Group-Adversarial Transfer for Causal Effect Estimation: Aligning Domains without Compromising Target-Domain Causal Information

Anonymous authors
Paper under double-blind review

Abstract

Estimating average treatment effects (ATEs) is challenging under target-domain data scarcity. Although related source domains can provide additional information, distributional shifts may cause negative transfer. Effective transfer therefore requires reducing cross-domain discrepancies while preserving treatment-effect-relevant information and the treatment–control structure.

We propose GADANet, a transfer causal learning framework that jointly learns representations for causal effect estimation and domain alignment. GADANet decomposes the learned representation into domain-alignment and treatment-effect subspaces and uses group-wise adversarial training to align source and target representations independently within treated and control groups. This design reduces cross-domain discrepancies while preserving treatment-effect-relevant information and avoiding cross-treatment alignment.

We evaluate GADANet across eight simulation settings with varying target-domain sample sizes and distribution shift magnitudes. GADANet consistently achieves lower ATE errors, reducing error by up to 50.8% relative to Baseline, 23.5% relative to Warmstart, and 24.5% relative to Pooling. Ablation studies confirm the contributions of both proposed components.

1 Introduction

To estimate the average treatment effect(ATE) in a target population, we often face a fundamental data-scarcity problem: the target population provides only a limited amount of data, which may be insufficient for reliable estimation of the target ATE (Steingrimsson et al., 2024). Transfer causal learning addresses this problem by leveraging information from a related source domain (Pan & Yang, 2010; Han et al., 2023). However, when target-domain data is scarce and the covariate distributions differ substantially between the source and target domains, distributional mismatch can undermine the effectiveness of knowledge transfer and lead to negative transfer (Wang et al., 2019).

Domain adaptation provides a principled approach to this problem by reducing distributional differences between the source and target domains (Ben-David et al., 2010; Wang et al., 2017). However, unconditional alignment that does not account for the causal structure of treatment assignment may align treated and control subjects across domains, thereby ignoring the treatment-control structure that are essential for accurate target-domain ATE estimation. Effective domain adaptation for causal inference therefore requires reducing domain discrepancies while preserving treatment-effect-relevant information.

To address this challenge, we propose the Group-Adversarial Domain-Adapted Causal Network(GADANet), a general transfer framework that integrates neural causal effect estimation with adversarial domain adaptation. Specifically, GADANet incorporates two key mechanisms. First, it employs orthogonal representation decomposition to separate nuisance and treatment-effect-relevant representations. Second, it uses grouped d omain discriminators to align the source and target domains independently within the treated and control

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

groups. These mechanisms reduce domain discrepancies while preserving the information required for target-domain treatment effect estimation.

Our contributions are summarized as follows:

- **Framework Generality.** We propose a novel transfer causal learning framework that can be integrated into any base causal effect estimation network. Our framework is capable of utilizing source-domain information to compensate for limited target-domain data, while alleviating negative transfer caused by covariate distribution shift. Our key insight is that, when unifying domain adaptation with transfer causal learning, preserving the target-domain causal structure is as important as eliminating the cross-domain distribution discrepancy.
- **Mechanisms.** We employ subspace separation with an orthogonality loss to learn domain-alignment and treatment-effect subspaces. Through adversarial learning, two independent discriminators are adopted to align the covariate distributions separately within each treatment group.
- **Empirics.** We evaluate GADANet across eight simulation settings that vary the target-domain sample size and the magnitude of distribution shift under limited target-domain data. These experiments assess the robustness of GADANet to increasing distribution shifts when target-domain data is scarce. Compared with three baseline models and two ablation models, GADANet consistently achieves lower estimation errors.

2 Related Work

An effective transfer causal learning approach should address two key requirements. First, it should reduce cross-domain discrepancies in nuisance covariates that are irrelevant to treatment effects, thereby reducing the adverse effects of distributional mismatch on target-domain causal effect estimation (Aloui et al., 2023). Second, it should preserve treatment-effect-relevant information and the treatment-control differences that are essential for accurate causal effect estimation. Satisfying both requirements enables the model to leverage source-domain information while preserving the causal information needed in the target domain, providing a principled basis for positive transfer under target-domain data scarcity. Existing approaches achieve neither goal.

Transfer causal learning. Pooling combines source and target-domain data to train a single model, but may become unstable under substantial distribution shifts. Warmstart (Künzel et al., 2018) pretrains on the source domain and fine-tunes on the target domain, but may suffer from residual source-domain bias and negative transfer when target-domain data are limited. More recently, ℓ_1-TCL (Wei et al., 2023) provides non-asymptotic recovery guarantees for generalized linear models, but relies on parametric nuisance models that may be less robust under substantial covariate shifts. Despite these advances, existing approaches do not explicitly align source and target-domain covariate distributions while accounting for target-domain data scarcity, leaving distributional mismatch insufficiently addressed in causal effect estimation.

Domain adaptation. The Domain-Adversarial Neural Network (DANN) (Ganin et al., 2016) employs adversarial training (Goodfellow et al., 2014) by inserting a gradient reversal layer between the representation encoder and a domain discriminator, such that the discriminator learns to distinguish the source and target domains while the encoder is encouraged to produce domain-alignment representations. These ideas have been adapted to causal inference: CFR and TARNet (Shalit et al., 2017) pioneered representation balancing for treatment effect estimation, and HINet (Caljon et al., 2025) employs domain-adversarial training to estimate treatment effects in network settings under covariate shift.

Nevertheless, when naively applied to causal effect estimation, even effective domain adaptation methods such as DANN face a fundamental limitation. On the one hand, they do not distinguish treatment-effect-relevant representations that should be preserved from nuisance representations that should be aligned. On the other hand, their single global discriminator minimizes domain discrepancy without accounting for the treatment–control

structure, which may lead to cross-treatment alignment and weaken treatment-related distinctions needed for accurate causal effect estimation.[1] Although related limitations have been discussed in prior work, existing approaches do not explicitly distinguish between nuisance and treatment-effect-relevant representations during domain adaptation or address the cross-treatment alignment induced by global discriminators.

Positioning of GADANet. This comparison reveals a gap: transfer causal learning methods do not explicitly address distributional shifts under target-domain data scarcity, while domain adaptation methods do not explicitly account for the treatment–control structure or preserve treatment-effect-relevant representations. To address this gap, GADANet integrates orthogonal representation decomposition with group-adversarial domain alignment.

3 Methodology: Group-Adversarial Domain Alignment for Causal Transfer

3.1 Problem Setup

Consider the tuple $(\mathbf{X}_t, T_t, Y_t)$ in the target domain and $(\mathbf{X}_s, T_s, Y_s)$ in the source domain, where $\mathbf{X}_t \in \mathbb{R}^k$ and $\mathbf{X}_s \in \mathbb{R}^k$ denote the pre-treatment covariates with distribution functions $P_t(\mathbf{X})$ and $P_s(\mathbf{X})$, respectively. The random variables $T_t, T_s \in \{0, 1\}$ are the treatment indicators, and Y_t and Y_s are the observed outcomes satisfying

$$Y_t = Y_t(1)\, T_t + (1 - T_t)\, Y_t(0),$$

where $Y_t(1)$ and $Y_t(0)$ are the potential outcomes. Let N_s and N_t denote the numbers of source and target samples, with the observed data denoted by

$$\mathcal{D}_s = \{(\mathbf{X}_i^s, T_i^s, Y_i^s)\}_{i=1}^{N_s}, \qquad \mathcal{D}_t = \{(\mathbf{X}_j^t, T_j^t, Y_j^t)\}_{j=1}^{N_t}.$$

We define the conditional outcome functions

$$\mu_t(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}, T = t], \quad t \in \{0, 1\}.$$

Under the standard unconfoundedness assumption (Pearl, 2009; Imbens & Rubin, 2015), the ATE is identified as:

$$\tau = \mathbb{E}_{\mathbf{X}}\big[\mu_1(\mathbf{X}) - \mu_0(\mathbf{X})\big].$$

In the target domain, both the inner conditional expectations and the outer expectation with respect to $P_t(\mathbf{X})$ must be estimated from finite samples. When N_t is small, estimation of the outer expectation can become a dominant source of error. Provided that the distributional shift from $P_s(\mathbf{X})$ to $P_t(\mathbf{X})$ is appropriately accounted for, source-domain samples can provide additional information for estimating the target-domain expectation, thereby reducing its sampling error (Ben-David et al., 2006; He et al., 2024; Zhang et al., 2013).

3.2 The GADANet Architecture

GADANet is designed as a general framework that can be built upon a broad range of foundational neural architectures, including TARNet and DragonNet (Shalit et al., 2017; Shi et al., 2019). To reduce cross-domain covariate discrepancies, we incorporate the domain discriminator and adversarial training mechanism of DANN (Ganin et al., 2016). To perform domain adaptation while preserving the causal information underlying target-domain treatment effects, GADANet further decomposes the learned representation into two orthogonal subspaces (Bousmalis et al., 2016). Building on this decomposition, we formulate group-specific discriminators to perform domain alignment independently within the treated and control groups. Concretely, GADANet integrates a shared encoder, grouped domain discriminators, and treatment-specific outcome heads into a unified transfer causal learning architecture. The complete pipeline is illustrated in Figure 1.

Shared Encoder. The shared encoder is a multi-layer perceptron that maps the covariates into a learned representation, $\phi : \mathbb{R}^k \to \mathbb{R}^d$, and applies the same parameterized mapping to

[1] This limitation arises because global adversarial training provides the encoder with a direct incentive to reduce domain discrepancy without distinguishing between treatment groups, potentially reducing the treatment–control contrast that is important for causal effect estimation.

