged information for efficient image super-resolution. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 465–482. Springer, 2020.

[29] D. Lopez-Paz, L. Bottou, B. Schölkopf, and V. Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

[30] A. Loquercio, E. Kaufmann, R. Ranftl, M. Müller, V. Koltun, and D. Scaramuzza. Learning high-speed flight in the wild. *Science Robotics*, 6(59):eabg5810, 2021.

[31] C. Lu, Y. Wu, J. M. Hernández-Lobato, and B. Schölkopf. Invariant causal representation learning for out-of-distribution generalization. In *International Conference on Learning Representations*, 2021.

[32] M. Makar, B. Packer, D. Moldovan, D. Blalock, Y. Halpern, and A. D'Amour. Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics*, pages 739–766. PMLR, 2022.

[33] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web*, pages 897–908, 2013.

[34] A. K. Menon, A. S. Rawat, S. Reddi, S. Kim, and S. Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning*, pages 7632–7642. PMLR, 2021.

[35] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[36] M. Nevitt, D. Felson, and G. Lester. The osteoarthritis initiative. *Protocol for the cohort study*, 1:2, 2006.

[37] S. Olsson, E. Akbarian, A. Lind, A. S. Razavian, and M. Gordon. Automating classification of osteoarthritis according to kellgren-lawrence in the knee using deep learning in an unfiltered adult population. *BMC musculoskeletal disorders*, 22:1–8, 2021.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[39] M. Phuong and C. Lampert. Towards understanding knowledge distillation. In *International conference on machine learning*, pages 5142–5151. PMLR, 2019.

[40] A. Ramdas, S. J. Reddi, B. Póczos, A. Singh, and L. Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[41] S. Reddi, A. Ramdas, B. Póczos, A. Singh, and L. Wasserman. On the high dimensional power of a linear-time two sample test under mean-shift alternatives. In *Artificial Intelligence and Statistics*, pages 772–780. PMLR, 2015.

[42] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[43] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

[44] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

[45] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[46] J. Su, D. V. Vargas, and K. Sakurai. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

[47] V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural networks*, 22(5-6):544–557, 2009.

[48] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations: Why and how to pass stress tests, 2021.

[49] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[50] R. Wang, M. Yi, Z. Chen, and S. Zhu. Out-of-distribution generalization with causal invariant transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 375–385, 2022.

[51] O. Wiles, S. Gowal, F. Stimberg, S. Alvise-Rebuffi, I. Ktena, K. Dvijotham, and T. Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

[52] D. Yin, R. Gontijo Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer. A fourier perspective on model robustness in computer vision. *Advances in Neural Information Processing Systems*, 32, 2019.

[53] H. Yuan, Y. Shi, N. Xu, X. Yang, X. Geng, and Y. Rui. Learning from biased soft labels. *Advances in Neural Information Processing Systems*, 36, 2024.

[54] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[55] J. Zheng and M. Makar. Causally motivated multi-shortcut identification and removal. *Advances in Neural Information Processing Systems*, 35:12800–12812, 2022.

[56] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A Related Work

Learning Using Privileged Information and Knowledge distillation Our work builds upon the learning using privileged information (LUPI) paradigm [47, 29, 28, 4, 30, 13, 5], which utilizes privileged information from a teacher to train a student. Most LUPI work focuses on gains in efficiency or improved interpretability, in contrast to our main focus: robustness. Most similar to our work are Concept Bottleneck Models (CMBs) [24], which use high-level concepts as PI. The concepts are predicted by the first model as intermediate values, and a second model uses the concepts to make a final prediction. Unlike CBMs, TIPMI does not require explicitly modeling $\boldsymbol{M}$ as a function of $\boldsymbol{X}$. This is important in settings where $\boldsymbol{M}$ is high dimensional, such as settings where $\boldsymbol{M}$ is an image or a segmentation (similar to our waterbirds experiment).

Related to LUPI is work on knowledge distillation primarily for the purpose of model compression and explainable AI [3, 6, 19, 34, 39, 53]. Most similar to our work is [9], which views knowledge distillation as a semiparametric inference problem to show how cross-fitting and loss correction can reduce the effects of teacher overfitting and underfitting. Our work extends their exposition by analyzing the robustness of the student models and highlighting the role of discouraging shortcuts in achieving additional efficiency gains.

Shortcut learning Previous work on building invariant models explicitly relies on observing a shortcut or environment variable that can be used to induce the desired invariances [1, 26]. Closest to our work is [32, 48], who present a causally-motivated shortcut removal regularization scheme to encourage robustness to a single shortcut by leveraging observed shortcut variables at training time. While [55] encourages invariance to multiple shortcuts, they still require these shortcuts to be observed at training time. As discussed in the introduction, our work avoids the limitations inherent to knowledge or observability of the shortcut/environment label and the full overlap assumption required for these methods to perform well. Discouraging shortcut learning by using data augmentation has also been suggested [18, 52, 7, 50]. This approach can work if we have access to a generator for shortcut transformations, an assumption that we do not make.

