STRUCTURALLY COMPATIBLE GENERATIVE AMBIGUITY SETS VIA EXOGENOUS-SPACE FACTORIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distributionally robust optimization (DRO) critically depends on the ambiguity set used to characterize plausible distribution shifts. Recent generative approaches parameterize adversarial distributions with trainable generative models, enabling flexible shifts beyond the empirical support, but they do not explicitly preserve known causal structure. In contrast, structural causal optimal transport restricts admissible distributions according to a structural causal model (SCM), but is not naturally formulated as a trainable generative adversary. We bridge these perspectives with exoDRO, which constructs generative ambiguity sets in the exogenous space of a bijective SCM. Rather than perturbing the endogenous distribution directly, we independently parameterize the exogenous marginals and push their product distribution through the fixed structural map. This construction guarantees structural compatibility for every adversarial parameter value by design. We further show that the inclusive KL divergence between the nominal and adversarial endogenous distributions decomposes exactly into the sum of the corresponding exogenous marginal KL divergences. This identity holds for any bimeasurable bijective structural map and does not depend on causal depth, Jacobian volume preservation, or additive-noise structure. Combined with likelihood-based generative consistency bounds, this identity yields a global reconstruction-based certificate and finite-sample control using a single dataset shared across all exogenous coordinates. To optimize the resulting coupled adversarial problem, we develop a cyclic block-coordinate scheme whose fixed-block subproblems reduce to single-model generative-DRO inner problems. Experiments verify the predicted structural properties and show that, on the predictive benchmark, exoDRO achieves the lowest prediction error across the evaluated nominal and held-out OOD settings compared with classical, generative, and structure-aware DRO baselines.

1 INTRODUCTION

Distributionally robust optimization (DRO) seeks decision rules that remain reliable under distribution shift by optimizing against the worst-case distribution within an ambiguity set around a nominal distribution P_0:

$$w^\star = \arg\min_{w\in\mathcal{W}} \sup_{Q\in\mathcal{B}(P_0)} \mathbb{E}_{X\sim Q}[f(w,X)]. \tag{1}$$

The choice of ambiguity set $\mathcal{B}(P_0)$ is therefore central: if it is too broad, robust optimization can become unnecessarily conservative, whereas if it excludes plausible shifts, the resulting robustness guarantee may have limited practical relevance.

Recent work has addressed this problem from two complementary directions. First, generative ambiguity modeling represents adversarial distributions using trainable generative models (Wen & Yang, 2026; 2025). This provides a flexible finite-dimensional parameterization of the inner adversary and allows it to generate shifts beyond the empirical support. Related approaches parameterize adversarial distributions or likelihood ratios within DRO (Michel et al., 2021; 2022), while generative models have also been used to construct adversarial environments in Wasserstein-based robust learning (Ren & Majumdar, 2022). These approaches, however, do not in general guarantee that the generated distributions remain compatible with known causal mechanisms.

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.

A complementary line of work incorporates structural information directly into optimal transport and ambiguity-set design (Veraguas et al., 2017; Cheridito & Eckstein, 2025). In particular, structural causal optimal transport characterizes distributions compatible with fixed structural equations through factorization in the exogenous space (Ehyaei et al., 2025). This provides a principled notion of structurally plausible distribution shift, but the resulting ambiguity sets are formulated through constrained transport plans rather than a trainable generative parameterization. This leaves open how to construct a trainable generative adversary whose entire parameterized family remains structurally compatible throughout optimization.

Key idea. We bridge these perspectives by constructing a generative ambiguity set directly in the exogenous space of a bijective structural causal model (SCM). We refer to the resulting framework as *exoDRO* (exogenous DRO). Let $U = (U_1, \dots, U_n)$ denote mutually independent exogenous variables and let $g : \mathcal{U} \to \mathcal{X}$ denote the reduced-form structural map, such that

$$X = g(U), \qquad P_0 = g_{\#} \left(\bigotimes_{i=1}^{n} P_{U_i} \right).$$

Rather than parameterizing a single adversarial distribution directly over the endogenous variables, we learn an adversarial marginal P_{θ_i} for each exogenous coordinate and define

$$Q_{\boldsymbol{\theta}} := g_{\#} \left(\bigotimes_{i=1}^{n} P_{\theta_i} \right). \tag{2}$$

Because the adversarial distribution remains factorized in the exogenous space for every value of $\boldsymbol{\theta}$, structural compatibility is guaranteed by construction rather than encouraged through an auxiliary penalty or imposed through a projection step.

A key consequence of this parameterization is an exact information-geometric identity. For any bimeasurable bijective structural map g, we show that

$$D_{\mathrm{KL}} \left(P_0 \,\|\, Q_{\boldsymbol{\theta}} \right) = \sum_{i=1}^{n} D_{\mathrm{KL}} \left(P_{U_i} \,\|\, P_{\theta_i} \right). \tag{3}$$

Thus, the endogenous-space KL discrepancy decomposes exactly into coordinatewise exogenous contributions, regardless of how deeply a perturbation propagates through the causal graph. The identity requires neither additive-noise structural equations nor a volume-preserving structural map; bijectivity is sufficient.

This decomposition connects naturally to likelihood-based generative ambiguity modeling. Existing consistency results provide coordinatewise bounds of the form

$$D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i}) \leq J_i(\theta_i) + c_i,$$

where J_i is a reconstruction or negative-log-likelihood objective and c_i is independent of the adversarial parameter (Wen & Yang, 2026). Combining these bounds with Equation 3 yields a global reconstruction-based certificate in the endogenous space. We further establish finite-sample control when the marginal objectives are estimated from a common dataset and develop a cyclic block-coordinate optimization scheme for the resulting coupled adversarial problem.

Contributions. Our main contributions are:

- We introduce an exogenous-space generative ambiguity set for a bijective SCM, such that every generated adversarial distribution is structurally compatible with the fixed mechanisms by construction.
- We establish an exact decomposition of the inclusive KL divergence between nominal and adversarial endogenous distributions into a sum of exogenous marginal divergences, yielding a global reconstruction-based KL certificate.
- We derive finite-sample control for empirically estimated reconstruction constraints and develop a cyclic block-coordinate optimization procedure whose fixed-block subproblems reduce to single-model generative-DRO inner problems.
- We evaluate the proposed ambiguity construction through controlled structural analyses and comparisons with classical, generative, and structure-aware DRO baselines under distribution shift.

2 Related Work

Classical distributionally robust optimization. Classical DRO constructs ambiguity sets around a nominal distribution using statistical divergences or optimal-transport distances. KL-DRO considers distributions within a KL-divergence neighborhood (Hu & Hong, 2013), whereas Wasserstein DRO uses transport-based ambiguity sets and admits principled adversarial-training formulations (Sinha et al., 2018).

Generative ambiguity modeling. Recent approaches parameterize the adversary in DRO using trainable generative models, allowing the ambiguity set to represent shifts beyond the empirical support. GAS-DRO constrains likelihood-based generative models through reconstruction objectives and optimizes the resulting adversary via dual learning and policy optimization (Wen & Yang, 2026); its diffusion-based predecessor develops a related construction for score-based models (Wen & Yang, 2025). Earlier parameterized DRO methods model adversarial distributions or likelihood ratios directly (Michel et al., 2021; 2022), while generative models have also been used to construct adversarial environments under Wasserstein-based robustness objectives (Ren & Majumdar, 2022). These approaches provide flexible adversarial parameterizations, but do not in general enforce compatibility with a known structural causal model. In contrast, our construction places the generative parameterization in the exogenous space so that structural compatibility follows directly from the model construction.

Causal and structural optimal transport. Adapted and causal optimal transport restrict admissible couplings according to temporal or graphical information (Veraguas et al., 2017; Cheridito & Eckstein, 2025). Structural causal optimal transport further incorporates the structural equations and characterizes compatible distributions through factorization after mapping to the exogenous space (Ehyaei et al., 2025). Our framework adopts this exogenous-space notion of structural compatibility, but differs in how the ambiguity set is represented: rather than optimizing over constrained transport plans, we parameterize independent exogenous marginals with trainable generative models and push their product distribution through the fixed structural map.

Additional connections to structured and factorized ambiguity sets, as well as entropic and Sinkhorn-regularized DRO, are discussed in Appendix A.

3 Problem Setup

3.1 Bijective Structural Causal Model

Let $X = (X_1, \dots, X_n) \in \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ be generated by a structural causal model $\mathcal{M} = \langle G, \mathcal{X}, F, \mathcal{U}, P_U \rangle$, where G is a directed acyclic graph, $F = \{f_i\}_{i=1}^n$ denotes the structural mechanisms, and $U = (U_1, \dots, U_n) \in \mathcal{U}$ contains the exogenous variables.

We assume causal sufficiency, so that the exogenous distribution factorizes as

$$P_U = \bigotimes_{i=1}^{n} P_{U_i}. \tag{4}$$

We further assume that the structural equations induce a bimeasurable bijection

$$g : \mathcal{U} \to \mathcal{X}, \qquad X = g(U).$$

The nominal endogenous distribution is therefore

$$P_0 = g_{\#} P_U. \tag{5}$$

This assumption is aligned with the bijective-generation viewpoint used in prior work on causal models and structural causal transport (Nasr-Esfahany et al., 2023; Ehyaei et al., 2025). Additive-noise SCMs under a known causal ordering provide an important example, although the results below do not require the structural equations themselves to be additive.

3.2 Structural Compatibility

Following Ehyaei et al. (2025), we characterize compatibility with the fixed structural mechanisms through factorization in the exogenous space.

Definition 1 (Structural compatibility). *A distribution $Q \in \mathcal{P}(\mathcal{X})$ is compatible with the fixed structural mechanisms F if*

$$(g^{-1})_{\#}Q = \bigotimes_{i=1}^{n} \widetilde{P}_i \qquad \text{for some } \widetilde{P}_i \in \mathcal{P}(\mathcal{U}_i). \tag{6}$$

We denote the set of all such distributions by

$$\mathcal{P}_F(\mathcal{X}) := \left\{ Q \in \mathcal{P}(\mathcal{X}) : \exists\, \widetilde{P}_i \in \mathcal{P}(\mathcal{U}_i) \text{ for } i = 1, \dots, n \text{ such that } (g^{-1})_{\#}Q = \bigotimes_{i=1}^{n} \widetilde{P}_i \right\}. \tag{7}$$

The structural map g is fixed, while the exogenous marginal distributions may change. Thus, a distribution can differ from P_0 while remaining compatible with the same structural mechanisms.

3.3 Generative Marginal Models

For each exogenous coordinate U_i, let $P_{\theta_i} \in \mathcal{P}(\mathcal{U}_i)$ denote the distribution of a likelihood-based generative model with parameters θ_i. Following the generative ambiguity framework of Wen & Yang (2026), let $J_i(\theta_i) := J(\theta_i, P_{U_i})$ denote its population reconstruction objective.

Our analysis uses the following per-coordinate consistency property.

Assumption 1 (Per-coordinate generative consistency). *For every $i \in [n]$, there exists a constant c_i, independent of θ_i, such that*

$$D_{\mathrm{KL}}\left(P_{U_i} \,\|\, P_{\theta_i}\right) \le J_i(\theta_i) + c_i. \tag{8}$$

For the likelihood-based models analyzed by Wen & Yang (2026), the constant c_i collects terms that do not depend on the adversarial parameter θ_i. The structural results below do not depend on the particular form of J_i; Assumption 1 is needed only when the exact KL identity is converted into a reconstruction-based bound.

4 Structural Generative Ambiguity Sets

We now define the proposed exogenous-space generative ambiguity set and characterize its structural and distributional properties.

4.1 Exogenous-Space Generative Ambiguity Set

For a vector of reconstruction budgets $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n) \in \mathbb{R}_+^n$, we define the structural generative ambiguity set as follows.