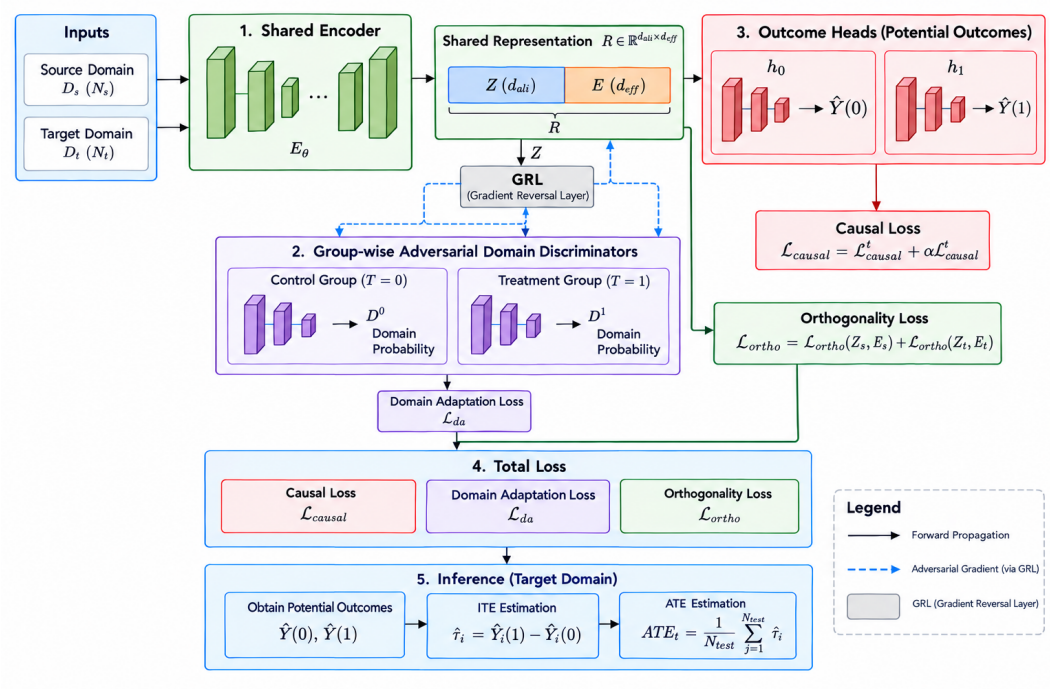

Figure 1: Group-Adversarial Domain-Adapted Causal Network (GADANet). Rather than simply incorporating a domain adaptation module into a conventional causal effect network, GADANet explicitly decomposes the learned representation into orthogonal subspaces and employs grouped discriminators for treatment-conditional domain alignment.

the source covariates $\mathbf{X}_s$ and target covariates $\mathbf{X}_t$. Following the neural causal effect estimation framework of (Shalit et al., 2017), the encoder learns a representation that supports treatment effect estimation. However, directly feeding the entire representation to a single domain discriminator may lead to inappropriate alignment because unconditional adversarial alignment does not distinguish nuisance representations from treatment-effect-relevant representations. It may therefore encourage the encoder to suppress cross-domain differences indiscriminately, including information that is important for target-domain treatment effect estimation.

To address this issue, motivated by the domain-separation framework of (Bousmalis et al., 2016), we decompose the learned representation into two separate subspaces:

$$\phi(\mathbf{X}) \;=\; \big[\, \phi_{\text{ali}}(\mathbf{X}) \,;\, \phi_{\text{eff}}(\mathbf{X}) \,\big] \;\in\; \mathbb{R}^{\,d_{\text{ali}}+d_{\text{eff}}}, \tag{1}$$

where $\mathbf{Z} = \phi_{\text{ali}}(\mathbf{X}) \in \mathbb{R}^{d_{\text{ali}}}$ denotes the domain-alignment subspace, which is provided to the group domain discriminators, and $\phi_{\text{eff}}(\mathbf{X}) \in \mathbb{R}^{d_{\text{eff}}}$ denotes the treatment-effect subspace, which retains treatment-effect-relevant information and is excluded from adversarial alignment. The two subspaces are parameterized by separate branches, so the domain adversarial loss is applied only to $\phi_{\text{ali}}(\mathbf{X})$ and does not directly update the parameters of $\phi_{\text{eff}}(\mathbf{X})$. In contrast, the outcome heads take the full representation $\phi(\mathbf{X})$ as input, allowing treatment effect estimation to use information from both subspaces. Finally, an orthogonality regularization term (Sec. 3.3) is introduced to encourage orthogonality between the two subspaces and promote a clearer separation between domain-alignment and treatment-effect-relevant information.

Grouped Domain Discriminators. A single global domain discriminator aligns the overall covariate distributions of the source and target domains without distinguishing between treatment groups. As a result, it may encourage treated samples in one domain to align with control samples in the other domain, confounding domain discrepancies with treatment-related differences and potentially distorting the representations used by the outcome heads.

To avoid such cross-treatment alignment, we perform treatment-conditional domain alignment by matching the source and target distributions separately within each treatment group. Specifically, for each $t \in \{0,1\}$, we align $P_s(\mathbf{Z} \mid T = t)$ with $P_t(\mathbf{Z} \mid T = t)$. Accordingly, we employ two independent domain discriminators $D^t : \mathbb{R}^{d_{\text{ali}}} \to [0,1]$, where each discriminator receives the domain-alignment representation $\mathbf{Z}$ from its corresponding treatment group and predicts whether the representation comes from the source or target domain.

Gradient Reversal layer. The Shared Encoder and Grouped Domain Discriminators are connected through a Gradient Reversal Layer (GRL), which establishes an adversarial optimization between the two components. The domain discriminators are optimized to distinguish the source and target domains, whereas the encoder is simultaneously encouraged to produce representations that make the two domains less distinguishable. Through this adversarial optimization, the encoder learns domain-alignment representation $\mathbf{Z}$ that reduce cross-domain covariate discrepancies.

Treatment-Specific Outcome Heads. Two separate outcome prediction heads, h_0 and h_1, map the learned representation $\mathbf{R} \in \mathbb{R}^d$ to the two potential outcomes $\hat{Y}(0) = h_0(\mathbf{R})$ and $\hat{Y}(1) = h_1(\mathbf{R})$, respectively. The parameters of h_0 and h_1 are shared across the source and target domains, ensuring that both domains benefit from the same learned outcome mappings.

3.3 Learning Objectives and Joint Optimization

Causal Loss. Taking the Tarnet loss as an example (the Dragonnet objective can be obtained by additionally incorporating a propensity score prediction head), the model jointly minimizes the factual outcome prediction errors in both the source and target domains:

$$\begin{aligned}\mathcal{L}_{\text{causal}} &= \mathcal{L}^s_{\text{causal}} + \alpha \mathcal{L}^t_{\text{causal}} \\ &= \frac{1}{N_s}\sum_{i=1}^{N_s}\Big[Y^s_i - h_{T^s_i}\big(\phi(\mathbf{X}^s_i)\big)\Big]^2 + \alpha \cdot \frac{1}{N_t}\sum_{j=1}^{N_t}\Big[Y^t_j - h_{T^t_j}\big(\phi(\mathbf{X}^t_j)\big)\Big]^2,\end{aligned}$$

where $\alpha = \max(1.0,\ \min(N_s/N_t,\ 3.0))$ is a target-domain reweighting factor that prevents the source-domain causal loss from dominating training when $N_s \gg N_t$.

Orthogonality Loss. The shared representation ϕ is decomposed into two separate subspaces to distinguish domain-alignment representations from treatment-effect-relevant representations, as detailed in Sec. 3.2, Eq. 1. Specifically, the encoder maps the input covariates into a domain-alignment (nuisance) subspace and a treatment-effect subspace. To encourage the two subspaces to capture distinct information, we introduce an orthogonality loss. For a mini-batch of n samples, let $\mathbf{R}_{\text{ali}} \in \mathbb{R}^{n \times d_{\text{ali}}}$ and $\mathbf{R}_{\text{eff}} \in \mathbb{R}^{n \times d_{\text{eff}}}$ denote the domain-alignment and treatment-effect representations, respectively. The orthogonality loss is defined as the squared Frobenius norm of their normalized cross-covariance matrix[2]:

$$\mathcal{L}_{\text{ortho}} = \Big\| \tfrac{1}{n}\, \widetilde{\mathbf{R}}^{\top}_{\text{ali}} \widetilde{\mathbf{R}}_{\text{eff}} \Big\|^2_F = \sum_{p=1}^{d_{\text{ali}}}\sum_{q=1}^{d_{\text{eff}}} \cos^2\Big(\mathbf{R}^{[:,p]}_{\text{ali}},\ \mathbf{R}^{[:,q]}_{\text{eff}}\Big),$$

where the loss is minimized separately for the source and target domains:

$$\mathcal{L}_{\text{ortho}} = \mathcal{L}^s_{\text{ortho}} + \mathcal{L}^t_{\text{ortho}}.$$

Grouped Domain Adversarial Loss. For $g \in \{0,1\}$, let $\mathcal{I}_g = \{i \mid T_i = g\}$ denote the set of sample indices belonging to treatment group g, and further partition this set into source-domain indices $\mathcal{I}^s_g$ and target-domain indices $\mathcal{I}^t_g$. Each group-wise discriminator D^g takes