B Proofs

Before introducing Lemma 4 from [9], we introduce the following notation. Let $\mathcal{G}$ and $\mathcal{F}$ be the student and teacher function classes. Define $f_0 = \arg\min_{f\in\mathcal{F}} R_P(f, g_0)$, where $R_P(f, g) = \mathbb{E}_{\boldsymbol{X}\sim P}[\ell_f(f(\boldsymbol{X}), g(\boldsymbol{X}))]$ and the norm $\|f\|_{\mathcal{F},P} = (\mathbb{E}_{\boldsymbol{X}\sim P}\|f(\boldsymbol{X})\|_{\mathcal{F}}^2)^{\frac{1}{2}}$. Finally, let $\nabla_{f,g}$ be the Jacobian cross partial derivative, $q_{f,g}(x) = \mathbb{E}[\nabla_{fg}\ell(f(\boldsymbol{X}), g(\boldsymbol{X}))|\boldsymbol{X} = x]$, and $\gamma_{f,g}(x) = \mathbb{E}_{U\sim\text{Unif}([0,1])}[q_{f,Ug+(1-U)g_0}(x)]$. In conjunction with Proposition B.3, Lemma B.1 can be used to obtain the full generalization bounds for TIPMI.

Lemma B.1 (Lemma 4 in [9]). *Consider any estimation algorithm that produces an estimate $\hat{f}$ with a small plug-in excess risk $R_{P^\circ}(\hat{f}, \hat{g}) - R_{P^\circ}(f_0, \hat{g}) \leq \epsilon(\hat{f}, \hat{g})$. If the loss R_{P° is σ-strongly convex with respect to f and $\mathcal{F}$ is a convex set, then*

$$\frac{\sigma}{4}\left\|\hat{f} - f_0\right\|_{\mathcal{F},P^\circ}^2 \leq \epsilon(\hat{f}, \hat{g}) + \frac{1}{\sigma}\left\|\gamma_{f_0,\hat{g}}^\top(\hat{g} - g_0)\right\|_{\mathcal{G},P^\circ}^2 \tag{2}$$

Proposition B.1 (Restated Proposition 1). *Let P_s be any source distribution defined under the causal DAG in Figure 1. Also, let $g_0(\boldsymbol{M}) = \mathbb{E}_{P_s}[Y|\boldsymbol{M}]$ and $f_0(\boldsymbol{X}) = \mathbb{E}_{P_s}[g_0(\boldsymbol{M})|\boldsymbol{X}]$. Then $g_0(\boldsymbol{M})$ is an optimal counterfactually invariant predictor and $f_0(\boldsymbol{X})$ under any distribution in $\mathcal{P}$ is counterfactually invariant.*

Proof. Let $\boldsymbol{X}, \boldsymbol{M}, V \sim P_s$ for some P_s. By Lemma 3.1 in [48], $g_0(\boldsymbol{M})$ must counterfactually invariant. Also, since any counterfactually invariant predictor must $\boldsymbol{M}$-measurable, g_0 must be optimal.

Under Assumption 1, we have that $\boldsymbol{M} = \pi(\boldsymbol{X})$, which means that all information in $\boldsymbol{M}$ can be recovered from $\boldsymbol{X}$. Since $g_0(\boldsymbol{M}) \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{M}$, f_0 is a function of $\mathbb{E}[g_0(\boldsymbol{M})|\boldsymbol{X}] = \mathbb{E}[g_0(\boldsymbol{M})|\pi(\boldsymbol{X})]$, which can be written as (with abuse of notation) $f(\boldsymbol{M})$. Since f_0 is only a function of $\boldsymbol{M}$, f_0 must be counterfactually invariant. □

Definition B.1. *Let $\boldsymbol{\epsilon} = \{\epsilon\}_{i=1}^n$ denote a vector of independent random variables where $P(\epsilon_i = 1) = P(\epsilon_i = -1) = \frac{1}{2}$. Then for the dataset $\mathcal{D} \sim P$ and the function class $\mathcal{F}$, the Rademacher complexity for a sample of size n is defined as: $\mathcal{R}(\mathcal{F}) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(\boldsymbol{x}_i)\right]\right]$.*

Proposition B.2 (Restated Proposition 2)**.** *Let $\boldsymbol{m}_{\|} := \Pi\boldsymbol{x}$ and $\boldsymbol{m}_{\perp} := (I-\Pi)\boldsymbol{x}$. For training data $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{m}_i, y_i)\}_{i=1}^n$ where $\mathcal{D} \sim P^\circ$, also let $\sup_{\boldsymbol{m}_\perp} \|\boldsymbol{m}_\perp\|_2 \leq B_\perp$, and $\sup_{\boldsymbol{m}_\|} \left\|\boldsymbol{m}_\|\right\|_2 \leq B_\|$. Then*

$$\mathcal{R}(\mathcal{G}_{SD}) \leq \frac{A\sqrt{B_\|^2 + B_\perp^2}}{\sqrt{n}} \quad \textit{and} \quad \mathcal{R}(\mathcal{G}_{TIPMI}) \leq \frac{A \cdot B_\|}{\sqrt{n}}$$

Proof. First, we derive the bound on $\mathcal{R}(\mathcal{F}_{\text{SD}})$:

$$\begin{aligned}
\mathcal{R}(\mathcal{G}_{\text{SD}}) &= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}^T\boldsymbol{x}_i] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}^T(\Pi\boldsymbol{x}_i + (I-\Pi)\boldsymbol{x}_i)] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}^T(\boldsymbol{m}_{\|i} + \boldsymbol{m}_{\perp i})] \\
&\leq \frac{A\sqrt{B_\|^2 + B_\perp^2}}{\sqrt{n}}
\end{aligned}$$