Definition 2 (Structural generative ambiguity set).

$$\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) := \left\{ g_{\#}\left(\bigotimes_{i=1}^{n} P_{\theta_i} \right) : J_i(\theta_i) \le \epsilon_i, \quad i = 1, \dots, n \right\}. \tag{9}$$

The corresponding DRO problem is

$$\min_{w \in \mathcal{W}} \sup_{\theta_1, \dots, \theta_n} \mathbb{E}_{U \sim \otimes_{i=1}^{n} P_{\theta_i}} \left[f(w, g(U)) \right] \quad \text{s.t.} \quad J_i(\theta_i) \le \epsilon_i \ \ \forall i. \tag{10}$$

The key distinction from a single generative adversary defined directly in the endogenous space is the location of the parameterization. The adversarial distribution is factorized where the SCM assumes independence—in the exogenous variables—and the resulting product distribution is subsequently mapped to the endogenous space through the fixed structural mechanism g.

4.2 Structural Compatibility by Construction

The exogenous factorization immediately ensures compatibility with the fixed structural mechanisms.

Proposition 1 (Automatic structural compatibility)**.** *Every distribution in $\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$ belongs to $\mathcal{P}_F(\mathcal{X})$:*

$$\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) \subseteq \mathcal{P}_F(\mathcal{X}). \tag{11}$$

Indeed, for $Q_{\boldsymbol{\theta}} = g_{\#}(\otimes_i P_{\theta_i})$, bijectivity of g gives $(g^{-1})_{\#} Q_{\boldsymbol{\theta}} = \otimes_i P_{\theta_i}$, which satisfies Definition 1. Thus, structural compatibility holds throughout adversarial optimization, without an additional causal regularizer or projection step. The formal proof is given in Appendix B.1.

4.3 Exact KL Decomposition

The exogenous parameterization also yields an exact characterization of the discrepancy between the nominal and adversarial distributions. The key observation is that KL divergence is invariant under a bimeasurable bijection:

$$D_{\mathrm{KL}}(g_{\#}\mu \,\|\, g_{\#}\nu) = D_{\mathrm{KL}}(\mu \,\|\, \nu).$$

A formal statement and proof are provided in Appendix B.2.

Combining this invariance with the additivity of KL divergence across product measures gives our central result.

Theorem 1 (Exact exogenous KL decomposition)**.** *Let*

$$P_0 = g_{\#}\left(\bigotimes_{i=1}^{n} P_{U_i}\right), \qquad Q_{\boldsymbol{\theta}} = g_{\#}\left(\bigotimes_{i=1}^{n} P_{\theta_i}\right),$$

and assume $P_{U_i} \ll P_{\theta_i}$ for every i. Then

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) = \sum_{i=1}^{n} D_{\mathrm{KL}}\left(P_{U_i} \,\|\, P_{\theta_i}\right). \tag{12}$$

Theorem 1 shows that a perturbation of an upstream exogenous variable does not acquire an additional KL penalty simply because it propagates through multiple downstream mechanisms. The equality depends only on the bijectivity of g and the factorization of the exogenous distributions; it requires neither $|\det J_g| = 1$ nor additive-noise structural equations. The complete proof is given in Appendix B.3.

4.4 Reconstruction-Based Global Certificate

Theorem 1 connects the structural parameterization directly to generative ambiguity modeling. Combining it with Assumption 1 yields the following global certificate.

Corollary 1 (Additive generative consistency bound)**.** *For any $Q_{\boldsymbol{\theta}}$ induced by the exogenous generative models,*

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) \leq \sum_{i=1}^{n} \left[J_i(\theta_i) + c_i\right]. \tag{13}$$

Consequently, every $Q_{\boldsymbol{\theta}} \in \mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$ satisfies

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) \leq \sum_{i=1}^{n} (\epsilon_i + c_i). \tag{14}$$

The equality in Theorem 1 is exact, whereas the inequality above inherits the approximation slack of the underlying generative consistency bound.

4.5 Relation to a Structural KL Ambiguity Set

To clarify the distributional neighborhood represented by our construction, define

$$\mathcal{B}_F^{\mathrm{KL}}(P_0, \rho) := \{Q \in \mathcal{P}_F(\mathcal{X}) : D_{\mathrm{KL}}(P_0 \| Q) \le \rho\}. \tag{15}$$

Proposition 2 (Containment in a structural KL ball)**.** *Let*

$$\rho(\boldsymbol{\epsilon}) := \sum_{i=1}^{n} (\epsilon_i + c_i).$$

Then

$$\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) \subseteq \mathcal{B}_F^{\mathrm{KL}}\left(P_0, \rho(\boldsymbol{\epsilon})\right). \tag{16}$$

This is a containment rather than an equality: the generative ambiguity set is additionally restricted by the expressiveness of the marginal generative models and by reconstruction-based constraints. It should therefore be viewed as a trainable parameterization of a subset of a structurally compatible KL neighborhood defined by $D_{\mathrm{KL}}(P_0 \| Q)$. The proof is given in Appendix B.5.

5 Finite-Sample Control

The ambiguity set in Definition 2 is defined through the population objectives $J_i(\theta_i)$, whereas in practice each marginal generative model is constrained using an empirical reconstruction objective. A useful consequence of the exogenous formulation is that these coordinatewise objectives do not require separate datasets. Given N endogenous observations $\{X^{(k)}\}_{k=1}^{N}$, bijectivity of g yields $U^{(k)} = g^{-1}(X^{(k)})$, providing N observations for every exogenous coordinate simultaneously.

Let $\widehat{\mathcal{B}}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$ denote the ambiguity set obtained by replacing the population objectives with their empirical counterparts. Under the bounded-loss uniform-convergence condition stated in Appendix C, simultaneous control of the n marginal objectives gives the following finite-sample guarantee.

Corollary 2 (Finite-sample endogenous KL guarantee)**.** *With probability at least* $1 - \delta$*, every* $Q_{\boldsymbol{\theta}} \in \widehat{\mathcal{B}}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$ *satisfies*

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) \le \sum_{i=1}^{n} \left[\epsilon_i + \Delta_i(N, \delta/n) + c_i\right], \tag{17}$$

where $\Delta_i(N, \delta/n)$ *is the generalization tolerance for the* i*th reconstruction objective.*

Importantly, the same N observations are used for all exogenous coordinates, so separate per-coordinate datasets are not required. Simultaneous control introduces the usual confidence adjustment through δ/n—equivalently a $\log(n/\delta)$ dependence for standard bounded-loss bounds—while the resulting KL certificate remains additive across coordinates. The precise empirical ambiguity set, uniform-convergence assumption, simultaneous coordinatewise bound, and the empirical–population ambiguity-set sandwich underlying Corollary 2 are given in Appendix C.

6 Block-Coordinate Optimization

Although the exogenous factorization makes the ambiguity constraints coordinatewise, the robust objective is generally not separable:

$$\mathbb{E}_{U \sim \otimes_{i=1}^{n} P_{\theta_i}}\left[f(w, g(U))\right]$$

couples all adversarial marginals through the downstream loss. We therefore optimize the exogenous marginal models using cyclic block-coordinate updates.

For fixed decision parameter w, the inner problem is

$$\sup_{\theta_1, \ldots, \theta_n} \; \mathbb{E}_{U \sim \otimes_{i=1}^{n} P_{\theta_i}}\left[f(w, g(U))\right] \quad \text{s.t.} \quad J_i(\theta_i) \le \epsilon_i \;\; \forall i. \tag{18}$$

Introducing multipliers $\mu_i \ge 0$ gives the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\mu}; w) = \mathbb{E}_{U \sim \otimes_i P_{\theta_i}}\left[f(w, g(U))\right] - \sum_{i=1}^{n} \mu_i \big(J_i(\theta_i) - \epsilon_i\big). \tag{19}$$

The constraint terms are separable across coordinates, whereas the expected robust loss is not.

Single-block reduction. Fix coordinate i and hold the remaining adversarial parameters θ_{-i} fixed. Define the averaged block reward

$$\widetilde{f}_i(w, u_i; \theta_{-i}) := \mathbb{E}_{U_{-i} \sim \otimes_{j \neq i} P_{\theta_j}} \left[f\big(w, g(U_{-i}, u_i)\big) \right]. \tag{20}$$

Up to terms independent of θ_i, the coordinate-i update reduces to

$$\sup_{\theta_i} \mathbb{E}_{U_i \sim P_{\theta_i}} \left[\widetilde{f}_i(w, U_i; \theta_{-i}) \right] - \mu_i J_i(\theta_i). \tag{21}$$

For fixed θ_{-i}, Equation 21 takes the same form as the single-model generative-DRO inner problem studied by Wen & Yang (2026), with nominal distribution P_{U_i} and reward $\widetilde{f}_i$. This motivates a cyclic procedure that updates one exogenous adversary at a time while holding the remaining blocks fixed.

Proposition 3 (Per-block optimization guarantee)**.** *Fix w, coordinate i, and θ_{-i}. If the coordinate-i subproblem is optimized using the single-model inner-maximization procedure analyzed by Wen & Yang (2026), and its corresponding assumptions hold for $\widetilde{f}_i(w, \cdot; \theta_{-i})$ and J_i, then the blockwise inner-maximization criterion inherits the same $\mathcal{O}(K^{-1/2})$ convergence rate as the single-model problem.*

The guarantee is conditional on the remaining blocks being fixed. It therefore does not imply global convergence of the full cyclic nonconvex problem, since updating one marginal changes the effective reward faced by subsequent blocks. Full update equations, policy-gradient and Monte Carlo estimators, dual optimization, pseudocode, and the proof of Proposition 3 are provided in Appendix D.

7 Experiments

We evaluate the proposed framework from two complementary perspectives. First, we compare predictive robustness against representative classical, generative, and structure-aware DRO baselines under controlled distribution shifts. Second, we use synthetic SCMs with known structural mechanisms to directly examine the structural properties of the proposed ambiguity construction. Full experimental protocols and additional results are provided in Appendix E.

7.1 Experimental Setup

Predictive robustness benchmark. We compare exoDRO with empirical risk minimization (ERM), Wasserstein DRO (W-DRO) (Sinha et al., 2018), KL-DRO (Hu & Hong, 2013), GAS-DRO (Wen & Yang, 2026), and structural causal optimal transport (hereafter SCOT) (Ehyaei et al., 2025). All methods use the same nominal training data, prediction architecture, OOD validation/test environments, and training-seed assignments. Method-specific adversarial parameterizations and optimization procedures follow the corresponding formulations.

Distribution shift is introduced exclusively through the exogenous distributions while keeping the structural mechanisms fixed. We construct three levels of shift severity—mild, moderate, and strong—with 10 held-out test environments at each level, yielding 30 OOD test environments in total. A separate set of OOD validation environments is used for hyperparameter and checkpoint selection; the held-out test environments are used only for final evaluation. Thus, the evaluation measures generalization to held-out environments from the specified exogenous-shift family, rather than robustness to arbitrary unseen shift mechanisms.

Each method is trained using five independent training seeds. For each severity level, performance is aggregated across its 10 test environments, and we report the mean and standard deviation across training seeds. We additionally report worst-environment OOD performance. Detailed implementation settings and model-selection procedures are provided in Appendix E.1. Code and configuration files are available in an anonymous repository.

Controlled structural analyses. To isolate the theoretical properties of the proposed ambiguity set, we additionally consider the three-variable causal chain

$$A := U_A, \qquad E := \alpha A + U_E, \qquad I := \beta E + U_I,$$

with $U_A, U_E, U_I \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$. Its analytically invertible structural map allows us to examine structural compatibility and the KL decomposition without confounding the analysis with SCM-estimation or inversion error.

Table 1: Predictive performance under distribution shift. Results are reported as mean ± standard deviation over five training seeds. Lower is better.

Method	ID MSE ↓	Mild OOD ↓	Moderate OOD ↓	Strong OOD ↓	Worst OOD ↓
ERM	0.0921 ± 0.0003	0.0937 ± 0.0004	0.0946 ± 0.0004	0.0973 ± 0.0010	0.1025 ± 0.0023
W-DRO	0.0926 ± 0.0002	0.0943 ± 0.0003	0.0950 ± 0.0004	0.0974 ± 0.0004	0.1023 ± 0.0003
KL-DRO	0.0921 ± 0.0004	0.0937 ± 0.0006	0.0945 ± 0.0005	0.0972 ± 0.0006	0.1022 ± 0.0015
SCOT	0.0906 ± 0.0005	0.0917 ± 0.0003	0.0918 ± 0.0002	0.0925 ± 0.0007	0.0952 ± 0.0009
GAS-DRO	0.0921 ± 0.0003	0.0936 ± 0.0002	0.0945 ± 0.0003	0.0972 ± 0.0008	0.1023 ± 0.0024
exoDRO	$\mathbf{0.0904 \pm 0.0002}$	$\mathbf{0.0916 \pm 0.0002}$	$\mathbf{0.0914 \pm 0.0002}$	$\mathbf{0.0909 \pm 0.0003}$	$\mathbf{0.0943 \pm 0.0003}$

Table 2: Finite-sample HSIC diagnostics on recovered exogenous residuals for a representative run ($n = 1{,}500$, 300 permutations, $\alpha = 0.7$, $\beta = -0.6$). The structural and joint adversaries have approximately matched realized inclusive KL.