[2] $\widetilde{\mathbf{R}}_{\text{ali}} = \big(\mathbf{R}_{\text{ali}} - \bar{\mathbf{R}}_{\text{ali}}\big)\,\text{diag}\Big(\|\mathbf{R}_{\text{ali}} - \bar{\mathbf{R}}_{\text{ali}}\|_{\text{col}} + \varepsilon\Big)^{-1}$, $\widetilde{\mathbf{R}}_{\text{eff}} = \big(\mathbf{R}_{\text{eff}} - \bar{\mathbf{R}}_{\text{eff}}\big)\,\text{diag}\Big(\|\mathbf{R}_{\text{eff}} - \bar{\mathbf{R}}_{\text{eff}}\|_{\text{col}} + \varepsilon\Big)^{-1}$, denote their column-wise centered and unit-normalized counterparts, where $\bar{\mathbf{R}}$ is the column-wise mean and $|\cdot|_{\text{col}}$ is the column-wise ℓ_2 norm, and ε is a small positive constant that prevents division by zero.

the domain-alignment subspace representations $\mathbf{Z}$ from its corresponding treatment group and is trained to minimize the binary cross-entropy loss:

$$\mathcal{L}_{\text{da}}^{T=g} = -\frac{1}{|\mathcal{I}_g|}\left(\sum_{i\in\mathcal{I}_g^s}\log\big(1-D^g(\mathbf{Z}_{s,i})\big)+\sum_{j\in\mathcal{I}_g^t}\log D^g(\mathbf{Z}_{t,j})\right),$$

where $\mathbf{Z}_{s,i}=\phi_{\text{ali}}(\mathbf{X}_i^s)$ and $\mathbf{Z}_{t,j}=\phi_{\text{ali}}(\mathbf{X}_j^t)$ are the domain-alignment subspace representations.

Joint Optimization Objective. The overall training objective of GADANet is formulated as a min–max problem. The encoder ϕ minimizes the causal prediction and orthogonality losses, while the outcome heads h_0, h_1 minimize the causal prediction loss. Meanwhile, the encoder maximizes the domain classification loss, whereas the domain discriminators minimize it, encouraging the source and target domains to become less distinguishable within each treatment group. Combining these components yields the following min–max objective:

$$\min_{\phi,h_0,h_1}\max_{D^0,D^1}\ \mathcal{L}_{\text{causal}}^s+\alpha\,\mathcal{L}_{\text{causal}}^t+\lambda_{\text{ortho}}\,\mathcal{L}_{\text{ortho}}-\beta\,\lambda\left(\mathcal{L}_{\text{da}}^{T=0}+\mathcal{L}_{\text{da}}^{T=1}\right), \tag{2}$$

where λ is the gradient-reversal strength applied inside the GRL, λ_{ortho} controls the strength of the orthogonality regularization, and β is the magnitude-calibrated balance factor, which is detailed in Appendix B.1.

Joint Mechanism of the Losses. The three losses in Eq. 2 jointly shape the two representation subspaces. Since the group-wise discriminators operate only on $\mathbf{R}_{\text{ali}}$, the adversarial loss directly affects the alignment branch, while the causal loss is computed from the full representation $[\mathbf{R}_{\text{ali}};\mathbf{R}_{\text{eff}}]$. The orthogonality loss further reduces redundancy between the two subspaces. Consequently, $\mathbf{R}_{\text{ali}}$ is encouraged to capture information that supports domain alignment, whereas $\mathbf{R}_{\text{eff}}$ remains free from direct adversarial alignment and can retain treatment-effect-relevant information. The resulting representation combines group-wise domain alignment with preservation of information relevant to target-domain treatment effect estimation.

4 Model Evaluation

Experimental Setup. We compare GADANet with three baselines: Baseline, which uses only target-domain data, Pooling, and Warm-start. We also consider two ablations: $\text{GADANet}_{\text{GA}}$, which replaces treatment-conditional alignment with global domain alignment, and $\text{GADANet}_{\text{NoOrth}}$, which removes the orthogonality regularization. To evaluate model performance under increasing target-domain data scarcity, its ability to handle shifts in nuisance covariates, and its robustness to negative transfer under general distribution shifts, we design three groups of experimental settings. The first group varies the target-domain sample size under full distribution shift, with all covariates subject to distributional shifts; the second varies the magnitude of nuisance-only shift, and the third varies the magnitude of full distribution shift. [3] Appendix C gives a detailed description of the data-generating processes for all eight configurations. The reported metrics are the absolute ATE, ATT, and ATU errors and the factual RMSE, each averaged over 50 runs and reported as mean ± standard error of the mean. [4]

Performance versus Target Sample Size. Under the full-shift setting, GADANet achieves the lowest ATE error among all baseline methods across both backbones as N_t decreases. Compared with Baseline, GADANet reduces the ATE error by 50.8% across the two backbones,

[3] Sets 1–3 share the full-shift setting and differ only in the target sample size N_t, which decreases from 500 to 100; Sets 4–6 share the nuisance-only-shift setting with $N_t = 100$ and differ only in the magnitude of the distribution shift; Sets 3, 7, and 8 share the same target sample size and differ in the magnitude of the full distribution shift.

[4] ATT and ATU denote the average treatment effect on the treated and the average treatment effect on the untreated, respectively. Factual RMSE is the root-mean-square error between the predicted and observed factual outcomes.

while the corresponding reductions relative to Warmstart and Pooling are 23.5% and 24.5%, respectively. GADANet also achieves the lowest factual RMSE among the three baselines across all target sample sizes. These results indicate that the learned aligned representation supports accurate outcome prediction even when the target domain has limited samples. Trends are shown in Fig. 2, and detailed values are reported in Table 1, 2, 3.

Performance versus Nuisance-only Shift Magnitude. At a fixed target sample size of $N_t = 100$, GADANet remains relatively stable as the magnitude of nuisance-only shift increases, achieves lower ATE error at the largest shift and lower factual RMSE across all settings. Under the largest shift (Set 6), GADANet reduces the ATE error by 10.3% and 64.1% relative to Warmstart, and by 16.6% and 19.0% relative to Pooling, for TARNet and DragonNet, respectively. In contrast, Warmstart becomes substantially less stable under large shifts: its DragonNet ATE error increases to 0.6778 on Set 6, which is 52.0% higher than the Baseline error, indicating substantial negative transfer. These results suggest that adversarial adaptation between the encoder and domain discriminators learns representations with reduced cross-domain covariate discrepancies, thereby mitigating the adverse effects of source-domain distributional mismatch on target-domain causal effect estimation.[5] Trends are shown in Fig. 3, and detailed values are reported in Table 4, 5, 6.

Performance versus Full Shift Magnitude. Figure 4 reports the estimation error as the full-shift magnitude increases from Set 3 to Set 8. GADANet attains the lowest error on all four metrics under both backbones across the three configurations, while its ATE error remains relatively stable as the shift magnitude increases, in contrast to the sharp increases observed for the baselines. Warmstart is particularly unstable on DragonNet, whose ATE error inflates from 0.3268 to 1.5694, even exceeding the non-adaptive Baseline (1.4958) at Set 8. These results suggest that GADANet is more robust to increasing distributional shifts than the transfer baselines. Detailed values are reported in Table 3, 7, 8.

Ablation Study. Across both shift settings, the performance gap between GADANet and its ablations generally increases with the magnitude of distribution shift, with a more pronounced trend under the TARNet backbone. On Set 6, GADANet with TARNet reduces the ATE error by 27.5% relative to $\text{GADANet}_{\text{NoOrth}}$, supporting the effectiveness of orthogonal representation decomposition when the shift primarily affects nuisance covariates. With DragonNet, GADANet reduces the ATE error by 11.0% relative to $\text{GADANet}_{\text{GA}}$, demonstrating the benefit of grouped treatment-conditional alignment. On Set 8, GADANet with TARNet further reduces the ATE error by 11.5% relative to $\text{GADANet}_{\text{NoOrth}}$ and by 11.0% relative to $\text{GADANet}_{\text{GA}}$, providing further evidence for the contributions of both components under substantial full distribution shift. Trends are shown in Figure 5, 6.

5 Conclusion

This work demonstrates the importance of reducing cross-domain distributional discrepancies while preserving the causal information required for treatment effect estimation. Our results show that effective transfer should distinguish between distributional differences induced by domain shift and information that is essential for estimating treatment effects. To address this requirement, GADANet jointly learns representations for causal effect estimation and domain adaptation through orthogonal representation decomposition and group-wise adversarial alignment. The learned representations separate domain-alignment and treatment-effect-relevant information, while group-wise adversarial training reduces distributional discrepancies within treated and control groups. These mechanisms jointly promote domain adaptation while preserving treatment-effect-relevant information, and the resulting framework can be incorporated into a broad range of neural causal effect estimators. Experiments across diverse simulation settings further demonstrate the effectiveness of GADANet for ATE estimation when target-domain data are limited.

[5] Under the nuisance-only shift setting, the source and target domains have approximately the same ATE, so Pooling shows no substantial increase in ATE error as the shift magnitude increases. GADANet consistently achieves lower ATE error and more stable factual RMSE, suggesting that separating nuisance and treatment-effect representations supports stable outcome prediction under nuisance shifts.

A natural question is whether this approach admits a rate-optimality guarantee; we leave a formal analysis to future work. A fundamental limitation of GADANet is that it is primarily designed for covariate shift, where the conditional outcome mechanism $P(Y \mid \mathbf{X}, T)$ remains stable across domains. Under the unconfoundedness assumption, the ATE is identified from the joint distribution characterized by $P(Y \mid \mathbf{X}, T)$ and $P(\mathbf{X})$. GADANet addresses shifts in $P(\mathbf{X})$, but when $P(Y \mid \mathbf{X}, T)$ also varies across domains, the causal outcome mechanism is no longer invariant. In this setting, transferring treatment effects across domains requires additional assumptions to ensure that the target-domain causal effects remain identifiable and transferable.