The result for $\mathcal{R}(\mathcal{G}_{\text{SD}})$ is followed by standard derivations (see [35]). Now we derive the bound on $\mathcal{R}(\mathcal{G}_{\text{TIPMI}})$. This proof follows as a result of $\Pi\boldsymbol{m}_i = \boldsymbol{m}_{i,\|}$ and $(I-\Pi)\boldsymbol{m}_i = 0$.

$$\begin{aligned}
\mathcal{R}(\mathcal{G}_{\text{TIPMI}}) &= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}^T\boldsymbol{m}_i] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i (\Pi\boldsymbol{w}^T\boldsymbol{m}_i + (I-\Pi)\boldsymbol{w}^T\boldsymbol{m}_i)] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i (\boldsymbol{w}_\perp^T\boldsymbol{m}_i + \boldsymbol{w}_\|^T\boldsymbol{m}_i)] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i (\boldsymbol{w}_\perp^T(\Pi\boldsymbol{m}_i + (I-\Pi)\boldsymbol{m}_i) \\
&\quad + \boldsymbol{w}_\|^T(\Pi\boldsymbol{m}_i + (I-\Pi)\boldsymbol{m}_i))] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i (\boldsymbol{w}_\perp^T\boldsymbol{m}_{\|i} + \boldsymbol{w}_\|^T\boldsymbol{m}_{\|i})] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}_\|^T\boldsymbol{m}_{\|i}] \\
&\leq \frac{A \cdot B_\|}{\sqrt{n}}
\end{aligned}$$

Once again, the result for $\mathcal{R}(\mathcal{G}_{\text{TIPMI}})$ is followed by standard derivations (see [35]).

□

Proposition B.3. *For training data $\mathcal{D}$ where $\mathcal{D} \sim P^\circ$, ℓ_g set to the squared loss, and A and $B_\|$ defined in Proposition 2, then*

$$\|\hat{g} - g_0\|^2_{\mathcal{G}_{TIPMI}, P^\circ} \leq \frac{2A \cdot B_\|}{\sqrt{n}} + 5\sqrt{\frac{2\ln(8/\delta)}{n}}$$

Proof. We start by decomposing the teacher estimation error:

$$\begin{aligned}\mathbf{E}[(\hat{g}(\boldsymbol{M}) - Y)^2] = \mathbf{E}[(\hat{g}(\boldsymbol{M}) - g_0(\boldsymbol{M}))^2] \\ + \mathbf{E}[(g_0(\boldsymbol{M}) - Y)^2]\end{aligned}$$

Since ℓ_g is the squared loss, i.e., $\ell_g(g(\boldsymbol{m}), Y) = (g(\boldsymbol{m}) - Y)^2$, we get that

$$\|\hat{g} - g_0\|^2_{\mathcal{G}_{\text{TIPMI}}, P^\circ} = R(\hat{g}(\boldsymbol{M}), Y) - R(g_0(\boldsymbol{M}), Y)$$

Finally, we use standard Rademacher complexity bounds [45]

$$\|\hat{g} - g_0\|^2_{\mathcal{G}_{\text{TIPMI}}, P^\circ} \leq 2\mathcal{R}(\mathcal{G}_{\text{TIPMI}}) + 5\sqrt{\frac{2\ln(8/\delta)}{n}}$$

and apply Proposition 2 to obtain the final result.

□

C Comparing TIPMI and MMD

To demonstrate how TIPMI can have better finite-sample efficiency compared to MMD, we consider a scenario where data is drawn from the unconfounded distribution P°. Under P°, we only need to enforce the signature $f(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{V}$ instead of $f(\boldsymbol{X}) \perp\!\!\!\perp \boldsymbol{V}|Y$, which allows us to expand upon the finite-sample analysis in [32]. In accordance with the DAGs in Figure 1, we assume the data contains two independent binary shortcuts V^1 and V^2, as well as a mediator $\boldsymbol{M}$ that is independent of the shortcuts. Here, we define the MMD function class as:

$$\mathcal{F}_{\text{MMD}} := \{f : \boldsymbol{x} \to \sigma(\boldsymbol{w}^T \boldsymbol{m}), \|\boldsymbol{w}\|_2 \leq A, \text{MMD}(P^\circ_{\phi^i_0}, P^\circ_{\phi^i_1}) \leq \tau_i \;\text{ for } i = 1, 2\}$$