Adversary	Test	HSIC	p-value
exoDRO	$\widehat{U}_E \perp A$	0.000430	0.0365
exoDRO	$\widehat{U}_I \perp (A, E)$	0.000202	0.5847
Joint generative	$\widehat{U}_E \perp A$	0.001558	0.0033
Joint generative	$\widehat{U}_I \perp (A, E)$	0.001156	0.0033

7.2 Predictive Robustness under Distribution Shift

Table 1 compares predictive performance under the nominal distribution and held-out distribution shifts of varying severity. Since W-DRO, KL-DRO, GAS-DRO, and SCOT define their ambiguity sets using different geometries, their robustness hyperparameters are selected separately using the designated OOD validation environments, rather than enforcing a common numerical radius across methods.

exoDRO has the lowest mean prediction error across all evaluated settings, including both the nominal distribution and held-out OOD environments. Among the implemented baselines, the SCOT reference yields the lowest mean prediction error across the reported settings, consistent with the potential benefit of incorporating structural information into the ambiguity set. Nevertheless, exoDRO is numerically lower than SCOT across all reported settings. The improvement is particularly pronounced under the strong OOD setting, where exoDRO reduces MSE from 0.0925 to 0.0909, corresponding to a 1.7% relative reduction. For the worst-case OOD metric, exoDRO achieves an MSE of 0.0943, compared with 0.0952 for SCOT.

In contrast, W-DRO, KL-DRO, and GAS-DRO remain close to ERM across the evaluated shifts. These results suggest that explicitly modeling structural variation in the exogenous space can yield more effective robust training than unstructured ambiguity modeling over the observed distribution. Complementary analyses based on environment-matched excess risk and coordinatewise exogenous shifts show the same trend: exoDRO achieves the lowest excess error and the smallest degradation under every single-coordinate shift as well as under the joint exogenous shift. Full results are reported in Appendices E.2 and E.3.

7.3 Structural Compatibility

The recovered exogenous residuals provide a finite-sample diagnostic of the structural independences implied by the SCM. We compare the proposed structural generative adversary with an unconstrained joint generative adversary at approximately matched realized inclusive KL divergence. For samples generated by either adversary, we recover

$$\hat{U}_E = E - \alpha A, \qquad \hat{U}_I = I - \beta E.$$

Under the structural construction, the corresponding population independences are $\hat{U}_E \perp A$ and $\hat{U}_I \perp (A, E)$. Finite-sample residual dependence is assessed using HSIC with permutation-based tests.

Table 2 shows substantially smaller HSIC values for exoDRO than for the unconstrained joint adversary in both relations. Because the structural independence is a population property enforced by the parameterization, we treat the permutation tests as finite-sample diagnostics rather than as the basis for the structural guarantee. In the representative run, one exoDRO permutation p-value falls below 0.05 despite the small HSIC value. A separate 10-seed repetition of $\hat{U}_E \perp A$ yields a mean p-value of 0.546, with 0/10 seeds below 0.05; the full results are reported in Appendix E.7.

7.4 Exact KL decomposition

We test Theorem 1 by perturbing exogenous variables at different causal depths and comparing $D_{\mathrm{KL}}(P_0 \| Q_{\boldsymbol{\theta}})$ with the sum of the corresponding marginal exogenous KL divergences. Across all 20 evaluated combinations of structural configuration and perturbation location, the ratio

$$r_{\mathrm{KL}} := \frac{D_{\mathrm{KL}}(P_0 \| Q_{\boldsymbol{\theta}})}{\sum_i D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i})}$$

is 1.0000 up to numerical precision. A separate k-nearest-neighbor estimate computed directly in the endogenous space provides an independent numerical cross-check. Full results are given in Appendix E.6.

8 Discussion and Limitations

Structural and statistical scope. Our analysis assumes a fixed, known, and bijective structural map g and mutually independent exogenous variables, enabling exact recovery of the exogenous coordinates. Estimating g, latent confounding, non-invertible mechanisms, and mechanism shifts $F \to F'$ are outside the present scope; when g is non-injective, the exact KL identity generally weakens to a data-processing inequality. Accordingly, our experiments use known invertible SCMs and do not model causal-discovery or mechanism-estimation errors. Further extensions, including possible blockwise treatments of dependent exogenous variables, are discussed in Appendix F.

Finite-sample scope. The finite-sample result controls coordinatewise empirical reconstruction constraints from a shared dataset rather than providing dimension-free concentration for the endogenous distribution itself. We assume bounded per-sample reconstruction losses; extensions to heavier-tailed settings require additional assumptions.

Optimization and comparison scope. Proposition 3 provides a conditional per-block guarantee and does not establish global convergence of the coupled nonconvex problem. Because reconstruction-, KL-, and Wasserstein-based ambiguity sets use different discrepancy geometries, their numerical robustness radii are not directly comparable; we therefore select method-specific hyperparameters independently using the same OOD-validation criterion. Additional limitations are discussed in Appendix F.

9 Conclusion

We introduced a generative ambiguity-set construction that parameterizes adversarial distributions in the exogenous space of a bijective structural causal model. By independently modeling the exogenous marginals and pushing their product distribution through the fixed structural map, the proposed construction guarantees structural compatibility for every adversarial parameter value by design. This parameterization yields an exact decomposition of the inclusive KL divergence between nominal and adversarial endogenous distributions into the sum of exogenous marginal KL divergences. Combined with likelihood-based generative consistency bounds, this identity provides a global reconstruction-based certificate and finite-sample control from a common dataset. We further showed that the coupled adversarial problem admits a cyclic block-coordinate optimization scheme whose fixed-block subproblems reduce to single-model generative-DRO inner problems. Our experiments examine both robustness under distribution shift and the structural properties predicted by the analysis, demonstrating that structural compatibility can be incorporated into generative ambiguity modeling by choosing a parameterization in which the desired structure is satisfied by design.

AI USE STATEMENT

In this work, we used generative AI tools to assist with proofreading and language editing, including improving readability and identifying typographical or formatting issues. We also used these tools for code review and debugging, and for limited feedback on experimental design and the interpretation of results. All substantive research decisions, method development, experiments, and final interpretations were made and verified by the authors. All AI-assisted code and text were reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Detailed assumptions and proofs are provided in Appendices B–D. Appendix E contains the experimental protocols, implementation details, hyperparameter and model-selection procedures, data-generation settings, and additional results used in our evaluation. Source code and configuration files, including the settings required to reproduce the reported experiments, are provided in the anonymous repository: `https://anonymous.4open.science/r/ExoDRO-Exogenous-space-DRO-EBCF/`.

REFERENCES

Eslam Abdelaleem, K. Michael Martini, and Ilya Nemenman. Accurate estimation of mutual information in high dimensional data. *arXiv preprint arXiv:2506.00330*, 2025.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. Mine: Mutual information neural estimation. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 531–540. PMLR, 2018.

Jose Blanchet, Daniel Kuhn, Jiajin Li, and Bahar Taskesen. Unifying Distributionally Robust Optimization via Optimal Transport Theory. 12 2025. URL http://arxiv.org/abs/2308.05414.

Lotfi M. Chaouach, Dimitris Boskos, and Tom Oomen. Uncertain uncertainty in data-driven stochastic optimization: towards structured ambiguity sets. In *Proceedings of the IEEE Conference on Decision and Control*, volume 2022-December, pp. 4776–4781. Institute of Electrical and Electronics Engineers Inc., 2022. ISBN 9781665467612. doi: 10.1109/CDC51059.2022.9992405.

Patrick Cheridito and Stephan Eckstein. Optimal transport and wasserstein distances for causal models. *Bernoulli*, 31(2):1351–1376, 2025. doi: 10.3150/24-BEJ1773.

Bao Duong and Thin Nguyen. Diffeomorphic Information Neural Estimation. 11 2022. URL http://arxiv.org/abs/2211.10856.

Ahmad-Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. Designing Ambiguity Sets for Distributionally Robust Optimization Using Structural Causal Optimal Transport. 10 2025. URL http://arxiv.org/abs/2510.00599.

Zhaolin Hu and L. Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Optimization Online*, 2013.

David McAllester and Karl Stratos. Formal Limitations on the Measurement of Mutual Information. 5 2020. URL http://arxiv.org/abs/1811.04251.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Modeling the Second Player in Distributionally Robust Optimization. 3 2021. URL http://arxiv.org/abs/2103.10282.

Paul Michel, Tatsunori Hashimoto, and Graham Neubig. Distributionally Robust Models with Parametric Likelihood Ratios. 4 2022. URL http://arxiv.org/abs/2204.06340.

Arash Nasr-Esfahany, Mohammad Alizadeh, and Devavrat Shah. Counterfactual Identifiability of Bijective Causal Models. 6 2023. URL http://arxiv.org/abs/2302.02228.

Allen Z. Ren and Anirudha Majumdar. Distributionally Robust Policy Learning via Adversarial Environment Generation. 7 2022. URL http://arxiv.org/abs/2107.06353.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Quang Huy Tran, Hicham Janati, Ievgen Redko, Rémi Flamary, and Nicolas Courty. Factored couplings in multi-marginal optimal transport via difference of convex programming. 12 2021. URL http://arxiv.org/abs/2110.00629.

Julio Backhoff Veraguas, Mathias Beiglböck, Yiqing Lin, and Anastasiia Zalashko. Causal transport in discrete time and applications. 3 2017. URL http://arxiv.org/abs/1606.04062.

Jie Wang, Rui Gao, and Yao Xie. Sinkhorn Distributionally Robust Optimization. 3 2025. URL http://arxiv.org/abs/2109.11926.

Qing Wang, Sanjeev R. Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55(5): 2392–2405, May 2009. doi: 10.1109/TIT.2009.2016060.

Jiaqi Wen and Jianyi Yang. Distributionally Robust Optimization via Diffusion Ambiguity Modeling. 10 2025. URL http://arxiv.org/abs/2510.22757.

Jiaqi Wen and Jianyi Yang. Distributionally Robust Optimization via Generative Ambiguity Modeling. 2 2026. URL http://arxiv.org/abs/2602.08976.

Yufeng Yang, Yi Zhou, and Zhaosong Lu. Nested Stochastic Algorithm for Generalized Sinkhorn distance-Regularized Distributionally Robust Optimization. 6 2025. URL http://arxiv.org/abs/2503.22923.

A Additional Related Work

Structured and factorized ambiguity sets. Prior work has exploited independence and factorization to construct structured ambiguity sets and multi-marginal optimal-transport formulations (Chaouach et al., 2022; Tran et al., 2021). These approaches restrict the class of admissible distribution shifts by incorporating known dependency structure, thereby avoiding an unrestricted model of the full joint distribution. Structural causal optimal transport further incorporates structural equations and shows that working with low-dimensional exogenous factors can improve the finite-sample scaling of Wasserstein ambiguity radii (Ehyaei et al., 2025).

exoDRO shares the motivation of exploiting structure, but differs in both the representation and the role of factorization. Rather than defining the ambiguity set through a structured transport cost or constrained coupling, we parameterize each exogenous marginal with a trainable generative model and form their product before applying the fixed structural map. Consequently, structural compatibility is satisfied for every adversarial parameter value by construction.

Statistical benefits of low-dimensional structure. Related work in information estimation has also emphasized that low-dimensional dependency structure can be substantially easier to estimate than an unrestricted high-dimensional joint distribution (Belghazi et al., 2018; McAllester & Stratos, 2020; Duong & Nguyen, 2022; Abdelaleem et al., 2025). Our finite-sample analysis is related in spirit but concerns a different statistical object. We do not estimate the endogenous-space KL divergence directly. Instead, we control the empirical reconstruction objectives of the coordinatewise generative models and combine these bounds through the exact KL decomposition established in Theorem 1. Accordingly, the dependence on the number of exogenous coordinates arises through simultaneous control of the marginal objectives rather than through direct estimation of a high-dimensional divergence.