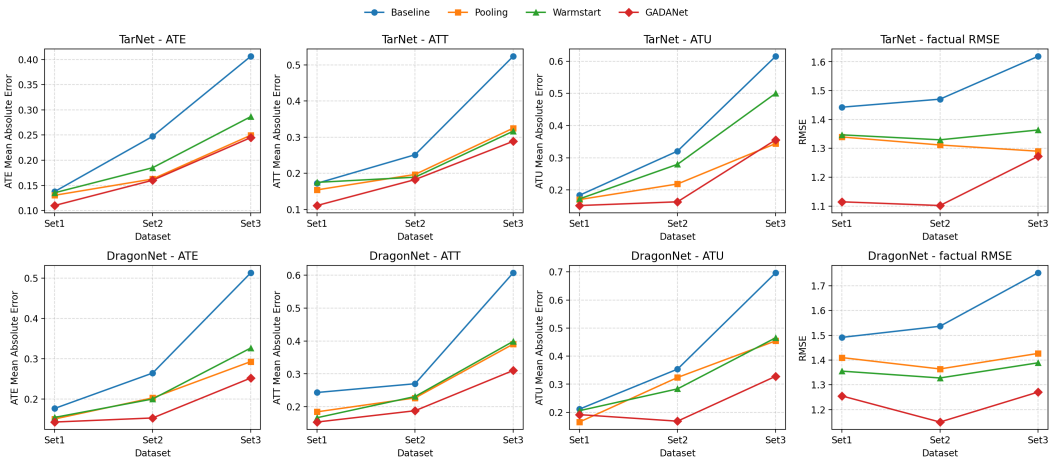

Figure 2: ATE,ATT,ATU and factual RMSE mean absolute error of four methods under TarNet and DragonNet across Set1–Set3 (N_t decreases from 500 to 100). GADANet substantially outperforms the other three models and achieves the lowest factual RMSE.

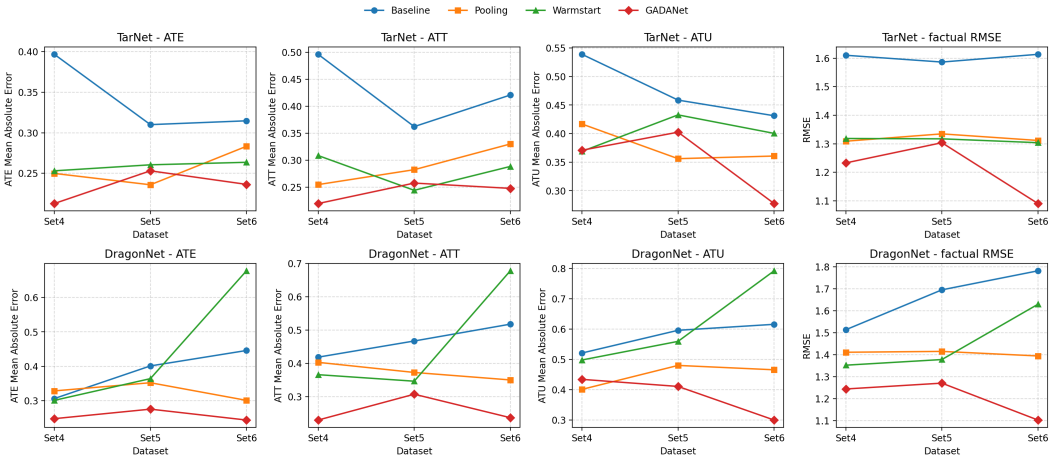

Figure 3: The same four methods evaluated on the nuisance-only shifted datasets in Sets 4–6, with $N_t = 100$ and $\delta \in \{1, 2, 5\}$. GADANet remains relatively stable as δ increases, indicating its robustness to nuisance distribution shifts between the source and target domains.

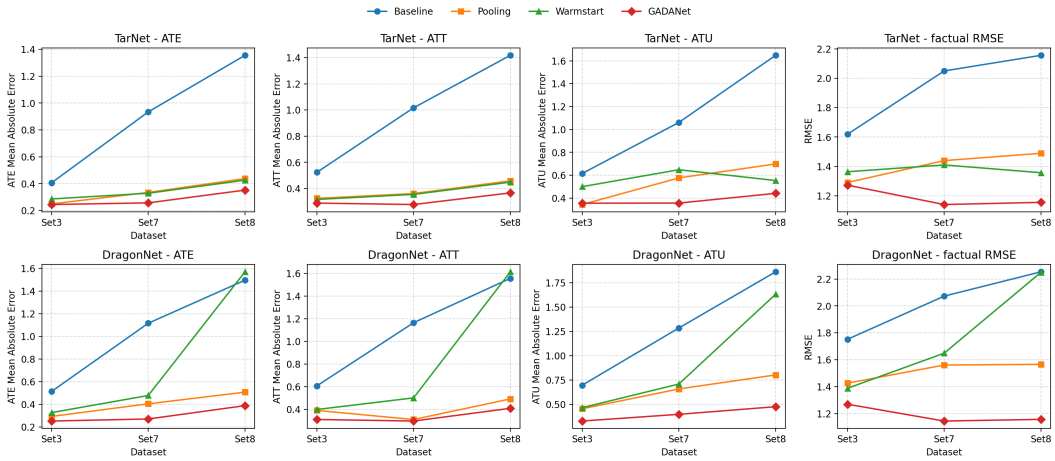

Figure 4: GADANet consistently achieves the lowest and most stable errors across all metrics and both backbones. GADANet maintains an ATE error of approximately 0.39, demonstrating strong robustness to severe distribution shifts.

AI use statement

In this work, we used generative AI tools to improve the language and readability of the text, summarizing existing literature, and searching for relevant information. We have not used generative AI tools for generating synthetic data, or interpreting experimental results. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI tools.

Reproducibility statement

The source code for GADANet, all ablation variants, and the complete datasets used in our experiments are provided in the supplementary material. Per-model results across all 50 independent runs, training configurations, full hyperparameter settings are also documented in Appendix A, Appendix B and Appendix C.

References

Ahmed Aloui, Juncheng Dong, Cat P. Le, and Vahid Tarokh. Transfer learning for individual treatment effect estimation. In Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, volume 216 of Proceedings of Machine Learning Research, pp. 56–66. PMLR, 2023.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In Advances in Neural Information Processing Systems, volume 19, pp. 137–144, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine Learning, 79(1):151–175, 2010.

Ioana Bica and Mihaela van der Schaar. Transfer learning on heterogeneous feature spaces for treatment effects estimation. In Advances in Neural Information Processing Systems, volume 35, pp. 37184–37198, 2022.

Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. In Advances in Neural Information Processing Systems, volume 29, pp. 343–351, 2016.

Daan Caljon, Jente Van Belle, and Wouter Verbeke. Estimating treatment effects in networks under unknown exposure mappings. arXiv preprint arXiv:2510.21457, 2025.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(59):1–35, 2016.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems, volume 27, pp. 2672–2680, 2014.

Larry Han, Zhu Shen, and Jose Zubizarreta. Multiply robust federated estimation of targeted average treatment effects. In Advances in Neural Information Processing Systems, volume 36, pp. 70453–70482, 2023.

Zelin He, Ying Sun, and Runze Li. Transfusion: Covariate-shift robust transfer learning for high-dimensional regression. In International Conference on Artificial Intelligence and Statistics, pp. 703–711. PMLR, 2024.

Guido W. Imbens and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015.

Sören R. Künzel, Bradly C. Stadie, Nikita Vemuri, Varsha Ramakrishnan, Jasjeet S. Sekhon, and Pieter Abbeel. Transfer learning for estimating causal effects using neural networks. arXiv preprint arXiv:1808.07804, 2018.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. IEEE Transactions on Knowledge and Data Engineering, 22(10):1345–1359, 2010.

Judea Pearl. Causality: Models, Reasoning and Inference. Cambridge University Press, 2nd edition, 2009.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In International Conference on Machine Learning, pp. 3076–3085, 2017.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In Advances in Neural Information Processing Systems, volume 32, pp. 2507–2517, 2019.

Jon A Steingrimsson, David H Barker, Ruofan Bie, and Issa J Dahabreh. Systematically missing data in causally interpretable meta-analysis. Biostatistics, 25(2):289–305, 2024.

Jindong Wang, Yiqiang Chen, Shuji Hao, Wenjie Feng, and Zhiqi Shen. Balanced distribution adaptation for transfer learning. In 2017 IEEE international conference on data mining (ICDM), pp. 1129–1134. IEEE, 2017.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime Carbonell. Characterizing and avoiding negative transfer. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11285–11294. IEEE, 2019.

Song Wei, Ronald Moore, Hanyu Zhang, Yao Xie, and Rishikesan Kamaleswaran. Transfer causal learning: Causal effect estimation with knowledge transfer. In ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH), 2023.

Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In International Conference on Machine Learning, pp. 819–827. PMLR, 2013.