where $f = h(\phi(\boldsymbol{x}))$ and $P^\circ_{\phi^i_v} := P^\circ(\phi(\boldsymbol{X})|V^i = v)$. For a realistic scenario, we assume our training data includes auxiliary labels for only shortcut V_1, such that $\mathcal{D} = \{\boldsymbol{x}_i, v^1_i, y_i\}$ where $\mathcal{D} \sim P^\circ$. Following [32], for each shortcut, we define $\boldsymbol{\Delta}_i : \boldsymbol{\mu}^i_1 - \boldsymbol{\mu}^i_0$ such that $\boldsymbol{\mu}^i_v := \mathbb{E}[\boldsymbol{X}|V^i = v]$. Using $\boldsymbol{\Delta}_i$, we define orthogonal projection matrices for each shortcut where $\Pi^i_\perp := \boldsymbol{\Delta}_i(\boldsymbol{\Delta}_i^T \boldsymbol{\Delta}_i)^{-1}\boldsymbol{\Delta}_i^T$, which projects any vector $\boldsymbol{x}$ onto the shortcut subspace. From here, we can define the projection matrix $\Pi_\perp = \Pi^1_\perp + \Pi^2_\perp$ which projects any vector onto the subspace of all shortcuts. Since $\boldsymbol{X}$ is generated only by independent components V_1, V_2, and $\boldsymbol{M}$, the projection matrix onto the subspace of mediators can be defined as $\Pi := (I - \Pi_\perp)$. Therefore, we can define $\boldsymbol{m}_\parallel := \Pi\boldsymbol{x}$ and $\boldsymbol{m}_{\perp^i} := \Pi^i_\perp \boldsymbol{x}$. Similarly, we define $\boldsymbol{w}_\parallel := \Pi\boldsymbol{w}$ and $\boldsymbol{w}_{\perp^i} := \Pi^i_\perp \boldsymbol{w}$.

Building off of Proposition 5 of [32], we show that the MMD bound can only produce tight finite-sample efficiency bounds if 1) all shortcuts are known and 2) the MMD regularizer can effectively minimize the difference between $P^\circ_{\phi^i_0}$ and $P^\circ_{\phi^i_1}$.

Proposition C.4. *Let $\boldsymbol{m}_\parallel := \Pi\boldsymbol{x}$ and $\boldsymbol{m}_{\perp^i} := \Pi^i_\perp \boldsymbol{x}$. For training data $\mathcal{D} = \{(\boldsymbol{x}_i, v^1_i, y_i)\}^n_{i=1}$ where $\mathcal{D} \sim P^\circ$, also let $\sup_{\boldsymbol{m}_{\perp^i}} \|\boldsymbol{m}_{\perp^i}\|_2 \leq B^i_\perp$, and $\sup_{\boldsymbol{m}_\parallel} \left\|\boldsymbol{m}_\parallel\right\|_2 \leq B_\parallel$. Then*

$$\mathcal{R}(\mathcal{F}_{MMD}) \leq \frac{A \cdot B_\parallel + \tau_1 \frac{B^1_\perp}{\|\Delta_1\|} + A \cdot B_{\perp^2}}{\sqrt{n}}$$

Proof.

$$
\begin{aligned}
\mathcal{R}(\mathcal{G}_{\text{MMD}}) &= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}^T \boldsymbol{x}_i] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i (\Pi \boldsymbol{w}^T \boldsymbol{x}_i + (I-\Pi)\boldsymbol{w}^T \boldsymbol{x}_i)] \\
&= \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\substack{\boldsymbol{w}_\parallel:\left\|\boldsymbol{w}_\parallel\right\|_2\leq A \\ \boldsymbol{w}_{\perp^1}:\left\|\boldsymbol{w}_{\perp^1}\right\|_2\leq A \\ \boldsymbol{w}_{\perp^2}:\left\|\boldsymbol{w}_{\perp^2}\right\|_2\leq A}} \frac{1}{n}\sum_i \epsilon_i (\boldsymbol{w}_\parallel^T \boldsymbol{x}_i + (\boldsymbol{w}_{\perp^1}^T + \boldsymbol{w}_{\perp^2}^T)\boldsymbol{x}_i)] \\
&\leq \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}_\parallel:\left\|\boldsymbol{w}_\parallel\right\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}_\parallel^T \boldsymbol{m}_{i,\parallel}] \\
&+ \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}_{\perp^1}:\left\|\boldsymbol{w}_{\perp^1}\right\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i \boldsymbol{w}_{\perp^1}^T \boldsymbol{m}_{\perp^1,i}] \\
&+ \mathbb{E}_{\mathcal{D}}\mathbb{E}_{\epsilon}[\sup_{\boldsymbol{w}_{\perp^2}:\left\|\boldsymbol{w}_{\perp^2}\right\|_2\leq A} \frac{1}{n}\sum_i \epsilon_i [\boldsymbol{w}_{\perp^2}^T \boldsymbol{m}_{\perp^2,i}] \\
&\leq \frac{A\cdot B_\parallel + \tau_1 \frac{B_\perp^1}{\|\Delta_1\|} + A\cdot B_{\perp^2}}{\sqrt{n}}
\end{aligned}
$$

The first and third terms in the final inequality are from standard derivations (see [35]), and the second term follows directly from Propositions 4 and 5 in [32].

□

From Proposition B.4, we can see that unless all shortcuts are known and $\tau_i = 0$, the finite-sample efficiency bounds for MMD will be dependent on spurious information. In contrast, the TIPMI bounds are only dependent on the mediators and how well the teacher can distill knowledge to the student. This analysis aligns aligns with our empirical findings that show TIPMI is significantly more efficient than MMD.

D Datasets

D.1 Waterbirds

The waterbirds dataset, which was first introduced by [44], uses labeled images of birds and their segmentations from the CUB-200-2011 dataset [49]. If the bird is an Albatross, Auklet, Cormorant, Frigatebird, Fulmar, Gull, Jaeger, Kittiwake, Pelican, Puffin, Tern, Gadwall, Grebe, Mallard, Merganser, Guillemot, or a Pacific Loon, we classify it as a waterbird; every other bird from the dataset we classify as a landbird. We use a subset of the places dataset [56], for the background images. Specifically, we use the 200 land backgrounds and 300 water backgrounds provided by Makar et al. [32]. Similar to Makar et al. [32], we derive other backgrounds from these images by applying rotations, manipulating brightness, and zooming in. The manipulated images are obtained from [32]. In total, our waterbirds dataset is comprised of 8,672 landbirds and 2,483 waterbirds. The bird images, along with the background images, are randomly split so that 80% of the images are used to develop the synthetic training sets, and 20% are used for the testing sets.