Entropic and Sinkhorn-regularized DRO. Entropic and Sinkhorn regularization provide another route to tractable continuous ambiguity sets (Yang et al., 2025; Wang et al., 2025; Blanchet et al., 2025). These methods primarily modify the transport geometry or the computational realization of transport-based DRO. Our focus is complementary: given a known structural map, we ask how to parameterize adversarial distributions so that compatibility with the structural mechanisms is automatic. Thus, exoDRO changes the *representation* of the adversarial distribution through exogenous-space factorization rather than introducing an alternative regularization of the transport problem.

B Additional Proof Details

This appendix provides complete proofs and additional discussion for the structural results in Section 4.

B.1 Automatic Structural Compatibility

Proposition 1 (restated). Every distribution in $\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$ belongs to $\mathcal{P}_F(\mathcal{X})$; equivalently,

$$\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) \subseteq \mathcal{P}_F(\mathcal{X}).$$

Proof. Let

$$Q_{\boldsymbol{\theta}} = g_{\#}\left(\bigotimes_{i=1}^{n} P_{\theta_i}\right) \in \mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}).$$

Because g is bijective with measurable inverse,

$$\begin{aligned}(g^{-1})_{\#} Q_{\boldsymbol{\theta}} &= (g^{-1})_{\#} g_{\#}\left(\bigotimes_{i=1}^{n} P_{\theta_i}\right) \\ &= \bigotimes_{i=1}^{n} P_{\theta_i}.\end{aligned}$$

The pullback therefore factorizes over the exogenous coordinates. By Definition 1, this is exactly the condition for $Q_{\boldsymbol{\theta}} \in \mathcal{P}_F(\mathcal{X})$. Since the argument holds for every admissible $\boldsymbol{\theta}$, we obtain

$$\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) \subseteq \mathcal{P}_F(\mathcal{X}).$$

□

Interpretation. The result does not depend on optimization reaching a particular solution. Structural compatibility holds at initialization, during intermediate adversarial updates, and at the final iterate because it is enforced by the parameterization itself. In particular, exoDRO does not require an auxiliary causal regularizer or a projection onto the structurally compatible set after each update.

B.2 KL Invariance under a Bimeasurable Bijection

The exact decomposition in Theorem 1 relies on the standard invariance of KL divergence under an invertible measurable transformation.

Lemma 1 (KL invariance under bijective pushforward)**.** *Let $h : \mathcal{Z} \to \mathcal{Y}$ be a bimeasurable bijection, and let $\mu, \nu \in \mathcal{P}(\mathcal{Z})$ with $\mu \ll \nu$. Then*

$$D_{\mathrm{KL}}\left(h_{\#}\mu \,\|\, h_{\#}\nu\right) = D_{\mathrm{KL}}\left(\mu \,\|\, \nu\right). \tag{22}$$

Proof. Let

$$r(z) := \frac{d\mu}{d\nu}(z)$$

denote the Radon–Nikodym derivative of μ with respect to ν. Since h is a bimeasurable bijection, $h_{\#}\mu \ll h_{\#}\nu$.

For any measurable set $A \subseteq \mathcal{Y}$,

$$\begin{aligned}(h_{\#}\mu)(A) &= \mu\big(h^{-1}(A)\big) \\ &= \int_{h^{-1}(A)} r(z)\, d\nu(z) \\ &= \int_A r\big(h^{-1}(y)\big)\, d(h_{\#}\nu)(y).\end{aligned}$$

Hence,

$$\frac{d(h_{\#}\mu)}{d(h_{\#}\nu)}(y) = r\big(h^{-1}(y)\big)$$

for $h_{\#}\nu$-almost every y.

Substituting this expression into the KL divergence gives

$$\begin{aligned}D_{\mathrm{KL}}(h_{\#}\mu \| h_{\#}\nu) &= \int_{\mathcal{Y}} \log \frac{d(h_{\#}\mu)}{d(h_{\#}\nu)}(y)\, d(h_{\#}\mu)(y) \\ &= \int_{\mathcal{Y}} \log r\big(h^{-1}(y)\big)\; d(h_{\#}\mu)(y) \\ &= \int_{\mathcal{Z}} \log r(z)\, d\mu(z) \\ &= D_{\mathrm{KL}}(\mu \| \nu),\end{aligned}$$

where the third equality follows from the definition of the pushforward measure. □

B.3 Exact Exogenous KL Decomposition

Theorem 1 (restated). Suppose

$$P_0 = g_{\#}\left(\bigotimes_{i=1}^{n} P_{U_i}\right), \qquad Q_{\boldsymbol{\theta}} = g_{\#}\left(\bigotimes_{i=1}^{n} P_{\theta_i}\right),$$

where $g : \mathcal{U} \to \mathcal{X}$ is a bimeasurable bijection. If $P_{U_i} \ll P_{\theta_i}$ for every i, then

$$D_{\mathrm{KL}}(P_0 \| Q_{\boldsymbol{\theta}}) = \sum_{i=1}^{n} D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i}).$$

Proof. Define

$$P_U := \bigotimes_{i=1}^{n} P_{U_i}, \qquad P_{\boldsymbol{\theta}} := \bigotimes_{i=1}^{n} P_{\theta_i}.$$

Because $P_0 = g_{\#} P_U$ and $Q_{\boldsymbol{\theta}} = g_{\#} P_{\boldsymbol{\theta}}$, Lemma 1 gives

$$D_{\mathrm{KL}}(P_0 \| Q_{\boldsymbol{\theta}}) = D_{\mathrm{KL}}(P_U \| P_{\boldsymbol{\theta}}). \tag{23}$$

For each coordinate, define

$$r_i(u_i) := \frac{dP_{U_i}}{dP_{\theta_i}}(u_i).$$

Since $P_{U_i} \ll P_{\theta_i}$ for all i, we also have $P_U \ll P_{\boldsymbol{\theta}}$. The Radon–Nikodym derivative of the product measures factorizes:

$$\frac{dP_U}{dP_{\boldsymbol{\theta}}}(u_1, \ldots, u_n) = \prod_{i=1}^{n} r_i(u_i).$$

Therefore,

$$\begin{aligned} D_{\mathrm{KL}}(P_U \| P_{\boldsymbol{\theta}}) &= \int_{\mathcal{U}} \log \left(\prod_{i=1}^{n} r_i(u_i) \right) dP_U(u) \\ &= \sum_{i=1}^{n} \int_{\mathcal{U}} \log r_i(u_i)\, dP_U(u). \end{aligned}$$

Since P_U is a product probability measure, integrating out all coordinates except u_i yields

$$\int_{\mathcal{U}} \log r_i(u_i)\, dP_U(u) = \int_{\mathcal{U}_i} \log r_i(u_i)\, dP_{U_i}(u_i).$$

Hence,

$$\begin{aligned} D_{\mathrm{KL}}(P_U \| P_{\boldsymbol{\theta}}) &= \sum_{i=1}^{n} \int_{\mathcal{U}_i} \log \frac{dP_{U_i}}{dP_{\theta_i}}(u_i)\, dP_{U_i}(u_i) \\ &= \sum_{i=1}^{n} D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i}). \end{aligned}$$

Combining this equality with Equation 23 proves the theorem. □

Why causal depth does not appear. The equality is independent of the number of structural mechanisms through which an exogenous perturbation propagates. An upstream perturbation can substantially change several downstream endogenous variables, but applying the same invertible map g to both the nominal and adversarial distributions leaves their KL divergence unchanged. This is different from geometric discrepancies such as transportation costs, whose numerical value can change as a perturbation is propagated through the geometry induced by g.

No volume-preservation assumption. The proof does not require $|\det J_g| = 1$. For differentiable density models, this can also be seen directly: the change-of-variables Jacobian appears in both transformed densities and cancels inside their likelihood ratio. The measure-theoretic argument above is more general and requires only that g be a bimeasurable bijection.

No additive-noise requirement. Likewise, the decomposition is not tied to an additive-noise representation of the structural equations. Additive-noise SCMs are one convenient class for which the inverse structural map is readily available, but the theorem applies to any bijective reduced-form map satisfying the stated measurability conditions.

B.4 Reconstruction-Based Global Certificate

Corollary 1 (restated). Under Assumption 1,

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) \leq \sum_{i=1}^{n} \left[J_i(\theta_i) + c_i\right].$$

Consequently, if $Q_{\boldsymbol{\theta}} \in \mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$, then

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) \leq \sum_{i=1}^{n} (\epsilon_i + c_i).$$

Proof. By Theorem 1,

$$D_{\mathrm{KL}}(P_0 \| Q_{\boldsymbol{\theta}}) = \sum_{i=1}^{n} D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i}).$$

Applying Assumption 1 coordinatewise gives

$$D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i}) \leq J_i(\theta_i) + c_i.$$

Summing these inequalities yields

$$D_{\mathrm{KL}}(P_0 \| Q_{\boldsymbol{\theta}}) \leq \sum_{i=1}^{n} [J_i(\theta_i) + c_i].$$

Finally, $Q_{\boldsymbol{\theta}} \in \mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$ implies $J_i(\theta_i) \leq \epsilon_i$ for every coordinate, giving the stated budget bound. □

Source of approximation. Theorem 1 is an exact identity. The inequality above is not: its slack comes entirely from the per-coordinate generative consistency bounds in Assumption 1. Thus, exogenous factorization itself introduces no additional approximation into the KL decomposition.

B.5 Containment in a Structural KL Ambiguity Set

Proposition 2 (restated). Let

$$\rho(\boldsymbol{\epsilon}) := \sum_{i=1}^{n} (\epsilon_i + c_i).$$

Then

$$\mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) \subseteq \mathcal{B}_F^{\mathrm{KL}}\left(P_0, \rho(\boldsymbol{\epsilon})\right).$$

Proof. Take any $Q \in \mathcal{B}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$. By Proposition 1,

$$Q \in \mathcal{P}_F(\mathcal{X}).$$

By Corollary 1,

$$D_{\mathrm{KL}}(P_0 \| Q) \leq \sum_{i=1}^{n} (\epsilon_i + c_i) = \rho(\boldsymbol{\epsilon}).$$

Thus Q satisfies both defining conditions of $\mathcal{B}_F^{\mathrm{KL}}(P_0, \rho(\boldsymbol{\epsilon}))$. Since Q was arbitrary, the desired set inclusion follows. □

Containment rather than equality. The proposition does not imply that every structurally compatible distribution within the KL ball defined by $D_{\mathrm{KL}}(P_0 \| Q)$ can be represented by exoDRO. The set $\mathcal{B}_{\mathrm{SG}}$ is additionally restricted by the expressiveness of the marginal generative families $\{P_{\theta_i}\}$ and by the reconstruction-based constraints used to control them. Accordingly, exoDRO should be interpreted as a trainable parameterization of a subset of the structurally compatible KL neighborhood defined by $D_{\mathrm{KL}}(P_0 \| Q)$, rather than an exact parameterization of the entire set.

C Additional Details for Finite-Sample Control

This appendix provides the assumptions, intermediate results, and proofs underlying the finite-sample guarantee summarized in Section 5.

Suppose that

$$X^{(1)}, \dots, X^{(N)} \overset{\text{i.i.d.}}{\sim} P_0.$$

Because the structural map g is bijective, each observation can be mapped to the exogenous space:

$$U^{(k)} = g^{-1}\left(X^{(k)}\right), \qquad k = 1, \dots, N.$$

Writing $U^{(k)} = (U_1^{(k)}, \dots, U_n^{(k)})$, the same N endogenous observations therefore provide N samples for every exogenous coordinate U_i. In particular, the coordinatewise generative models do not require separate datasets of sizes $N_1, \dots, N_n$; they are trained using different coordinate projections of the same N exogenous samples.