A Results and Analysis

Table 1: Experimental results on Set1. GADANet achieves the lowest ATE and factual RMSE under both backbones, with ATE errors of 0.1100 and 0.1433 for TarNet and DragonNet, respectively. Its consistent improvement over $\text{GADANet}_{\text{NoOrth}}$ supports the contribution of orthogonal representation decomposition.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.1376 ± 0.0160	0.1723 ± 0.0200	0.1832 ± 0.0229	1.4426 ± 0.1054
	DragonNet	0.1769 ± 0.0216	0.2434 ± 0.0256	0.2113 ± 0.0204	1.4915 ± 0.1075
Pooling	TarNet	0.1303 ± 0.0139	0.1542 ± 0.0152	0.1699 ± 0.0182	1.3394 ± 0.1075
	DragonNet	0.1508 ± 0.0158	0.1846 ± 0.0216	0.1653 ± 0.0181	1.4099 ± 0.1241
Warmstart	TarNet	0.1353 ± 0.0139	0.1750 ± 0.0165	0.1720 ± 0.0202	1.3467 ± 0.1117
	DragonNet	0.1547 ± 0.0152	0.1667 ± 0.0182	0.2057 ± 0.0204	1.3552 ± 0.0976
$\text{GADANet}_{\text{GA}}$	TarNet	0.1481 ± 0.0189	0.1602 ± 0.0216	0.1895 ± 0.0228	1.2164 ± 0.1775
	DragonNet	0.1814 ± 0.0234	0.2223 ± 0.0262	0.2226 ± 0.0234	1.3754 ± 0.1344
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.1255 ± 0.0148	0.1537 ± 0.0139	0.1655 ± 0.0227	1.2754 ± 0.1718
	DragonNet	0.1520 ± 0.0162	0.1793 ± 0.0185	0.1978 ± 0.0236	1.3703 ± 0.1930
GADANet	TarNet	0.1100 ± 0.0137	0.1109 ± 0.0119	0.1514 ± 0.0189	1.1146 ± 0.1181
	DragonNet	0.1433 ± 0.0153	0.1536 ± 0.0162	0.1917 ± 0.0220	1.2551 ± 0.1640

Table 2: Experimental results on Set2. GADANet consistently outperforms the three baselines in ATE error under both backbones, while achieving the lowest factual RMSE with TarNet (1.1017). Although $\text{GADANet}_{\text{NoOrth}}$ yields a lower TarNet ATE (0.1496), the full model improves factual outcome prediction, supporting the contribution of the orthogonality-based representation decomposition.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.2474 ± 0.0261	0.2514 ± 0.0296	0.3206 ± 0.0342	1.4704 ± 0.1393
	DragonNet	0.2647 ± 0.0266	0.2700 ± 0.0279	0.3537 ± 0.0348	1.5364 ± 0.1280
Pooling	TarNet	0.1627 ± 0.0182	0.1969 ± 0.0180	0.2184 ± 0.0324	1.3116 ± 0.1124
	DragonNet	0.2032 ± 0.0215	0.2269 ± 0.0228	0.3240 ± 0.0274	1.3640 ± 0.1281
Warmstart	TarNet	0.1851 ± 0.0207	0.1893 ± 0.0228	0.2795 ± 0.0277	1.3295 ± 0.1237
	DragonNet	0.2004 ± 0.0236	0.2312 ± 0.0257	0.2834 ± 0.0298	1.3276 ± 0.0863
$\text{GADANet}_{\text{GA}}$	TarNet	0.1597 ± 0.0164	0.1637 ± 0.0205	0.1992 ± 0.0194	1.1342 ± 0.1414
	DragonNet	0.2126 ± 0.0220	0.2238 ± 0.0259	0.2620 ± 0.0292	1.2876 ± 0.1280
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.1496 ± 0.0159	0.1872 ± 0.0187	0.1557 ± 0.0185	1.1322 ± 0.1138
	DragonNet	0.1665 ± 0.0174	0.1820 ± 0.0201	0.1938 ± 0.0188	1.1481 ± 0.1400
GADANet	TarNet	0.1602 ± 0.0170	0.1827 ± 0.0222	0.1630 ± 0.0181	1.1017 ± 0.1167
	DragonNet	0.1534 ± 0.0177	0.1880 ± 0.0208	0.1684 ± 0.0210	1.1496 ± 0.1301

Table 3: Experimental results on Set3 under full distribution shift. GADANet achieves the lowest ATE under both backbones (0.2449 for TarNet and 0.2524 for DragonNet), reducing the Baseline error by 39.7% and 50.8%, respectively.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.4063 ± 0.0491	0.5236 ± 0.0577	0.6152 ± 0.0683	1.6189 ± 0.2913
	DragonNet	0.5131 ± 0.0461	0.6071 ± 0.0604	0.6966 ± 0.0593	1.7522 ± 0.3514
Pooling	TarNet	0.2496 ± 0.0332	0.3249 ± 0.0372	0.3440 ± 0.0479	1.2900 ± 0.2511
	DragonNet	0.2931 ± 0.0289	0.3905 ± 0.0377	0.4544 ± 0.0592	1.4267 ± 0.2349
Warmstart	TarNet	0.2865 ± 0.0331	0.3165 ± 0.0317	0.5002 ± 0.0514	1.3636 ± 0.2199
	DragonNet	0.3268 ± 0.0413	0.3988 ± 0.0426	0.4653 ± 0.0609	1.3888 ± 0.2284
$\text{GADANet}_{\text{GA}}$	TarNet	0.2807 ± 0.0324	0.3391 ± 0.0345	0.3650 ± 0.0488	1.2885 ± 0.2093
	DragonNet	0.2715 ± 0.0281	0.3049 ± 0.0326	0.3816 ± 0.0441	1.2880 ± 0.2185
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.2599 ± 0.0287	0.2976 ± 0.0321	0.3843 ± 0.0453	1.2412 ± 0.2093
	DragonNet	0.2581 ± 0.0296	0.2978 ± 0.0355	0.3639 ± 0.0411	1.2412 ± 0.2284
GADANet	TarNet	0.2449 ± 0.0288	0.2886 ± 0.0370	0.3557 ± 0.0390	1.2722 ± 0.2220
	DragonNet	0.2524 ± 0.0271	0.3104 ± 0.0318	0.3282 ± 0.0432	1.2706 ± 0.2185

Table 4: Experimental results on Set4 under a moderate distribution shift. GADANet achieves the lowest TarNet ATE (0.2128) and outperforms all three baselines with DragonNet (0.2475). Its performance is not consistently better than the two ablations, suggesting that the benefits of grouped alignment and orthogonal representation decomposition are limited when the source-target distribution shift is relatively small.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.3966 ± 0.0407	0.4962 ± 0.0527	0.5389 ± 0.0600	1.6102 ± 0.3330
	DragonNet	0.3060 ± 0.0319	0.4186 ± 0.0435	0.5211 ± 0.0655	1.5135 ± 0.2942
Pooling	TarNet	0.2499 ± 0.0326	0.2549 ± 0.0348	0.4168 ± 0.0437	1.3090 ± 0.2276
	DragonNet	0.3283 ± 0.0346	0.4030 ± 0.0373	0.4013 ± 0.0474	1.4112 ± 0.2249
Warmstart	TarNet	0.2531 ± 0.0352	0.3087 ± 0.0364	0.3689 ± 0.0375	1.3185 ± 0.2051
	DragonNet	0.3008 ± 0.0389	0.3659 ± 0.0517	0.4981 ± 0.0472	1.3525 ± 0.2135
$\text{GADANet}_{\text{GA}}$	TarNet	0.2392 ± 0.0288	0.2607 ± 0.0330	0.3953 ± 0.0433	1.2363 ± 0.2312
	DragonNet	0.2066 ± 0.0275	0.2423 ± 0.0288	0.3813 ± 0.0525	1.2293 ± 0.2355
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.2268 ± 0.0233	0.2446 ± 0.0236	0.3654 ± 0.0409	1.1931 ± 0.2058
	DragonNet	0.2000 ± 0.0273	0.2179 ± 0.0261	0.3993 ± 0.0432	1.1810 ± 0.2150
GADANet	TarNet	0.2128 ± 0.0256	0.2198 ± 0.0261	0.3709 ± 0.0454	1.2329 ± 0.2234
	DragonNet	0.2475 ± 0.0280	0.2295 ± 0.0289	0.4343 ± 0.0512	1.2444 ± 0.2348

Table 5: Experimental results on Set5 under an increased distribution shift. GADANet achieves lower ATE errors than all three baselines with both backbones. Although $\text{GADANet}_{\text{NoOrth}}$ remains competitive and yields lower factual RMSE, the full model shows increasing gains in ATE estimation as the distribution shift becomes stronger.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.3100 ± 0.0331	0.3623 ± 0.0400	0.4587 ± 0.0543	1.5864 ± 0.3451
	DragonNet	0.4010 ± 0.0488	0.4672 ± 0.0514	0.5959 ± 0.0665	1.6947 ± 0.3840
Pooling	TarNet	0.2359 ± 0.0317	0.2825 ± 0.0352	0.3561 ± 0.0363	1.3347 ± 0.2095
	DragonNet	0.3521 ± 0.0348	0.3726 ± 0.0417	0.4802 ± 0.0549	1.4151 ± 0.2652
Warmstart	TarNet	0.2605 ± 0.0350	0.2443 ± 0.0354	0.4328 ± 0.0501	1.3173 ± 0.2319
	DragonNet	0.3640 ± 0.0565	0.3463 ± 0.0511	0.5594 ± 0.0757	1.3781 ± 0.2376
$\text{GADANet}_{\text{GA}}$	TarNet	0.2916 ± 0.0309	0.3217 ± 0.0329	0.4644 ± 0.0554	1.3290 ± 0.2333
	DragonNet	0.2818 ± 0.0293	0.3067 ± 0.0323	0.4549 ± 0.0478	1.3236 ± 0.2440
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.2780 ± 0.0279	0.2824 ± 0.0305	0.3793 ± 0.0434	1.1987 ± 0.2121
	DragonNet	0.2631 ± 0.0297	0.2603 ± 0.0295	0.4287 ± 0.0501	1.1982 ± 0.2234
GADANet	TarNet	0.2531 ± 0.0273	0.2577 ± 0.0291	0.4025 ± 0.0462	1.3037 ± 0.2397
	DragonNet	0.2752 ± 0.0254	0.3072 ± 0.0274	0.4110 ± 0.0436	1.2709 ± 0.2567