The teacher models are trained on the segmented images of birds with a black background, which is the privileged mediation information, as is shown in Figure 4. The student models are trained on images of birds containing either land or water backgrounds. Background and bird images are only used once to generate samples for the training and testing datasets. The size of the images is 256×256 pixels, except for the finite-sample efficiency experiments, which use 128×128. Example images taken from one of the generated waterbirds datasets are shown in Figure 5. For an additional shortcut, we add small black squares randomly placed within an image to simulate camera artifacts. An example of a landbird with this shortcut is given in Figure 6.

(a) Waterbird with no background

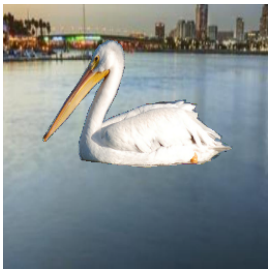
(b) Landbird with no background

Figure 4: An example of the teacher's training data for the waterbirds experiments. It consists of waterbird and landbird images with black backgrounds.

Landbird over land
$y = 0, v = 0$

Waterbird over land
$y = 1, v = 0$

Landbird over water
$y = 0, v = 1$



Waterbird over water
$y = 1, v = 1$

Figure 5: Examples of generated images used in the waterbirds dataset, where the only shortcut is the background.

D.2 Food Review

The food review dataset is created from 20,000 randomly selected reviews from the Amazon Food Review dataset [33]. It contains reviews, the amount of stars (1-5), and a summary of each review. Each synthetic training dataset contains 16,000 samples, whereas the testing set contains 4,000. To inject a shortcut into the reviews, we add perturbation to the words "the" and "be" such that "the"

Figure 6: Example image from waterbirds dataset with two shortcuts. The first shortcut is the background, and the second shortcut is the black squares that simulate camera artifacts.

Figure 7: An example of the teacher and student's training data for the CMNIST experiments. The bottom row of images shows the original CMNIST images that the teacher leverages and the top row shows the images that the student uses, which have been colored to create a spurious correlation.

becomes "thexxxxx" and "be" becomes "bexxxxx", similar to what is done in [48]. For instance, the review "They make the best gummies hands down" is turned into "They make thexxxxx best gummies hands down". We add a similar shortcut by adding perturbations to the words "a" and "to" such that "a" becomes "ayyyyy" and "to" becomes "toyyyyy". The labels are binarized such that $Y = 0$ for 1-3 stars and $Y = 1$ for 4-5 stars. We use the summary as the mediator, which is typically much shorter than the full review.

D.3 Colored MNIST

The colored MNIST dataset is derived from the original MNIST dataset [27], which contains images of handwritten digits 0-9. The dataset contains 60,000 greyscaled training images and 10,000 test images that are 28x28 pixels in size. The Colored MNIST dataset is made by injecting color into the images so that each digit is correlated with one of ten colors. The colors are red, blue, green, brown, purple, tan, cyan, yellow, orange, and pink. The teacher model is trained on the original black and white MNIST images, which are the privileged mediation information, whereas the other models are trained on a colored version of the dataset. Examples of the colored images can be seen on the top row of Figure 7, and the corresponding original MNIST images are shown on the bottom row.

D.4 KOA

The knee osteoarthritis dataset [36] is comprised of knee X-ray images and joint space width measurements from the Osteoarthritis Initiative, which is publicly available for download at https://nda.nih.gov/oai. The knee X-ray images are obtained from the "OAI12MonthImages" sub-dataset. They are stored as a DICOM image and include both of the patient's knees. We convert each image to a PNG and split it in half so that each image only contains a patient's left or right knee. Each image is then centered and resized to 256x256 pixels. The training dataset is comprised of 80% of the total 4982 samples, whereas the testing set is the other 20%. The datasets are split so that one patient's X-rays could only be in one of the two datasets. To determine if a knee in the X-ray had osteoarthritis, we use the Kellgren-Lawrence (KL) grades (0-4) [25] that were provided for each knee. If the knee had a grade of 2, 3, or 4, we classified the sample as having OA (Y=1); if it had a grade of 0 or 1, we classified it as normal (Y=0). For evaluation, we discard all samples with a KL grade

Table 1: The distribution of KL grades in the KOA dataset.

KL Grade	Number of Samples
0	978
1	671
2	1928
3	1082
4	323

of 2 as it is vaguely defined, and KOA diagnoses labeled with this grade are subjective [37]. The distribution of the KL grades is provided in Table 1.

To inject spurious correlations into the dataset, we overlay a black box over the X-rays such that the presence of the black box was correlated with an OA diagnosis. This was done to mimic a metallic token, which may act as a spurious feature in real-world X-ray classification problems [54]. Examples of the spuriously correlated KOA dataset are shown in Figure 8. To evaluate how well TIPMI and the baselines perform when multiple spurious features are present, we also create a dataset with an additional spuriously correlated white box, as is shown in Figure 9. For the privileged mediation information, we use 16 joint space width measurements that are provided with each knee X-ray.