C.1 Empirical Reconstruction Objectives

For objectives involving auxiliary training randomness, such as a noise level or Gaussian perturbation in diffusion training, let ξ_i denote this randomness and define the conditional expected per-sample loss

$$\bar{\ell}_i(\theta_i, u_i) := \mathbb{E}_{\xi_i}\left[\ell_i(\theta_i, u_i; \xi_i)\right].$$

The corresponding population reconstruction objective is

$$J_i(\theta_i) = \mathbb{E}_{U_i \sim P_{U_i}}\left[\bar{\ell}_i(\theta_i, U_i)\right].$$

When no auxiliary randomness is present, $\bar{\ell}_i = \ell_i$.

For coordinate i, define the empirical objective

$$\widehat{J}_i(\theta_i) := \frac{1}{N}\sum_{k=1}^{N} \bar{\ell}_i\left(\theta_i, U_i^{(k)}\right).$$

We use the following standard bounded-loss assumption.

Assumption 2 (Bounded reconstruction loss)**.** *For every coordinate i, the conditional expected per-sample loss satisfies*

$$0 \le \bar{\ell}_i(\theta_i, u_i) \le B_i$$

almost surely for every admissible parameter $\theta_i \in \Theta_i$. Let

$$\mathcal{L}_i := \left\{u_i \mapsto \bar{\ell}_i(\theta_i, u_i) : \theta_i \in \Theta_i\right\}$$

denote the induced loss class, and let p_i denote an appropriate complexity measure such as its pseudo-dimension.

C.2 Coordinatewise Uniform Convergence

We first control each reconstruction objective separately.

Proposition 4 (Coordinatewise uniform convergence)**.** *Under Assumption 2, for every $\delta_i \in (0, 1)$, with probability at least $1 - \delta_i$,*

$$\sup_{\theta_i \in \Theta_i} \left|J_i(\theta_i) - \widehat{J}_i(\theta_i)\right| \le \Delta_i(N, \delta_i), \tag{24}$$

where, up to universal constants,

$$\Delta_i(N, \delta_i) = \mathcal{O}\left(B_i\sqrt{\frac{p_i \log N + \log(1/\delta_i)}{N}}\right). \tag{25}$$

Proof. This follows from the standard pseudo-dimension uniform-convergence bound applied to the loss class $\mathcal{L}_i$ under the marginal distribution P_{U_i}. The resulting complexity term depends on the coordinatewise loss class rather than directly on the dimension of the full endogenous vector. □

The relevant complexity is therefore associated with each marginal generative model rather than an unrestricted model of the full joint distribution. This is consistent with the broader observation that estimation can become easier when low-dimensional dependency structure is available (Belghazi et al., 2018; McAllester & Stratos, 2020; Duong & Nguyen, 2022; Abdelaleem et al., 2025).

C.3 Simultaneous Control Across Exogenous Coordinates

All marginal constraints must hold simultaneously. Applying Proposition 4 with $\delta_i = \delta/n$ and taking a union bound gives the following result.

Corollary 3 (Simultaneous coordinatewise control)**.** *Under Assumption 2, with probability at least* $1 - \delta$*,*

$$\left| J_i(\theta_i) - \widehat{J}_i(\theta_i) \right| \le \Delta_i(N, \delta/n) \qquad \forall i \in [n], \tag{26}$$

uniformly over the coordinatewise parameter spaces. In particular,

$$\Delta_i(N, \delta/n) = \mathcal{O}\left(B_i \sqrt{\frac{p_i \log N + \log(n/\delta)}{N}} \right). \tag{27}$$

Proof. For each coordinate i, define

$$E_i := \left\{ \sup_{\theta_i \in \Theta_i} \left| J_i(\theta_i) - \widehat{J}_i(\theta_i) \right| \le \Delta_i(N, \delta/n) \right\}.$$

By Proposition 4, $\Pr(E_i^c) \le \delta/n$. The events $E_1, \dots, E_n$ need not be independent, even though the coordinates are obtained from the same observations. By the union bound,

$$\begin{aligned} \Pr\left(\bigcap_{i=1}^n E_i \right) &= 1 - \Pr\left(\bigcup_{i=1}^n E_i^c \right) \\ &\ge 1 - \sum_{i=1}^n \Pr(E_i^c) \\ &\ge 1 - \delta. \end{aligned}$$

□

Thus, simultaneous control introduces the usual $\log(n/\delta)$ dependence through the confidence term. Importantly, this does not require n separate datasets: all coordinatewise bounds are obtained from the same N observations.

C.4 Empirical–Population Ambiguity-Set Sandwich

Define the empirically constrained ambiguity set

$$\widehat{\mathcal{B}}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon}) := \left\{ g_{\#} \left(\bigotimes_{i=1}^n P_{\theta_i} \right) : \widehat{J}_i(\theta_i) \le \epsilon_i, \quad \forall i \right\}. \tag{28}$$

For $\boldsymbol{\Delta} = (\Delta_1, \dots, \Delta_n)$, all vector inequalities below are interpreted coordinatewise.

Proposition 5 (Empirical–population ambiguity-set sandwich)**.** *On the simultaneous event of Corollary 3,*

$$\mathcal{B}_{\mathrm{SG}}\left(P_0, \boldsymbol{\epsilon} - \boldsymbol{\Delta}\right) \subseteq \widehat{\mathcal{B}}_{\mathrm{SG}}\left(P_0, \boldsymbol{\epsilon}\right) \subseteq \mathcal{B}_{\mathrm{SG}}\left(P_0, \boldsymbol{\epsilon} + \boldsymbol{\Delta}\right), \tag{29}$$

provided $\boldsymbol{\epsilon} - \boldsymbol{\Delta} \ge 0$ *coordinatewise.*

Proof. Suppose first that

$$J_i(\theta_i) \le \epsilon_i - \Delta_i.$$

On the simultaneous uniform-convergence event,

$$\widehat{J}_i(\theta_i) \le J_i(\theta_i) + \Delta_i \le \epsilon_i,$$

which proves the first inclusion.

Conversely, if

$$\widehat{J}_i(\theta_i) \le \epsilon_i,$$

then

$$J_i(\theta_i) \le \widehat{J}_i(\theta_i) + \Delta_i \le \epsilon_i + \Delta_i,$$

which proves the second inclusion. □

Proposition 5 gives a direct finite-sample interpretation of the empirically trained adversary: an empirical budget ϵ_i corresponds, with high probability, to a population budget inflated by at most the coordinatewise estimation error Δ_i.

C.5 Proof of the Finite-Sample Endogenous KL Guarantee

We now prove Corollary 2 from the main paper.

Proof of Corollary 2. On the simultaneous event of Corollary 3, Proposition 5 implies that every

$$Q_{\boldsymbol{\theta}} \in \widehat{\mathcal{B}}_{\mathrm{SG}}(P_0, \boldsymbol{\epsilon})$$

also belongs to

$$\mathcal{B}_{\mathrm{SG}}\left(P_0, \boldsymbol{\epsilon} + \boldsymbol{\Delta}\right).$$

Applying Corollary 1 with the inflated coordinatewise budgets gives

$$D_{\mathrm{KL}}\left(P_0 \,\|\, Q_{\boldsymbol{\theta}}\right) \le \sum_{i=1}^{n} \left[\epsilon_i + \Delta_i(N, \delta/n) + c_i\right]. \tag{30}$$

The simultaneous event holds with probability at least $1 - \delta$, which completes the proof. □

The result avoids direct estimation of a high-dimensional endogenous-space KL divergence. Instead, the only empirical quantities being controlled are the coordinatewise reconstruction objectives that already define the generative ambiguity set. We emphasize, however, that the result should not be interpreted as having only logarithmic overall dependence on the number of coordinates. The term $\log(n/\delta)$ appears inside each coordinatewise confidence radius $\Delta_i(N, \delta/n)$, while the final KL certificate remains additive across the n coordinates.

The analysis also treats the expectation over any auxiliary generative-model training randomness as part of the conditional per-sample objective $\tilde{\ell}_i$. If this expectation is itself approximated using finitely many Monte Carlo samples in practice, the additional Monte Carlo estimation error is not included in the bound above.

C.6 Relation to Structural Causal Optimal Transport

The finite-sample guarantee above should not be interpreted as reproducing the dimension-free Wasserstein concentration result of structural causal optimal transport (Ehyaei et al., 2025). The two analyses control different statistical objects.

Structural causal optimal transport studies concentration of a structurally constrained transport discrepancy and exploits low-dimensional exogenous structure to characterize the scaling of the corresponding Wasserstein ambiguity radius. In contrast, our analysis does not estimate a transport distance or the endogenous-space KL divergence directly. It controls the empirical reconstruction objectives of learned marginal generative models and then translates these coordinatewise bounds into an endogenous-space KL certificate through Theorem 1.

Accordingly, our claim is more specific: a single dataset containing N endogenous observations provides N samples for every exogenous coordinate simultaneously, and enforcing all coordinatewise generative constraints introduces a $\log(n/\delta)$ confidence adjustment rather than requiring separate sample budgets for the n marginal models. The final KL certificate nevertheless remains additive across coordinates, as shown in Equation 30.

D Additional Optimization Details

The exogenous factorization makes the reconstruction constraints coordinatewise, but the downstream adversarial objective remains coupled across coordinates. This appendix provides the detailed single-block reduction, gradient estimators, dual updates, complete cyclic algorithm, and the scope of the per-block optimization guarantee summarized in Section 6.

D.1 Exact Reduction of a Single Block

Fix w, a coordinate i, and the remaining generative parameters θ_{-i}. Recall the averaged block reward

$$\widetilde{f}_i(w, u_i; \theta_{-i}) := \mathbb{E}_{U_{-i} \sim \bigotimes_{j \neq i} P_{\theta_j}} \left[f\big(w, g(U_{-i}, u_i)\big) \right]. \tag{31}$$

By iterated expectation,

$$\begin{aligned} &\mathbb{E}_{U \sim \bigotimes_{j=1}^{n} P_{\theta_j}} \left[f(w, g(U)) \right] \\ &\quad = \mathbb{E}_{U_i \sim P_{\theta_i}} \left[\mathbb{E}_{U_{-i} \sim \bigotimes_{j \neq i} P_{\theta_j}} \left[f\big(w, g(U_{-i}, U_i)\big) \right] \right] \\ &\quad = \mathbb{E}_{U_i \sim P_{\theta_i}} \left[\widetilde{f}_i(w, U_i; \theta_{-i}) \right]. \end{aligned} \tag{32}$$

Therefore, with w and θ_{-i} fixed, optimization over θ_i reduces exactly to

$$\sup_{\theta_i} \ \mathbb{E}_{U_i \sim P_{\theta_i}} \left[\widetilde{f}_i(w, U_i; \theta_{-i}) \right] \qquad \text{s.t.} \qquad J_i(\theta_i) \leq \epsilon_i.$$

For a fixed multiplier $\mu_i \geq 0$, the θ_i-dependent part of the Lagrangian is

$$\mathbb{E}_{U_i \sim P_{\theta_i}} \left[\widetilde{f}_i(w, U_i; \theta_{-i}) \right] - \mu_i J_i(\theta_i),$$

since $\mu_i \epsilon_i$ is constant with respect to θ_i.

Importantly, this reduction does not fix a particular realization of U_{-i}. The effect of the remaining coordinates is integrated into $\widetilde{f}_i$; Monte Carlo sampling is used only to estimate this expectation during optimization.

D.2 Policy-Gradient Block Update

For concreteness, consider a diffusion parameterization of P_{θ_i}, and let $U_i^{0:T}$ denote a reverse diffusion trajectory sampled from the current coordinate model. For fixed θ_{-i}, a score-function estimator of the reward gradient is

$$\nabla_{\theta_i} \mathbb{E} \left[\widetilde{f}_i(w, U_i^0; \theta_{-i}) \right] = \mathbb{E} \left[\widetilde{f}_i(w, U_i^0; \theta_{-i}) \nabla_{\theta_i} \log P_{\theta_i}(U_i^{0:T}) \right]. \tag{33}$$

The expectation over the remaining coordinates in Equation 31 can be approximated using samples from their current adversarial models:

$$\widehat{\widetilde{f}}_i(w, u_i) = \frac{1}{M} \sum_{m=1}^{M} f\left(w, g\left(U_{-i}^{(m)}, u_i\right)\right), \qquad U_{-i}^{(m)} \sim \bigotimes_{j \neq i} P_{\theta_j}. \tag{34}$$

The corresponding gradient of the block Lagrangian is

$$\nabla_{\theta_i} \mathcal{L} = \nabla_{\theta_i} \mathbb{E} \left[\widetilde{f}_i \right] - \mu_i \nabla_{\theta_i} J_i(\theta_i). \tag{35}$$

A PPO-style clipped objective may be used in place of the raw score-function update for improved empirical stability. The convergence statement in Proposition 3, however, applies only when the single-model inner-maximization assumptions and procedure analyzed by Wen & Yang (2026) are satisfied; it should not be interpreted as a convergence guarantee for an arbitrary PPO or gradient implementation.