Table 6: Experimental results on Set6 under a large distribution shift. GADANet achieves the lowest ATE under both backbones and the lowest TarNet factual RMSE. Its advantage over both ablations becomes more pronounced, suggesting increasing benefits from orthogonal representation decomposition and group-wise alignment under larger shifts. Warmstart exhibits negative transfer with DragonNet, with its ATE increasing to 0.6778 versus 0.4459 for Baseline.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.3147 ± 0.0329	0.4208 ± 0.0504	0.4313 ± 0.0547	1.6135 ± 0.3585
	DragonNet	0.4459 ± 0.0463	0.5178 ± 0.0565	0.6156 ± 0.0630	1.7811 ± 0.4073
Pooling	TarNet	0.2833 ± 0.0360	0.3301 ± 0.0350	0.3607 ± 0.0428	1.3116 ± 0.2096
	DragonNet	0.3008 ± 0.0380	0.3501 ± 0.0385	0.4658 ± 0.0517	1.3944 ± 0.2219
Warmstart	TarNet	0.2636 ± 0.0345	0.2885 ± 0.0365	0.4004 ± 0.0456	1.3038 ± 0.2503
	DragonNet	0.6778 ± 0.1243	0.6782 ± 0.1170	0.7918 ± 0.1344	1.6292 ± 0.5119
$\text{GADANet}_{\text{GA}}$	TarNet	0.2467 ± 0.0290	0.2353 ± 0.0311	0.2992 ± 0.0331	1.0990 ± 0.1945
	DragonNet	0.2737 ± 0.0287	0.2804 ± 0.0301	0.3351 ± 0.0337	1.1035 ± 0.1874
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.3260 ± 0.0324	0.3351 ± 0.0343	0.3355 ± 0.0329	1.1062 ± 0.1966
	DragonNet	0.2777 ± 0.0297	0.2640 ± 0.0325	0.3225 ± 0.0325	1.1137 ± 0.1853
GADANet	TarNet	0.2364 ± 0.0277	0.2479 ± 0.0290	0.2775 ± 0.0304	1.0905 ± 0.1853
	DragonNet	0.2435 ± 0.0284	0.2366 ± 0.0299	0.3003 ± 0.0337	1.1033 ± 0.1916

Table 7: Experimental results on Set7. GADANet attains the lowest errors under both backbones, reducing ATE error by about 72% (TarNet) and 76% (DragonNet) relative to Baseline. Unlike Warmstart, which degrades markedly under DragonNet, GADANet remains stable, and its two ablations consistently underperform the full model, confirming the contribution of both components.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	0.9344 ± 0.1186	1.0168 ± 0.1375	1.0616 ± 0.1683	2.0505 ± 0.8332
	DragonNet	1.1156 ± 0.1334	1.1653 ± 0.1477	1.2853 ± 0.1739	2.0728 ± 0.7601
Pooling	TarNet	0.3349 ± 0.0420	0.3614 ± 0.0439	0.5766 ± 0.0656	1.4394 ± 0.2441
	DragonNet	0.4045 ± 0.0439	0.3102 ± 0.0465	0.6585 ± 0.0687	1.5613 ± 0.2622
Warmstart	TarNet	0.3290 ± 0.0501	0.3553 ± 0.0505	0.6483 ± 0.0730	1.4098 ± 0.2276
	DragonNet	0.4781 ± 0.0676	0.5012 ± 0.0946	0.7090 ± 0.0917	1.6495 ± 0.9193
$\text{GADANet}_{\text{GA}}$	TarNet	0.2846 ± 0.0377	0.3039 ± 0.0378	0.3744 ± 0.0511	1.1283 ± 0.2133
	DragonNet	0.3114 ± 0.0390	0.3182 ± 0.0386	0.4390 ± 0.0541	1.1574 ± 0.2267
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.3130 ± 0.0394	0.3342 ± 0.0379	0.4048 ± 0.0539	1.1438 ± 0.2012
	DragonNet	0.3394 ± 0.0423	0.3524 ± 0.0410	0.4606 ± 0.0615	1.2288 ± 0.2551
GADANet	TarNet	0.2573 ± 0.0380	0.2776 ± 0.0370	0.3569 ± 0.0552	1.1400 ± 0.2137
	DragonNet	0.2706 ± 0.0347	0.2967 ± 0.0347	0.3974 ± 0.0554	1.1457 ± 0.1993

Table 8: On Set8, GADANet again attains the lowest ATE/ATT/ATU errors and factual RMSE, reducing ATE error by about 74% relative to Baseline under both backbones. Notably, Warmstart collapses under DragonNet, whereas GADANet remains stable.

Method		ATE $\mu \pm \sigma$	ATT $\mu \pm \sigma$	ATU $\mu \pm \sigma$	factual RMSE $\mu \pm \sigma$
Baseline	TarNet	1.3556 ± 0.1909	1.4183 ± 0.2000	1.6484 ± 0.2731	2.1570 ± 0.7050
	DragonNet	1.4958 ± 0.1876	1.5573 ± 0.1950	1.8630 ± 0.2635	2.2532 ± 0.6813
Pooling	TarNet	0.4382 ± 0.0482	0.4586 ± 0.0486	0.6989 ± 0.0883	1.4892 ± 0.2068
	DragonNet	0.5072 ± 0.0528	0.4920 ± 0.0541	0.8025 ± 0.0936	1.5664 ± 0.2299
Warmstart	TarNet	0.4263 ± 0.0539	0.4487 ± 0.0535	0.5538 ± 0.0671	1.3569 ± 0.2270
	DragonNet	1.5694 ± 0.2881	1.6150 ± 0.2925	1.6324 ± 0.2703	2.2493 ± 1.2740
$\text{GADANet}_{\text{GA}}$	TarNet	0.3959 ± 0.0502	0.4167 ± 0.0531	0.4363 ± 0.0592	1.1634 ± 0.1661
	DragonNet	0.3625 ± 0.0481	0.3877 ± 0.0515	0.4582 ± 0.0626	1.1777 ± 0.1823
$\text{GADANet}_{\text{NoOrth}}$	TarNet	0.3980 ± 0.0458	0.4156 ± 0.0494	0.4625 ± 0.0571	1.1720 ± 0.1896
	DragonNet	0.3588 ± 0.0458	0.3777 ± 0.0478	0.4453 ± 0.0617	1.1607 ± 0.1765
GADANet	TarNet	0.3523 ± 0.0509	0.3666 ± 0.0518	0.4417 ± 0.0634	1.1549 ± 0.1748
	DragonNet	0.3883 ± 0.0521	0.4091 ± 0.0532	0.4757 ± 0.0671	1.1582 ± 0.1730

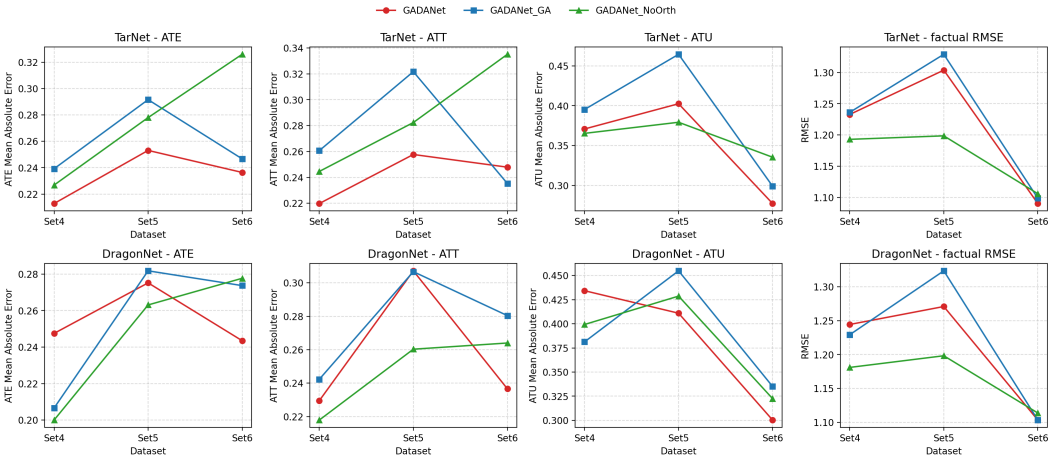

Figure 5: Same comparison on Set4–Set6. GADANet remains the most accurate and stable; the two ablations deteriorate more under distribution shift, with $\text{GADANet}_{\text{NoOrth}}$ showing the largest error at Set6.