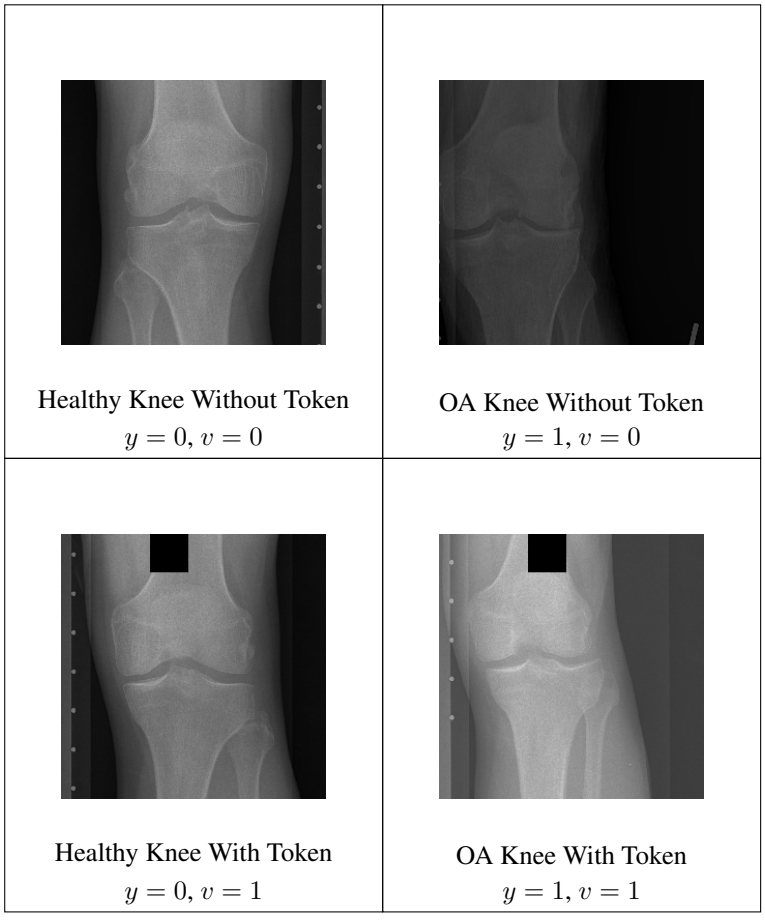


Figure 8: Examples of generated images used in the KOA dataset with a single shortcut.

Figure 9: Example image from KOA dataset with two shortcuts represented by a white and black square spuriously associated with osteoarthritis.

E Experiment Setup

Models. For the waterbirds experiments, both the teacher and the student models are ResNet-50 models [17] pre-trained using the ImageNet-1k dataset [43]. For the KOA experiments, only the student model is a ResNet-50 model pre-trained using the ImageNet-1k dataset, whereas the teacher is a neural network with a single hidden layer with 1024 neurons. Both the student and teacher in the CMNIST experiments are models similar to LeNet-5 27 with additional channels added for the RGB images. For the food review experiments, both the student and teacher models are fine-tuned BERT-tiny models [2]. All models use the cross entropy loss, except for the TIPMI teacher models, which use the mean squared error.

TIPMI Implemenation. The TIPMI teacher was implemented as described in Section 3, "Teacher (Step 1)". However, the student was trained by using the MSE to match the student logits with the teacher logits, not the final probabilities.

Hyperparameter Selection. The hyperparameters for the TIPMI and TIPMI-NC teacher models are chosen through 5-fold cross-validation using the training set. The hyperparameters for the SD, TIPMI/TIPMI-NC students, MMD, GDRO, and L2 models are chosen using cross-validation for each of the 10 simulations. For GDRO and MMD, hyperparameters are chosen based on the highest worst-group accuracy, whereas those with the best overall AUROC are used for the TIMPI/TIPMI-NC, L2, and SD models.

For the waterbirds experiments, we perform cross-validation with a learning rate of $1e^{-5}$ and L_2 penalty parameters with the values $[0, 1e^{-3}, 1e^{-5}]$ for all models except MMD. For MMD, we set the learning rate to $1e^{-5}$, $\sigma = 10$, and cross-validate over $\alpha = [1e0, 1e1, 1e2, 1e3]$. We train each model using Adam [22] and a batch size of 32, except for MMD, which uses a batch size of 256.

For the food review experiments, we perform cross-validation with a learning rate of $1e^{-5}$ and L_2 penalty parameters with the values $[0, 1e^{-3}, 1e^{-5}]$ for all models except MMD. For MMD, we cross-validate over a learning rate of $1e^{-5}$, $\sigma = 10$, and $\alpha = [1e0, 1e1, 1e2, 1e3]$. We train each model using Adam and a batch size of 32, except for MMD, which uses a batch size of 256.

For the KOA experiments, we perform cross-validation with a learning rate of $1e^{-4}$ and L_2 penalty parameters with the values $[0, 1e^{-3}, 1e^{-5}]$ for all models except MMD. For MMD, we set the learning rate to $1e^{-4}$, $\sigma = 10$, and cross-validate over $\alpha = [1e0, 1e1, 1e2, 1e3]$. We train each model using Adam and a batch size of 32, except for MMD, which uses a batch size of 256.