D.3 Dual-Variable Update

For fixed w, the Lagrangian of the joint inner maximization is

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\mu}; w) = \mathbb{E}_{U \sim \bigotimes_i P_{\theta_i}} \left[f(w, g(U)) \right] - \sum_{i=1}^{n} \mu_i \left(J_i(\theta_i) - \epsilon_i \right),$$

with $\mu_i \geq 0$.

Because the primal problem is a constrained maximization problem, the associated dual problem minimizes the dual function over $\boldsymbol{\mu} \geq 0$. Since

$$\nabla_{\mu_i} \mathcal{L} = - \left(J_i(\theta_i) - \epsilon_i \right),$$

projected gradient descent gives

$$\mu_i \leftarrow \left[\mu_i + \eta_\mu \left(J_i(\theta_i) - \epsilon_i \right) \right]_+ . \tag{36}$$

Thus, when the reconstruction constraint is violated, $J_i(\theta_i) > \epsilon_i$, the corresponding multiplier increases and places a larger penalty on subsequent adversarial updates.

D.4 Complete Cyclic exoDRO Algorithm

Algorithm 1 summarizes the complete cyclic procedure. In practical finite-sample training, the population reconstruction objectives J_i are replaced by their empirical counterparts $\widehat{J}_i$.

The cyclic ordering is not essential for implementing the method, but it matches the conditional single-block analysis: when coordinate i is updated, the remaining adversarial models are treated as fixed. Parallel or Jacobi-style variants are also possible, but they no longer correspond directly to the fixed-block argument because multiple adversaries change simultaneously.

D.5 Conditional Per-Block Optimization Guarantee

Proposition 3 (restated). Fix w, coordinate i, and θ_{-i}. Then the coordinate-i subproblem is exactly a single-model generative-DRO inner problem with nominal distribution P_{U_i}, reward

$$u_i \mapsto \widetilde{f}_i(w, u_i; \theta_{-i}),$$

and reconstruction constraint $J_i(\theta_i) \leq \epsilon_i$. If the subproblem is optimized using the single-model inner-maximization procedure analyzed by Wen & Yang (2026), and the assumptions required by their convergence result hold for $\widetilde{f}_i(w, \cdot; \theta_{-i})$ and J_i, then the corresponding blockwise inner-maximization criterion inherits the same $\mathcal{O}(K^{-1/2})$ convergence rate.

Proof. For fixed w and θ_{-i}, Equation 32 shows that the joint adversarial objective reduces exactly to

$$\mathbb{E}_{U_i \sim P_{\theta_i}} \left[\widetilde{f}_i(w, U_i; \theta_{-i}) \right].$$

The nominal distribution is therefore P_{U_i}, the adversarial distribution is P_{θ_i}, the reward is $\widetilde{f}_i(w, \cdot; \theta_{-i})$, and the reconstruction constraint is $J_i(\theta_i) \leq \epsilon_i$.

These are exactly the ingredients of the corresponding single-model generative-DRO inner problem. Once θ_{-i} is fixed, no additional coupling remains in the optimization variable θ_i. Consequently, when the assumptions and inner-maximization procedure apply, their $\mathcal{O}(K^{-1/2})$ convergence result transfers directly to this fixed-block subproblem. □

Algorithm 1 exoDRO with Cyclic Block-Coordinate Updates

Require: Nominal samples $\{X^{(k)}\}_{k=1}^{N}$, bijective structural map g, budgets $\boldsymbol{\epsilon}$, outer learning rate η_w, dual learning rate η_μ.

Compute $U^{(k)} = g^{-1}(X^{(k)})$ for all k.
Initialize decision parameter w, generative parameters $\theta_1, \dots, \theta_n$, and multipliers $\mu_1, \dots, \mu_n \geq 0$.
for $r = 1, \dots, R$ **do**
  for $i = 1, \dots, n$ **do**
    Hold θ_{-i} fixed.
    for $k = 1, \dots, K$ **do**
      Sample U_{-i} from $\bigotimes_{j \neq i} P_{\theta_j}$.
      Estimate the block reward $\widetilde{f}_i$ using Equation 34.
      Update θ_i using Equation 35.
      Update μ_i using Equation 36.
    end for
  end for
  Sample $U \sim \bigotimes_i P_{\theta_i}$ and set $X = g(U)$.
  Update the decision parameter:

$$w \leftarrow w - \eta_w \nabla_w f(w, X).$$

end for
return w.

Scope of the guarantee. The proposition is conditional on w and θ_{-i} remaining fixed during the coordinate update. It does not establish convergence of the full cyclic sequence $(\theta_1, \dots, \theta_n)$ to a global maximizer of the coupled nonconvex inner problem, since updating one coordinate changes the averaged reward faced by the subsequent coordinates. The stated rate also applies specifically to the corresponding single-model inner-maximization procedure. It is not a general convergence guarantee for arbitrary gradient, PPO, cyclic, or parallel implementations.

E Experimental Details

This appendix provides the complete experimental protocols and additional results for Section 7. We first describe the predictive-robustness benchmark and its additional analyses, followed by the controlled experiments used to examine the structural properties of exoDRO.

E.1 Predictive Robustness Comparison Protocol

Data generation. The predictive benchmark is generated from the fixed structural causal model

$$\begin{aligned} X_1 &= U_1, \\ X_2 &= 0.8X_1 + U_2, \\ X_3 &= -0.6X_1 + U_3, \\ X_4 &= 0.7X_2 - 0.5X_3 + U_4, \end{aligned}$$

with target

$$Y = 0.7X_2 - 0.5X_3 + 0.6X_4 + 0.8\tanh(0.5X_1X_4) + 0.3U_y.$$

Under the nominal distribution,

$$U_1, U_2, U_3, U_4, U_y \overset{\text{ind}}{\sim} \mathcal{N}(0, 1).$$

The nominal training, ID-validation, and ID-test sets contain 10,000, 2,000, and 5,000 samples, respectively, and are generated using independent random seeds.

Distribution shift is introduced exclusively through the exogenous distributions of $U_1, \dots, U_4$, while the structural equations and the distribution of U_y remain fixed. In each OOD environment,

$$U_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad i = 1, \dots, 4,$$

where the mean and standard deviation parameters are sampled independently according to the shift severity:

Severity	μ_i	σ_i
Mild	$[-0.5, 0.5]$	$[0.8, 1.2]$
Moderate	$[-1.0, 1.0]$	$[0.6, 1.4]$
Strong	$[-1.5, 1.5]$	$[0.5, 1.5]$.

To obtain separated shift severities, a sampled environment is retained only when its analytic exogenous KL divergence

$$D_{\mathrm{KL}}\left(\mathcal{N}(0, I_4) \,\middle\|\, \bigotimes_{i=1}^{4} \mathcal{N}(\mu_i, \sigma_i^2)\right) = \sum_{i=1}^{4} \left[\log \sigma_i + \frac{1+\mu_i^2}{2\sigma_i^2} - \frac{1}{2}\right]$$

falls within the corresponding interval: $[0.08, 0.45]$ for mild, $[0.55, 1.45]$ for moderate, and $[1.60, 3.30]$ for strong shifts. For each severity level, we generate five OOD validation environments and ten held-out OOD test environments, yielding 15 validation and 30 test environments in total. Each OOD environment contains 5,000 samples. The validation and test environments are generated using disjoint random seeds and are used exclusively for model selection and final evaluation, respectively. For every generated environment, the target exogenous distribution parameters and their realized empirical statistics are recorded to verify the intended shift severity while keeping the SCM fixed.

Prediction model. All compared methods use the same prediction architecture whenever their formulations permit: a multilayer perceptron with four input features, two hidden layers of 64 units with ReLU activations, and one scalar output, i.e., $4 \to 64 \to 64 \to 1$. Models are trained using mean-squared error. Using a common predictor isolates differences due to the robust training objectives rather than predictive model capacity.

Compared methods. We compare six training strategies: empirical risk minimization (ERM), Wasserstein DRO (W-DRO) (Sinha et al., 2018), KL-DRO (Hu & Hong, 2013), structural causal optimal transport (SCOT), GAS-DRO, and the proposed exoDRO. ERM serves as the nominal empirical-risk baseline. W-DRO and KL-DRO provide transport- and divergence-based robust baselines, respectively. GAS-DRO (Wen & Yang, 2026) provides an endogenous-space generative adversary, whereas SCOT (Ehyaei et al., 2025) is implemented using a finite-sample approximation and serves as a structure-aware transport-based reference. exoDRO instead parameterizes independent adversarial marginals in the exogenous space and maps their product distribution through the fixed SCM.

Hyperparameter and model selection. Method-specific hyperparameters are selected using the designated OOD validation environments. Because parameters associated with different ambiguity geometries have different numerical meanings, tuning is performed independently for each method rather than enforcing a common numerical radius.

For each hyperparameter candidate, the best checkpoint is selected independently for each training seed according to the worst-environment MSE over the 15 OOD validation environments. Hyperparameters are then selected by minimizing this worst-environment validation MSE averaged across the five training seeds.

For all robust methods initialized from ERM, the seed-matched ERM checkpoint is included as the step-0 candidate for checkpoint selection. For each training seed, the step-0 checkpoint and all subsequent candidate checkpoints are evaluated on the designated OOD validation environments, and the checkpoint with the lowest worst-environment OOD-validation MSE is selected. This ensures that all robust methods start from the same seed-matched predictor checkpoint, while the corresponding ERM solution remains an eligible step-0 model under the same validation criterion. The final hyperparameter configuration and checkpoint are fixed before evaluation on the held-out OOD test environments, which are never used for hyperparameter or checkpoint selection. Table 3 summarizes the final optimization settings used in the predictive comparison.

Baseline implementation details. ERM is trained from a fresh, seed-specific MLP initialization rather than from a pretrained checkpoint. The selected configuration uses $\eta = 3 \times 10^{-4}$, $\lambda_{\mathrm{wd}} =$

Table 3: Final optimization settings for the predictive robustness benchmark. Hyperparameters and checkpoints are selected using the common OOD-validation environments and worst-environment validation criterion described in Appendix E.1.

Method	Optimization	Final setting	Predictor init.
ERM	batch 128	$\eta = 3 \times 10^{-4}$, $\lambda_{\mathrm{wd}} = 10^{-5}$	Fresh MLP
W-DRO	Adam, $\eta = 10^{-3}$, batch 64, 100 epochs	$\gamma = 100$	ERM checkpoint
KL-DRO	Adam, $\eta = 10^{-3}$, full-batch, 2000 updates	$\epsilon = 0.01$	ERM checkpoint
SCOT	Adam, $\eta = 10^{-3}$, batch 64, 50 epochs	$r = 0.5$	ERM checkpoint
GAS-DRO	endogenous generative adversary	budget $= 0.010$	ERM checkpoint
exoDRO	batch 128, 16000 outer steps	$b = 6,\ \epsilon_i = 0.1,\ K = 10$	ERM checkpoint

10^{-5}, and batch size 128. The resulting seed-specific ERM checkpoints are used to initialize the robust methods that require ERM initialization.

W-DRO follows a fixed-penalty Wasserstein robust objective. The selected penalty coefficient is $\gamma = 100$. For each seed, training is initialized from the corresponding seed-matched ERM checkpoint.

KL-DRO constructs its ambiguity set around the full empirical distribution and performs 2,000 full-batch optimization updates. The selected KL radius is $\epsilon = 0.01$. As with W-DRO, training is initialized from the corresponding seed-matched ERM checkpoint.