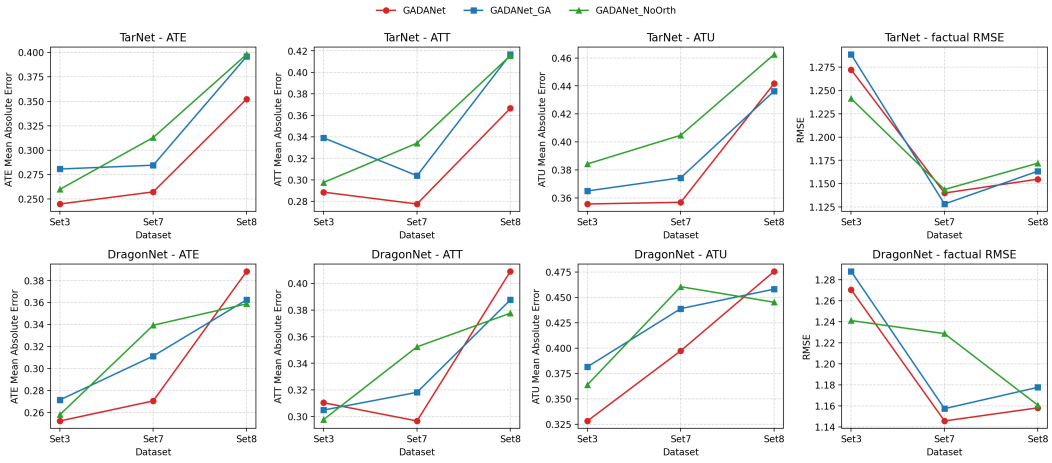

Figure 6: Same comparison on Set3,Set7,Set8. GADANet remains the most accurate and stable. Under Set 8, the increases in ATE, ATT, and ATU are more pronounced for DragonNet than for TARNet, while their factual RMSE values remain comparable. This pattern suggests that the larger degradation in DragonNet may be associated with its propensity-score head, rather than indicating a failure of GADANet's core adaptation mechanism.

B Model Training

Algorithm 1: GADANet Training and Causal Effect Estimation

Input : Source data $\mathcal{D}_s$ (N_s samples); target data $\mathcal{D}_t$ (N_t samples); backbone $\in \{\text{TARNet}, \text{DragonNet}\}$; λ (GRL strength), E (max epochs), λ_{ortho} (orthogonality); d_{ali}, d_{eff} (subspace dims), α (target reweighting); β (loss balancing factor).

Output: $\widehat{\text{ATE}}_t$, $\widehat{\text{ATT}}_t$, $\widehat{\text{ATU}}_t$, factual RMSE.

Split $\mathcal{D}_t$ into training $\mathcal{D}_t^{\text{train}}$ and test $\mathcal{D}_t^{\text{test}}$; $\alpha = \max\big(1.0,\ \min(N_s/N_t,\ 3.0)\big)$;
Initialize alignment branch $\phi_{\text{ali}} : \mathbb{R}^k \to \mathbb{R}^{d_{\text{ali}}}$ and effect branch $\phi_{\text{eff}} : \mathbb{R}^k \to \mathbb{R}^{d_{\text{eff}}}$;
Initialize outcome heads $h_0, h_1 : \mathbb{R}^{d_{\text{ali}}+d_{\text{eff}}} \to \mathbb{R}$;
group discriminators $D^0, D^1 : \mathbb{R}^{d_{\text{ali}}} \to \mathbb{R}^2$;
Initialize β for balancing causal and domain-adaptation losses;
for epoch $= 1$ to E **do**
  for each mini-batch $(\mathbf{X}_s, T_s, Y_s) \sim \mathcal{D}_s$, $(\mathbf{X}_t, T_t, Y_t) \sim \mathcal{D}_t^{\text{train}}$ **do**
    $\mathbf{Z}_s = \phi_{\text{ali}}(\mathbf{X}_s)$, $\mathbf{E}_s = \phi_{\text{eff}}(\mathbf{X}_s)$, $\mathbf{R}_s = [\mathbf{Z}_s; \mathbf{E}_s]$;
    $\mathbf{Z}_t = \phi_{\text{ali}}(\mathbf{X}_t)$, $\mathbf{E}_t = \phi_{\text{eff}}(\mathbf{X}_t)$, $\mathbf{R}_t = [\mathbf{Z}_t; \mathbf{E}_t]$;
    $\hat{Y}_s = h_{T_s}(\mathbf{R}_s)$, $\hat{Y}_t = h_{T_t}(\mathbf{R}_t)$;
    $\mathcal{L}^s_{\text{causal}} = \frac{1}{|\text{batch}|}\sum(Y_s - \hat{Y}_s)^2$, $\mathcal{L}^t_{\text{causal}} = \frac{1}{|\text{batch}|}\sum(Y_t - \hat{Y}_t)^2$;
    $\mathcal{L}_{\text{causal}} = \mathcal{L}^s_{\text{causal}} + \alpha\mathcal{L}^t_{\text{causal}}$;
    $\mathcal{L}_{\text{ortho}} = \mathcal{L}_{\text{ortho}}(\mathbf{Z}_s, \mathbf{E}_s) + \mathcal{L}_{\text{ortho}}(\mathbf{Z}_t, \mathbf{E}_t)$;
    $\mathcal{L}_{\text{da}} = 0$;
    for $g \in \{0, 1\}$ **do**
      if $|\{i \mid T^s_i = g\}| > 0$ and $|\{j \mid T^t_j = g\}| > 0$ **then**
        $\mathbf{Z}^g_s = \text{GRL}_\lambda\big(\mathbf{Z}_s[T_s = g]\big)$, $\mathbf{Z}^g_t = \text{GRL}_\lambda\big(\mathbf{Z}_t[T_t = g]\big)$;
        $\mathbf{Z}^g = \mathbf{Z}^g_s \cup \mathbf{Z}^g_t$;
        $\mathcal{L}_{\text{da}}$ += CrossEntropy$\big(D^g(\mathbf{Z}^g),$ domain labels$\big)$;
      end
    end
    $\beta_t \leftarrow \text{AdaptiveBalance}(\mathcal{L}_{\text{causal}}, \mathcal{L}_{\text{da}})$;
    $\mathcal{L} = \mathcal{L}_{\text{causal}} + \beta_t\mathcal{L}_{\text{da}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}}$;
    Update $\phi_{\text{ali}}, \phi_{\text{eff}}, h_0, h_1, D^0, D^1$ by descending $\nabla\mathcal{L}$;
  end
end
for all $\mathbf{X}^t_j \in \mathcal{D}_t^{\text{test}}$ **do**
  $\mathbf{R}_j = \big[\phi_{\text{ali}}(\mathbf{X}^t_j);\ \phi_{\text{eff}}(\mathbf{X}^t_j)\big]$; $\hat{Y}^t_j(0) = h_0(\mathbf{R}_j)$, $\hat{Y}^t_j(1) = h_1(\mathbf{R}_j)$;
  $\hat{\tau}(\mathbf{X}^t_j) = \hat{Y}^t_j(1) - \hat{Y}^t_j(0)$;
end
$\widehat{\text{ATE}}_t = \frac{1}{N_{\text{test}}}\sum_{j=1}^{N_{\text{test}}}\hat{\tau}(\mathbf{X}^t_j)$;
return $\widehat{\text{ATE}}_t$ and the evaluation metrics.

In practice, the GRL acts as an identity mapping in the forward pass and multiplies the incoming gradients by $-\lambda$ in the backward pass, establishing an adversarial optimization between the encoder and domain discriminators: the encoder is encouraged to produce representations that make the two domains less distinguishable, while the discriminators learn to distinguish between them. To restrict domain alignment to the domain-alignment subspace, the encoder decomposes its representation into two separate branches, ϕ_{ali} and ϕ_{eff}, corresponding to the domain-alignment and treatment-effect subspaces, respectively. The group-wise discriminators operate only on $\phi_{\text{ali}}(\mathbf{X})$, whereas the outcome heads use the full representation $[\phi_{\text{ali}}(\mathbf{X}); \phi_{\text{eff}}(\mathbf{X})]$. An orthogonality loss $\mathcal{L}_{\text{ortho}}$ further encourages the two subspaces to capture distinct information. The causal loss promotes accurate outcome prediction from the full representation, while the domain adversarial loss encourages the alignment subspace to reduce cross-domain discrepancies. Together, these objectives encourage representations that support potential-outcome prediction while reducing domain-specific

discrepancies in the alignment subspace. The balance between the causal and adversarial objectives is controlled by a scale factor β, which is either calibrated before training by matching gradient norms on the alignment branch or updated dynamically using the instantaneous loss ratio with exponential moving-average smoothing. Algorithm 1 summarizes the complete learning procedure of GADANet.

B.1 Hyperparameter

B.1.1 Hyperparameter Selection per Dataset

Table 9 reports the three task-specific hyperparameters of GADANet for each simulated configuration: the gradient-reversal strength λ(grl_lambda), the balance-factor calibration scheme β(balance_method), and the discriminator learning-rate multiplier (disc_lr_multiplier). The remaining hyperparameters are fixed across all configurations, $\lambda_{\text{ortho}} = 0.05$; for all five models on the simulated data, the learning rate and the number of training epochs are set to 0.001 and 300, respectively.

Table 9: Hyperparameter configuration of GADANet for each simulated dataset. N_t is the target-domain sample size; "Full" and "Nuisance-only" denote the full shift $P_s(\mathbf{X}_{\text{em}}) \neq P_t(\mathbf{X}_{\text{em}}), P_s(\mathbf{X}_{\text{nui}}) \neq P_t(\mathbf{X}_{\text{nui}})$ and the nuisance-only shift $P_s(\mathbf{X}_{\text{nui}}) \neq P_t(\mathbf{X}_{\text{nui}})$, respectively; δ is the shift magnitude.

Config.	N_t	Shift	δ	λ	balance_method	disc_lr_mult
Set 1	500	Full	1	3	gradient_norm	1
Set 2	300	Full	1	5	gradient_norm	1
Set 3	100	Full	1	5	auto_beta	5
Set 4	100	Nuisance-only	1	3	auto_beta	1
Set 5	100	Nuisance-only	2	5	auto_beta	5
Set 6	100	Nuisance-only	5	8	gradient_norm	1
Set 7	100	Full	4	6	auto_beta	1
Set 8	100	Full	6	8	auto_beta	1

The gradient-reversal strength λ grows with the shift magnitude and with data scarcity. Within the nuisance-only series (Sets 4–6), λ increases monotonically from 3 to 8 as δ grows from 1 to 5: a larger covariate shift requires a stronger adversarial pull to align the nuisance subspace. Within the full-shift series (Sets 1–3), λ increases from 3 to 5 as N_t shrinks from 500 to 100; within the full-shift series (Set3, Set7,Set8), λ increases from 5 to 8.