For the CMNIST experiments, we perform cross-validation with a learning rate of $1e^{-3}$ and L_2 penalty parameters with the values $[0, 1e^{-3}, 1e^{-5}]$. We train each model using Adam. Batch sizes used are given for each experiment.

Training Environment. All models used throughout this paper are implemented in PyTorch [38]. We train each model using a Nvidia A40 GPU and 36 GB of memory on a Linux operating system. All experiments took approximately a month with two Nvidia A40 GPUs.

Table 2: CMNIST batch size performance for TIPMI, L2, and GDRO

	AUROC (STE)		
Batch Size	TIPMI	L2	GDRO
1024	**0.985 (0.001)**	0.956 (0.019)	0.103 (0.008)
2048	**0.986 (0.001)**	0.957 (0.006)	0.373 (0.155)
4096	**0.985 (0.001)**	0.948 (0.039)	0.887 (0.039)
8192	**0.982 (0.001)**	0.933 (0.026)	0.974 (0.002)

F Additional Empirical Results

F.1 Batch Size Requirements

We analyze how the batch size affects the performance of models that require overlap. Even when overlap is present, methods such as GDRO require large batch sizes, especially as the number of classes and shortcuts increases. For instance, when both the main label and shortcut label represent 10 classes, the number of groups increases to 100. If there are not enough examples for each group in each batch, GDRO will fail to promote invariance. Table 2 shows the performance of TIPMI, L2, and GDRO with respect to the batch size for the CMNIST dataset. Here, the training distribution is $P(Y = y|\boldsymbol{V} = 1) = P(Y \neq y|\boldsymbol{V} = 0) = 0.95$. At test time, we set $P(Y = 1|\boldsymbol{V} = 1) = P(Y = 0|\boldsymbol{V} = 0) = 0.0$. In Table 2, the performance of TIPMI varies little with respect to batch size, however, GDRO fails to promote invariance without large batch sizes. This limits GDRO and other methods that require overlap to settings with a small number of classes or where large batch sizes are feasible.

F.2 Additional Finite Sample Efficiency Results

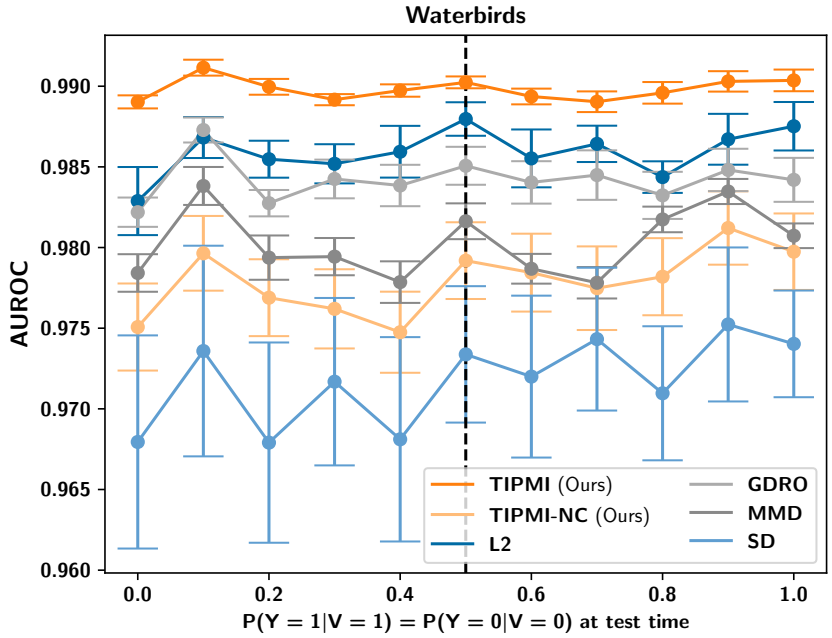


Figure 10: The x-axis shows $P(Y|V)$ at test time under different shifted distributions for a single shortcut, the y-axis shows the AUROC over the test data, and the dashed vertical line shows $P(Y|V)$ at training time. TIPMI outperforms all baselines, showing that TIPMI can be used to increase finite sample efficiency.