SCOT uses a finite-sample approximation of its structurally compatible transport ambiguity set. A bank of 512 candidate perturbations is evaluated in chunks of 64, and the predictor is updated against the selected worst-case candidate. All four exogenous coordinates are active. The selected transport radius is $r = 0.5$. Training is initialized from the corresponding seed-matched ERM checkpoint.

GAS-DRO parameterizes its generative adversary directly in the endogenous space and is initialized from the corresponding seed-matched ERM checkpoint. The ambiguity budget is selected using the designated OOD validation environments, yielding a final budget of 0.010. After each predictor optimizer step, the current predictor is evaluated on the 15 OOD validation environments, and the checkpoint with the lowest worst-environment validation MSE is retained for final evaluation.

exoDRO implementation. Each exogenous coordinate is parameterized using a one-dimensional rational-quadratic spline generative model with eight bins. Each generator is initialized from a nominal model pretrained on the corresponding exogenous marginal, while the predictor is initialized from the corresponding seed-matched ERM checkpoint. We fix the number of inner updates per coordinate to $K = 10$. The spline bound and KL budget are selected using the designated OOD validation environments, yielding $b = 6$ and a coordinatewise budget of $\epsilon_i = 0.1$. For this tractable-density instantiation, we directly use the coordinatewise inclusive KL divergence, $J_i(\theta_i) := D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i})$, as the ambiguity constraint; thus, Assumption 1 holds with $c_i = 0$. The final configuration uses learning rates 3×10^{-3}, 10^{-3}, and 10^{-2} for the predictor, generator, and dual variables, respectively, with batch size 128 and 16,000 outer optimization steps. The common predictor contains 4,545 trainable parameters. exoDRO additionally uses four one-dimensional spline generators, with 92 generator parameters in total.

After every predictor update, the current checkpoint is evaluated on the 15 OOD validation environments. For each seed, the checkpoint minimizing the maximum MSE across these environments is retained.

Realized KL at the selected checkpoints. The KL budget in exoDRO is imposed separately on each exogenous coordinate, with a target of $\epsilon_i = 0.1$ for $i = 1, \ldots, 4$. At the selected checkpoints, the sum of the four realized coordinatewise inclusive KL divergences is 0.3943 ± 0.0134 across the five training seeds, close to the sum of the four target budgets, 0.4. Because the constraints are enforced through dual updates rather than exact projection, small numerical deviations from the target may occur; among the 20 coordinate–seed pairs, the largest observed excess above the target 0.1 was 0.0089.

Evaluation. For each method, we report ID MSE, mild-OOD MSE, moderate-OOD MSE, strong-OOD MSE, and worst-environment OOD MSE. The worst-environment metric is defined as

$$\text{Worst OOD MSE} = \max_{e \in \mathcal{E}_{\text{test}}} \text{MSE}_e.$$

All methods are evaluated over five independent training seeds. For each severity level, performance is aggregated across the corresponding 10 held-out environments, and we report the mean and standard deviation across training seeds. The same dataset-generation and evaluation protocol is used across methods whenever their formulations permit.

E.2 Excess-Risk Analysis

To complement the absolute MSE results in Section 7, we evaluate prediction error above the irreducible noise level. Since U_y is independent of (X_1, X_2, X_3, X_4) and remains distributed as $\mathcal{N}(0, 1)$ in every environment, the Bayes predictor is

$$f^\star(X) = \mathbb{E}[Y \mid X] = 0.7X_2 - 0.5X_3 + 0.6X_4 + 0.8\tanh(0.5X_1X_4),$$

with population Bayes MSE

$$\mathbb{E}\big[(Y - f^\star(X))^2\big] = \text{Var}(0.3U_y) = 0.09.$$

Although the population Bayes MSE is identical across environments, its empirical value varies across finite test sets. We therefore report an environment-matched empirical excess MSE. For environment e with N_e test samples, define

$$\widehat{R}_e^\star = \frac{1}{N_e} \sum_{k=1}^{N_e} \left(Y^{(k)} - f^\star(X^{(k)})\right)^2,$$

and

$$\widehat{\text{ExcessMSE}}_e(f) = \widehat{\text{MSE}}_e(f) - \widehat{R}_e^\star.$$

Thus, the model and Bayes predictor are evaluated on exactly the same samples before taking their difference. This controls for environment-specific finite-sample variation in the irreducible prediction error. Across the 30 held-out OOD environments, the empirical Bayes MSE ranges from 0.086773 to 0.093344, despite having the same population value of 0.09.

Figure 1 provides a visual summary of the environment-matched excess MSE across the three shift severities, while Table 4 reports the corresponding numerical results. For each shift severity, excess MSE is first averaged across the 10 environments of that severity for each training seed. Average OOD denotes the mean over all 30 held-out OOD environments. Worst OOD denotes the maximum environment-matched excess MSE among the 30 environments for each training seed. We then report the mean and sample standard deviation across the five training seeds.

The environment-matched analysis preserves the overall ordering observed with raw MSE while more clearly separating prediction error above the irreducible noise level. exoDRO achieves the lowest excess MSE in every reported setting. Its average OOD excess MSE is 0.0011, compared with 0.0017 for SCOT and approximately 0.0049–0.0053 for ERM, W-DRO, KL-DRO, and GAS-DRO.

The separation becomes larger as the distribution shift becomes stronger. Under the strong OOD setting, exoDRO attains an excess MSE of 0.0013, compared with 0.0028 for SCOT and approximately 0.0075–0.0077 for the remaining methods. The worst-environment analysis shows the same pattern: exoDRO achieves a worst OOD excess MSE of 0.0023, compared with 0.0057 for SCOT and approximately 0.0127–0.0131 for the other methods.

These results also clarify the interpretation of the raw worst-environment MSE. Because the empirical Bayes error varies across finite test environments, the environment with the largest raw MSE need not be the environment in which a predictor has the largest error above the Bayes predictor. Environment-matched excess MSE removes this finite-sample variation and therefore provides a complementary measure of degradation under distribution shift.

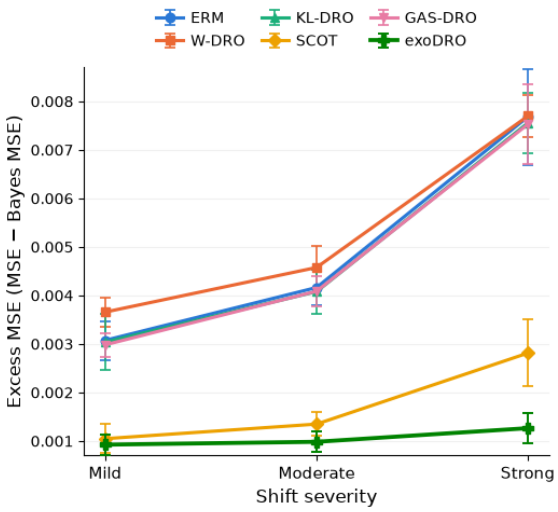

Figure 1: Excess MSE relative to the Bayes predictor under mild, moderate, and strong distribution shifts. For each severity level, performance is averaged across the corresponding 10 held-out environments; curves show the mean and error bars show ± 1 sample standard deviation across five training seeds. Lower is better.

Table 4: Environment-matched excess MSE under distribution shift. For each test environment, the empirical Bayes-predictor MSE on the same samples is subtracted from the model MSE. Results are mean $\pm$ sample standard deviation over five training seeds. Lower is better.

Method	ID	Mild	Moderate	Strong	Avg. OOD	Worst OOD
ERM	0.0024 ± 0.0003	0.0031 ± 0.0004	0.0042 ± 0.0004	0.0077 ± 0.0010	0.0050 ± 0.0005	0.0131 ± 0.0023
W-DRO	0.0029 ± 0.0002	0.0037 ± 0.0003	0.0046 ± 0.0004	0.0077 ± 0.0004	0.0053 ± 0.0003	0.0130 ± 0.0004
KL-DRO	0.0024 ± 0.0004	0.0030 ± 0.0006	0.0041 ± 0.0005	0.0076 ± 0.0006	0.0049 ± 0.0005	0.0127 ± 0.0014
SCOT	0.0009 ± 0.0005	0.0010 ± 0.0003	0.0013 ± 0.0002	0.0028 ± 0.0007	0.0017 ± 0.0004	0.0057 ± 0.0023
GAS-DRO	0.0023 ± 0.0003	0.0030 ± 0.0002	0.0041 ± 0.0003	0.0075 ± 0.0008	0.0049 ± 0.0004	0.0130 ± 0.0023
exoDRO	$\mathbf{0.0007 \pm 0.0002}$	$\mathbf{0.0009 \pm 0.0002}$	$\mathbf{0.0010 \pm 0.0002}$	$\mathbf{0.0013 \pm 0.0003}$	$\mathbf{0.0011 \pm 0.0002}$	$\mathbf{0.0023 \pm 0.0004}$

E.3 COORDINATEWISE EXOGENOUS SHIFT ANALYSIS

To further examine how predictive robustness depends on the source of distribution shift, we decompose the held-out OOD environments according to the shifted exogenous coordinate. For each of the 30 OOD environments used in the predictive benchmark, we construct variants in which only one of U_1, U_2, U_3, U_4 follows the shifted distribution specified by that environment, while the remaining exogenous variables retain their nominal $\mathcal{N}(0,1)$ distributions. The target-noise variable U_y is also kept at its nominal distribution. We additionally construct a no-shift reference and an all-coordinate variant in which all four exogenous shifts are applied simultaneously.

All derived environments contain 5,000 samples. We use independent base random seeds across shift severities and environments, while the no-shift, single-coordinate, and all-coordinate variants corresponding to the same environment share the same underlying random realization. This paired construction isolates the effect of the specified exogenous shift while reducing variation due to test-sample realization.

No additional training, hyperparameter tuning, or checkpoint selection is performed for this analysis. Each method is evaluated using the final checkpoint selected in the main predictive comparison. For method m and shifted coordinate set S, we report the increase in average OOD MSE relative to the paired no-shift reference:

$$\Delta_S^{(m)} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left[\mathrm{MSE}_{e,S}^{(m)} - \mathrm{MSE}_{e,\varnothing}^{(m)} \right],$$

where $\varnothing$ denotes the no-shift variant and $\mathcal{E}$ contains the 30 held-out OOD environments. Smaller values therefore indicate less predictive degradation under the specified exogenous shift.

Table 5: coordinatewise sensitivity to exogenous distribution shift. Each entry reports the increase in average OOD MSE relative to the paired no-shift reference, in units of 10^{-3}. For each environment, shifted and no-shift variants share the same underlying random realization. Lower is better.

Method	Δ_{U_1}	Δ_{U_2}	Δ_{U_3}	Δ_{U_4}	Δ_{all}
ERM	0.386	0.237	0.415	0.926	2.056
W-DRO	0.337	0.243	0.435	0.867	1.799
KL-DRO	0.353	0.272	0.563	0.863	2.093
SCOT	0.039	0.105	0.118	0.222	0.646
GAS-DRO	0.381	0.235	0.421	0.915	2.031
exoDRO	**0.003**	**0.049**	**0.025**	**0.098**	**0.189**

Table 5 shows that exoDRO exhibits the smallest performance degradation for every single-coordinate shift as well as when all four exogenous coordinates are shifted jointly. Among the individual coordinates, shifts in U_4 produce the largest degradation for all methods. This is consistent with the structural role of U_4: it enters X_4 directly through

$$X_4 = 0.7X_2 - 0.5X_3 + U_4,$$

while X_4 appears both linearly and through the interaction term $0.8\tanh(0.5X_1X_4)$ in the prediction target.

The difference is particularly clear under the joint exogenous shift. The increase in average OOD MSE is only 0.189×10^{-3} for exoDRO, compared with 0.646×10^{-3} for SCOT and 1.799–2.093×10^{-3} for ERM, W-DRO, KL-DRO, and GAS-DRO. Thus, the robustness advantage observed in the aggregate OOD benchmark is also reflected when distribution shifts are decomposed according to their exogenous source.

E.4 Synthetic Structural Causal Model

The controlled structural experiments use the three-variable causal chain

$$A = U_A, \qquad E = \alpha A + U_E, \qquad I = \beta E + U_I,$$

where

$$U_A, U_E, U_I \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Unless otherwise specified, $\alpha, \beta \in \{-1, -0.5, 0, 0.5, 1\}$.