The balance-factor scheme trades off stability and adaptivity. The static gradient-norm calibration is used when the target sample is large (Sets 1–2) or the scenario is extreme (Set 6 with $\delta = 5$), where a fixed β is more reliable than a dynamically fluctuating one. The dynamic auto_beta scheme is used for the intermediate regime ($N_t = 100$, $\delta \leq 2$), where the loss magnitudes evolve smoothly and an exponentially smoothed β can track them.

The discriminator learning rate is coupled with the balance scheme. When auto_beta is used (Sets 3 and 5), the discriminator learning rate is increased fivefold, since β depends on the magnitude of the domain loss, which in turn requires the discriminator to converge rapidly to a discriminative state. Under gradient-norm calibration, β is computed from natural gradient norms that are independent of the discriminator's training state, so the discriminator uses the base learning rate. Set 4 is the exception: despite using auto_beta, the smallest shift ($\delta = 1$) makes the domain easily separable and thus needs no accelerated discriminator.

B.1.2 Magnitude Calibration of the Balance Factor

The domain adversarial loss is typically several orders of magnitude smaller than the causal loss and must be rescaled to a comparable magnitude. We provide two complementary mechanisms for calibrating the balance factor β.

Dynamic loss-ratio balancing. At every mini-batch, a raw balance factor is first computed from the instantaneous loss ratio:

$$\beta^{\text{raw}} = \frac{\mathcal{L}^{s}_{\text{causal}} + \alpha\, \mathcal{L}^{t}_{\text{causal}}}{\mathcal{L}^{T=0}_{\text{da}} + \mathcal{L}^{T=1}_{\text{da}} + \varepsilon}, \quad \varepsilon = 10^{-8},$$

where both adversarial losses are computed without the GRL to estimate their magnitudes. The raw factor is then clipped and rescaled,

$$\beta^{\text{raw}} \leftarrow \text{clip}\big(\beta^{\text{raw}};\ \beta_{\min},\ \beta_{\max}\big) \cdot \beta_{\text{scale}},$$

and finally smoothed with an exponential moving average to dampen mini-batch fluctuations,

$$\beta_t = \rho\, \beta_{t-1} + (1-\rho)\, \beta^{\text{raw}}_t,$$

where ρ is the EMA momentum (default $\rho = 0.9$), $\beta_{\min} = 0.1$, $\beta_{\max} = 50.0$, and $\beta_{\text{scale}} = 0.2$ are hyperparameters.

Gradient-norm calibration. As an alternative, β is estimated once before training by matching the gradient norms of the causal and domain losses on the nuisance branch ϕ_{inv}. Over K stratified mini-batches, each of which guarantees non-empty source/target cells within every treatment group, the mean gradient norms

$$\bar{g}_{\text{causal}} = \frac{1}{K}\sum_{k=1}^{K} \big\|\nabla_{\theta_{\text{inv}}} \mathcal{L}^{(k)}_{\text{causal}}\big\|_2, \qquad \bar{g}_{\text{domain}} = \frac{1}{K}\sum_{k=1}^{K} \big\|\nabla_{\theta_{\text{inv}}} \mathcal{L}^{(k)}_{\text{da}}\big\|_2,$$

are computed with respect to the nuisance-branch parameters θ_{inv}, and the balance factor is set to

$$\beta = \text{clip}\left(\frac{\bar{g}_{\text{causal}}}{\bar{g}_{\text{domain}} + \varepsilon};\ 0.05,\ 50.0\right),$$

which is then kept fixed throughout training.

C Description of Data Shift Configuration

The simulated data follow a dual-domain covariate-shift design in which a source population and a target population are drawn from the same structural model but with different covariate distributions. Each observation consists of twelve covariates $\mathbf{X} \in \mathbb{R}^{12}$, composed of six binary and six continuous variables, partitioned into two groups according to their causal roles:

- Effect modifiers (confounders) $\mathbf{X}_{\text{em}} \in \mathbb{R}^7$ (three binary and four continuous variables) drive the treatment assignment, the baseline outcome, and the treatment effect $\tau(\cdot)$;
- Nuisance covariates $\mathbf{X}_{\text{nui}} \in \mathbb{R}^5$ (three binary and two continuous variables) contribute only weakly to the baseline outcome μ_0 and do not influence the treatment effect.

Binary covariates are generated through a Gaussian copula with correlation $\rho_{\text{bin}} = 0.2$ and marginal probability 0.5; continuous covariates are drawn from a zero-mean, unit-variance multivariate Gaussian with correlation $\rho_{\text{con}} = 0.3$. Given the covariates, the treatment is assigned by a logistic propensity model

$$\text{logit}\, P(T = 1 \mid \mathbf{X}) = \kappa\Big(\boldsymbol{\beta}^{\top}\mathbf{X}_{\text{em}} + \psi(\mathbf{X}_{\text{em}})\Big),$$

where $\kappa = 0.5$ is the treatment strength, $\boldsymbol{\beta}$ collects the linear coefficients, and $\psi(\cdot)$ is a nonlinear term. The potential outcomes are

$$\mu_0(\mathbf{X}) = \boldsymbol{\alpha}^{\top}\mathbf{X}_{\text{em}} + \boldsymbol{\gamma}^{\top}\mathbf{X}_{\text{nui}}, \mu_1(\mathbf{X}) = \mu_0(\mathbf{X}) + \tau(\mathbf{X}_{\text{em}}),$$

and the observed outcome is

$$Y = \mu_0(\mathbf{X}) + \tau(\mathbf{X}_{\text{em}})\, T + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0,1).$$

Crucially, the conditional treatment effect $\tau(\cdot)$ depends only on the effect modifiers $\mathbf{X}_{\text{em}}$; this property is what distinguishes the two shift regimes described next.

Domain Shift Mechanisms. The shift is applied by perturbing the marginal distribution of the target-domain covariates while keeping the conditional mechanisms $P(T \mid \mathbf{X})$ and $P(Y \mid \mathbf{X}, T)$ identical across domains. This is therefore a covariate (marginal) shift, not a conditional shift. For continuous covariates the shift is a mean translation

$$\mathbf{X}_t^{(j)} = \mathbf{X}_s^{(j)} + \delta\, \Delta^{(j)},$$

and for binary covariates it is a shift in the Bernoulli probability

$$P_t(X^{(j)} = 1) = \text{clip}\Big(P_s(X^{(j)} = 1) + \delta\, \Delta_{\text{bin}}^{(j)};\ 0.15,\ 0.85\Big),$$

where δ is the shift magnitude (domain_shift) and $\Delta^{(j)}$, $\Delta_{\text{bin}}^{(j)}$ are per-variable shift coefficients. Two regimes are considered.

Full shift. Both the effect modifiers and the nuisance covariates shift,

$$P_s(\mathbf{X}_{\text{em}}) \neq P_t(\mathbf{X}_{\text{em}}), \qquad P_s(\mathbf{X}_{\text{nui}}) \neq P_t(\mathbf{X}_{\text{nui}}).$$

Because $\tau(\cdot)$ is a function of $\mathbf{X}_{\text{em}}$, the target ATE $\mathbb{E}_{P_t}[\tau(\mathbf{X}_{\text{em}})]$ differs from the source ATE $\mathbb{E}_{P_s}[\tau(\mathbf{X}_{\text{em}})]$ even though the function $\tau(\cdot)$ itself is unchanged.

Nuisance-only shift. Only the nuisance covariates shift,

$$P_s(\mathbf{X}_{\text{nui}}) \neq P_t(\mathbf{X}_{\text{nui}}), \qquad P_s(\mathbf{X}_{\text{em}}) = P_t(\mathbf{X}_{\text{em}}).$$

Since τ depends solely on $\mathbf{X}_{\text{em}}$, whose marginal distribution is unchanged, the source and target ATEs coincide; the shift perturbs only the nuisance structure available for domain alignment. In both regimes the shift is marginal: the conditional distributions $P(T \mid \mathbf{X})$ and $P(Y \mid \mathbf{X}, T)$ are identical across domains, and the observed downstream changes (e.g., the ATE difference under the full shift) are induced solely by the change in $P(\mathbf{X})$.

Table 10 summarizes the generator parameters and their values used throughout the experiments.

Table 10: Generator parameters and their values used in the experiments.

Parameter	Value	Description
n_{binary}, $n_{\text{continuous}}$	6, 6	Number of binary / continuous covariates
ρ_{bin}, ρ_{con}	0.2, 0.3	Within-type covariate correlation
κ (treatment_strength)	0.5	Treatment propensity strength
effect_type	heterogeneous	Functional form of $\tau(\cdot)$
$\Delta_{\text{con}}^{\text{em}}$	1.0	Mean-shift coefficient, continuous effect modifiers
$\Delta_{\text{bin}}^{\text{em}}$	0.2	Probability-shift coefficient, binary effect modifiers
$\Delta_{\text{con}}^{\text{nui}}$	2.0	Mean-shift coefficient, continuous nuisance covariates
$\Delta_{\text{bin}}^{\text{nui}}$	0.3	Probability-shift coefficient, binary nuisance covariates
N_s	1000	Source-domain sample size
N_t	100–500	Target-domain sample size
δ (domain_shift)	1, 2,4,5,6	Shift magnitude
R (n_replications)	50	Independent replications per configuration