F.3 Single Shortcut Performance Results

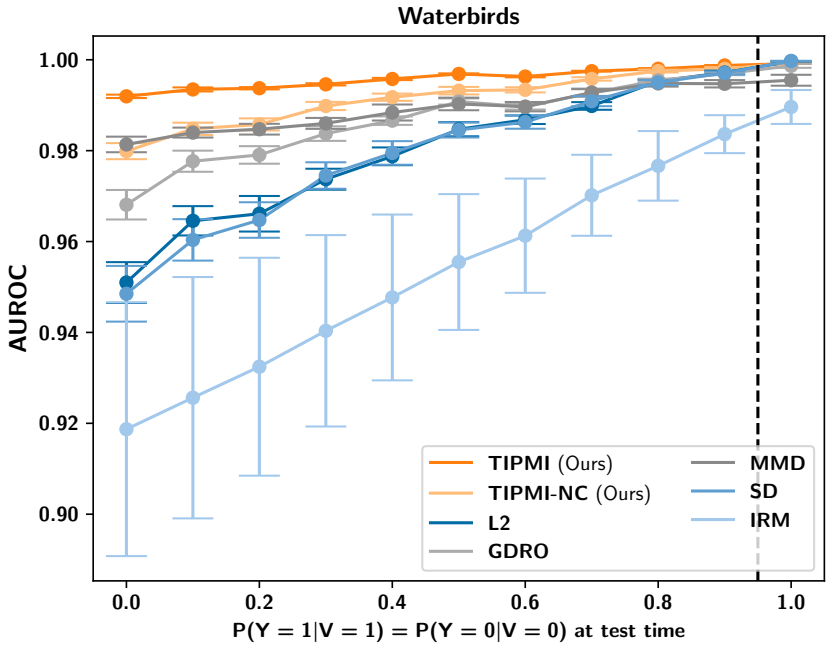


Figure 11: The x-axis shows $P(Y|V)$ at test time under different shifted distributions for a single shortcut, the y-axis shows the AUROC over the test data, and the dashed vertical line shows $P(Y|V)$ at training time. TIPMI outperforms all baselines, showing that TIPMI can be used to effectively promote invariance to single shortcuts.

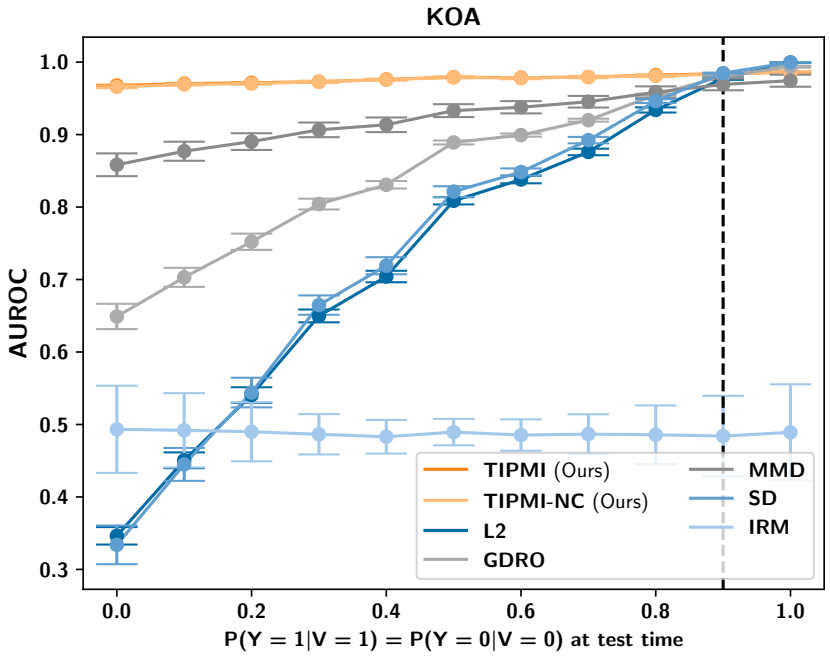


Figure 12: The x-axis shows $P(Y|V)$ at test time under different shifted distributions for a single shortcut, the y-axis shows the AUROC over the test data, and the dashed vertical line shows $P(Y|V)$ at training time. Both TIPMI and TIPMI-NC outperform all baselines, showing that TIPMI can be used to effectively promote invariance to single shortcuts.

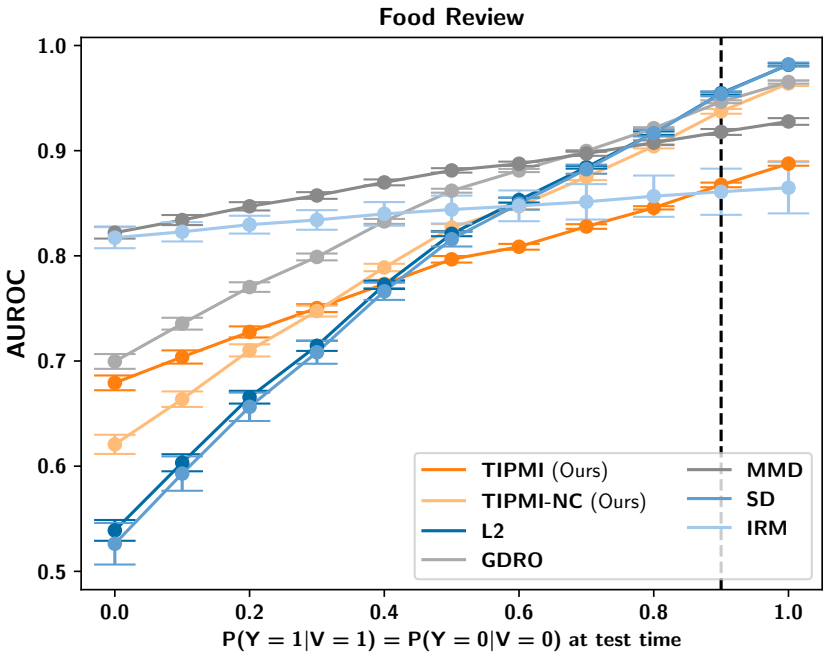


Figure 13: The x-axis shows $P(Y|V)$ at test time under different shifted distributions for a single shortcut, the y-axis shows the AUROC over the test data, and the dashed vertical line shows $P(Y|V)$ at training time. Here, MMD out performs TIPMI. This is due to a violation of our assumptions that the mediator fully mediates the causal effects of Y onto $\boldsymbol{X}$.