The corresponding reduced-form structural map is

$$g(u_A, u_E, u_I) = \left(u_A,\, \alpha u_A + u_E,\, \alpha\beta u_A + \beta u_E + u_I\right).$$

Its inverse is

$$u_A = A, \qquad u_E = E - \alpha A, \qquad u_I = I - \beta E.$$

The analytic inverse eliminates numerical inversion error from the controlled structural analyses.

E.5 Marginal Generative Models

Each adversarial exogenous marginal P_{θ_i} is represented by a one-dimensional likelihood-based density model. We use monotone spline flows rather than standard coupling-based normalizing flows, since coupling transformations are degenerate in one dimension. The base distribution is standard Gaussian. Each marginal model is initialized from a model fitted to its nominal distribution P_{U_i}, and all coordinate models use the same architecture and optimization settings.

Because these models have tractable densities, the marginal inclusive KL divergence used as the coordinatewise ambiguity constraint can be evaluated directly:

$$D_{\mathrm{KL}}(P_{U_i} \| P_{\theta_i}) = \mathbb{E}_{U_i \sim P_{U_i}}\left[\log \frac{p_{U_i}(U_i)}{p_{\theta_i}(U_i)}\right].$$

We estimate the expectation using high-accuracy Monte Carlo integration.

Table 6: Verification of the exact KL decomposition across five structural configurations. D_{exo} is the sum of marginal exogenous KL divergences, $D_{\text{endo}} = D_{\text{KL}}(P_0 \| Q_\theta)$, and $r_{\text{KL}} = D_{\text{endo}}/D_{\text{exo}}$. Values are Monte Carlo estimates using $n = 500{,}000$ samples per configuration and seed 0; standard errors for D_{endo} are shown in parentheses.

α	β	Perturbed coordinate(s)	D_{exo}	D_{endo}	r_{KL}
-1.0	-1.0	A (2 descendants)	1.8499	1.8499 (0.00211)	1.0000
-1.0	-1.0	E (1 descendant)	0.0787	0.0787 (0.00055)	1.0000
-1.0	-1.0	I (0 descendants)	0.1956	0.1956 (0.00085)	1.0000
-1.0	-1.0	A, E, I	2.1243	2.1243 (0.00234)	1.0000
-0.5	0.5	A	1.8499	1.8499 (0.00211)	1.0000
-0.5	0.5	E	0.0787	0.0787 (0.00055)	1.0000
-0.5	0.5	I	0.1956	0.1956 (0.00085)	1.0000
-0.5	0.5	A, E, I	2.1243	2.1243 (0.00234)	1.0000
0.0	1.0	A	1.8499	1.8499 (0.00211)	1.0000
0.0	1.0	E	0.0787	0.0787 (0.00055)	1.0000
0.0	1.0	I	0.1956	0.1956 (0.00085)	1.0000
0.0	1.0	A, E, I	2.1243	2.1243 (0.00234)	1.0000
0.5	-0.5	A	1.8499	1.8499 (0.00211)	1.0000
0.5	-0.5	E	0.0787	0.0787 (0.00055)	1.0000
0.5	-0.5	I	0.1956	0.1956 (0.00085)	1.0000
0.5	-0.5	A, E, I	2.1243	2.1243 (0.00234)	1.0000
1.0	1.0	A	1.8499	1.8499 (0.00211)	1.0000
1.0	1.0	E	0.0787	0.0787 (0.00055)	1.0000
1.0	1.0	I	0.1956	0.1956 (0.00085)	1.0000
1.0	1.0	A, E, I	2.1243	2.1243 (0.00234)	1.0000

For the structural-compatibility experiment, the unconstrained joint adversary uses a correlated three-dimensional Gaussian base $Z \sim \mathcal{N}(0, \Sigma)$ with non-diagonal covariance, followed by the same coordinatewise marginal spline transformations. This preserves comparable marginal transformations while introducing cross-coordinate dependence unavailable to the structurally factorized adversary.

E.6 Verification of the Exact KL Decomposition

For each structural configuration, we perturb one or more exogenous marginals while keeping g fixed:

$$Q_{\boldsymbol{\theta}} = g_{\#}\left(P_{\theta_A} \otimes P_{\theta_E} \otimes P_{\theta_I}\right).$$

We compute

$$D_{\text{exo}} := \sum_{i \in \{A,E,I\}} D_{\text{KL}}(P_{U_i} \| P_{\theta_i})$$

and

$$D_{\text{endo}} := D_{\text{KL}}(P_0 \| Q_{\boldsymbol{\theta}}),$$

and report

$$r_{\text{KL}} := \frac{D_{\text{endo}}}{D_{\text{exo}}}.$$

Theorem 1 predicts $r_{\text{KL}} = 1$. We perturb U_A, U_E, and U_I individually to examine perturbations with two, one, and zero descendants, respectively, and also perturb all three coordinates jointly. The resulting KL-decomposition results are reported in Table 6.

As an independent numerical cross-check, we estimate the endogenous-space KL directly using the non-parametric k-nearest-neighbor estimator of Wang et al. (2009), without using g^{-1}. Let $x_1, \ldots, x_n \overset{\text{i.i.d.}}{\sim} P$ and $y_1, \ldots, y_m \overset{\text{i.i.d.}}{\sim} Q$ denote samples in $\mathbb{R}^d$, where $P = P_0$ and $Q = Q_\theta$. Let

Table 7: Independent k-NN cross-check of the endogenous-space KL divergence. D_{exo} is the exogenous KL sum, Analytic D_{endo} is the exact endogenous-space KL, and k-NN D_{endo} is estimated directly in the endogenous space using $n = 20{,}000$ samples and $k = 5$, without using g^{-1}.

α	β	D_{exo}	Analytic D_{endo}	k-NN D_{endo}
-1.0	-1.0	2.1243	2.1243	2.0505
0.0	1.0	2.1243	2.1243	2.0617
0.5	-0.5	2.1243	2.1243	2.0637
1.0	1.0	2.1243	2.1243	2.0528

Table 8: Multi-seed HSIC permutation-test p-values for $\widehat{U}_E \perp A$ under exoDRO ($n = 1{,}500$; 150 permutations per seed).

Seed	100	101	102	103	104	105	106	107	108	109	Mean
p-value	0.483	0.993	0.344	0.762	0.073	0.424	0.185	0.431	0.993	0.768	0.546

$\rho_k(x_i)$ denote the Euclidean distance from x_i to its k-th nearest neighbor among $\{x_j : j \neq i\}$, and let $\nu_k(x_i)$ denote the Euclidean distance from x_i to its k-th nearest neighbor among $\{y_j\}_{j=1}^m$. We use

$$\widehat{D}_{\text{KL}}(P\|Q) = \frac{d}{n}\sum_{i=1}^{n} \log \frac{\nu_k(x_i)}{\rho_k(x_i)} + \log \frac{m}{n-1}. \tag{37}$$

For the three-dimensional SCM considered here, $d = 3$, and we use $n = m = 20{,}000$ samples and $k = 5$. The independent k-NN cross-check is reported in Table 7. The estimator agrees with the exact value within approximately 3–4%.

E.7 Structural Compatibility Diagnostics

For samples from each adversary, we recover

$$\widehat{U}_A = A, \qquad \widehat{U}_E = E - \alpha A, \qquad \widehat{U}_I = I - \beta E.$$

Structural compatibility requires

$$\widehat{U}_E \perp A, \qquad \widehat{U}_I \perp (A, E).$$

Dependence is evaluated using HSIC with Gaussian kernels and permutation-based significance tests. Kernel bandwidths are selected using the median heuristic. The joint and structural generative adversaries are calibrated to approximately matched realized inclusive KL rather than matched raw reconstruction budgets. The representative results are reported in Table 2. We additionally repeat the $\widehat{U}_E \perp A$ test for exoDRO over 10 independent random seeds; the corresponding multi-seed results are reported in Table 8.

E.8 Evaluation and Reproducibility Protocol

The same nominal datasets, random-seed conventions, prediction architecture, and evaluation code are shared across competing methods whenever their formulations permit.

For the controlled KL analyses, tractable densities are used rather than learned divergence estimators, except for the independent k-nearest-neighbor cross-check in Appendix E.6. Structural dependence is evaluated using the same recovered residuals and HSIC procedure across adversaries.

For ambiguity sets based on different discrepancy geometries, raw robustness radii are not directly equated. Predictive-comparison hyperparameters are instead selected using OOD validation performance, while controlled diagnostic experiments report realized distributional discrepancies whenever calibration across ambiguity geometries is required.

F Additional Limitations

This section expands on the scope and limitations discussed in Section 8.

F.1 Structural-Model Misspecification

Our analysis assumes that the structural map g is known and fixed. When g must instead be estimated from data, errors in the estimated mechanisms or in the inverse map can directly affect the recovered exogenous variables. In particular, imperfect inversion may induce apparent dependencies among the recovered coordinates even when the true exogenous variables are independent. Extending the present analysis to an estimated structural map therefore requires explicitly accounting for structural-model and inversion error, potentially through a perturbation analysis similar in spirit to that considered in structural causal optimal transport (Ehyaei et al., 2025).

More generally, different SCMs may be observationally compatible with the same endogenous distribution while inducing different notions of structurally compatible distribution shift. exoDRO treats the SCM as given and therefore does not address uncertainty over the causal model itself. Distributional uncertainty within a fixed SCM and structural uncertainty over the SCM are distinct sources of uncertainty that we leave separate in the present work.

F.2 Non-Bijective Mechanisms and Latent Confounding

The exact KL decomposition requires the reduced-form structural map g to be bijective. If g is measurable but non-injective, the data-processing inequality gives only

$$D_{\mathrm{KL}}\left(g_{\#}P_U \,\|\, g_{\#}P_{\boldsymbol{\theta}}\right) \le D_{\mathrm{KL}}\left(P_U \,\|\, P_{\boldsymbol{\theta}}\right),$$

and the exact equality of Theorem 1 need not hold. Consequently, the exogenous divergence may still provide an upper bound, but it no longer exactly characterizes the endogenous-space KL divergence.

Our formulation also assumes mutually independent exogenous coordinates. Latent confounding violates this coordinatewise factorization. One possible extension is to partition the exogenous variables into mutually independent blocks while allowing dependence within each block:

$$P_U = \bigotimes_{b=1}^{B} P_{U_{\mathcal{G}_b}}.$$

The same argument would then yield KL additivity across independent blocks rather than across individual coordinates. Developing such a block-structured version of exoDRO is left for future work.

F.3 Mechanism Shifts

The ambiguity set considered in this work changes the exogenous distribution while keeping the structural mechanisms fixed:

$$P_U \longrightarrow Q_U, \qquad F \text{ fixed}.$$

It therefore models distribution shifts that are compatible with the same causal mechanisms, but does not cover interventions or environmental changes that alter the mechanisms themselves, $F \to F'$.

Handling simultaneous uncertainty in both the exogenous distribution and the structural mechanisms would require an ambiguity set over structural maps in addition to the generative ambiguity set over exogenous distributions. The exact KL decomposition established here does not directly transfer when the nominal and adversarial distributions are pushed forward through different maps.

F.4 Computational Scalability

exoDRO uses one adversarial generative model per exogenous coordinate in its basic form. Although this decomposition can make the statistical constraints easier to control, the computational cost may increase when the SCM contains many variables. In addition, each block update requires estimating an expectation over the remaining $n-1$ adversarial coordinates.

Several extensions may reduce this cost, including grouping related coordinates into blocks, sharing parameters across marginal generative models, or updating multiple approximately independent blocks in parallel. Such modifications may substantially improve scalability, but their optimization guarantees would require separate analysis because the fixed-block argument in Proposition 3 assumes that the remaining adversarial models are held fixed during each coordinate update.

F.5 Scope of the Empirical Evaluation

The controlled experiments are designed to isolate the theoretical properties of the proposed ambiguity construction under an exactly known and invertible SCM. They therefore avoid confounding the evaluation with errors from causal-structure discovery or mechanism estimation.

The predictive robustness experiments evaluate the method under held-out distribution shifts, but they do not resolve the separate problem of obtaining a reliable SCM in a real-world application. Evaluating exoDRO when the causal structure itself is learned, partially known, or misspecified is an important direction for future work.