Extended Wasserstein-GAN Approach to Causal Distribution Learning: Density-Free Estimation and Minimax Optimality

Shu Tamano[1,2] **Masaaki Imaizumi**[1,3,4]

[1] The University of Tokyo
[2] Japan Institute for Health Security
[3] RIKEN Center for Advanced Intelligence Project
[4] Kyoto University

tamano-shu212@g.ecc.u-tokyo.ac.jp, imaizumi@g.ecc.u-tokyo.ac.jp

Abstract

Distributional causal inference requires estimating not only average treatment effects but also interventional outcome distributions, including quantiles, tail risks, and policy-dependent uncertainty. As a method for distributional causal inference, generative adversarial network (GAN)-based counterfactual methods are flexible tools for this task. However, these methods have several limitations. First, the objectives of certain techniques do not coincide with the statistical risk of the identifiable causal target, and therefore provide limited theoretical guarantees regarding estimable counterfactual distributions or optimality. Second, implementing target-design risks under observational sampling can require density-ratio estimation. In this paper, we propose GANICE (GAN for Interventional Conditional Estimation) with several advantages: it (i) specifies the conditional interventional distribution for each treatment–covariate state as the causal estimation target; (ii) estimates the conditional distribution such that its averaged Wasserstein risk is minimized; (iii) attains minimax optimality. GANICE achieves these advantages through the extended Wasserstein distance and cell-normalized estimation, with minimax bounds based on Wasserstein GAN oracle inequalities and an anisotropic Assouad construction; Besov theory characterizes the cellwise critics. Our experiments show that GANICE outperforms existing methods on most benchmarks.

1 Introduction

Causal inference often targets average treatment effects, while many scientific and policy questions are distributional. Classical and recent work on counterfactual distributions, quantile treatment effects, conditional distributional treatment effects, and semiparametric counterfactual density estimation shows that such tasks require more than conditional means [Fir07, CFVM13, PSSM21, KBW23, KO23, BOY24, NPS26, JL26]. Under standard potential-outcomes assumptions, the identifiable target of a causal generative model is the family of conditional interventional outcome distributions, one for each treatment–covariate state. This family can be identified from observed outcomes, but the joint distribution of all potential outcomes cannot be identified without additional assumptions.

Generative adversarial networks (GANs) are promising for this problem because they can represent complex, non-Gaussian, and structured outcome distributions. For treatment-effect estimation, GANITE introduced adversarial counterfactual imputation for discrete treatments [YJvdS18b], SCIGAN extended this to continuous interventions [BJvdS20], and MCGAN generalized it to bi-

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison with representative GAN-based causal inference methods. Cond. Dist. marks explicit conditional-distribution targets, and No Ratio Est. marks methods avoiding density-ratio or propensity-score estimation. CE denotes task-specific adversarial cross-entropy objective; W_1 and eW_1 denote Wasserstein-1 and extended Wasserstein distances. Empty entries indicate unspecified properties.

Regime	Study	Cond. Dist.	No Ratio Est.	Metric	Optimality
(Baseline)	WGAN [SAL24]	–	–	W_1	✓
Disc.	GANITE [YJvdS18b]			CE	
Disc.	MCGAN [GHF$^+$20]	✓		CE	
Disc.	ITE-CAN [AZQ26]	✓		CE	
Disc.	GANICE (ours)	✓	✓	eW_1	✓
Cont.	SCIGAN [BJvdS20]			CE	
Cont.	MCGAN [GHF$^+$20]	✓		CE	
Cont.	GAD [LKL$^+$20]		✓	CE	
Cont.	GANICE (ours)	✓	✓	eW_1	✓

nary, categorical, and continuous treatments [GHF$^+$20]. Additionally, GAD uses adversarial deconfounding to learn balancing weights [LKL$^+$20]. Further models address sequential treatments and dosages, compromised nets, and mediation [HSS24, NI25, AZQ26, ZLZ$^+$26].

Despite this progress, the statistical foundations of GAN-based causal distribution learning remain limited. First, existing adversarial objectives for causal inference (e.g., [YJvdS18b, BJvdS20]) are algorithmically well defined, but their population values are rarely formulated as risks for the identifiable causal target. As a result, these objectives neither characterize the estimable counterfactual distributions nor guarantee optimality for the identifiable interventional distribution. Second, exact identification of a target-design interventional conditional distribution is naturally expressed through overlap reweighting. Direct empirical implementation may require density-ratio or propensity-score estimation, which can be unstable in continuous or high-dimensional treatment–covariate spaces. Similar limitations apply beyond adversarial methods. While powerful, flow- and diffusion-based methods such as INFs [MFF23], DiffPO [MMSF24], and PO-Flow [WIX25] lack minimax guarantees for the target-design-averaged causal risk and typically require tractable densities, invertibility, or conditional score modeling.

To address these issues, we propose GANICE (Generative Adversarial Network for Interventional Conditional Estimation), an extended Wasserstein GAN for causal distribution learning. It has three statistical advantages. (i) It specifies the estimable causal target: the generator learns the conditional interventional outcome law for each treatment–covariate state. (ii) It estimates the conditional distribution such that its averaged Wasserstein risk is minimized. The stratified estimator normalizes within treatment–covariate cells and aggregates their discrepancies under the target design, avoiding observational density, propensity score, or overlap density-ratio estimation. (iii) It attains minimax-optimal convergence rates, up to logarithmic factors, for the resulting causal distributional risk. The theory is density-free in the outcome space: it uses latent pushforward models and Wasserstein risk, with no Lebesgue density, likelihood, or conditional score. Overlap and latent-generator smoothness are retained; continuous conditioning additionally requires Wasserstein regularity across states.

GANICE achieves these advantages via three theoretical designs. First, the extended Wasserstein distance [CHSW25] uses a diagonal coupling to compare outcomes only at identical treatment–covariate states, aligning the loss causally and preventing cross-state transport. Second, its dual yields finite-resolution cellwise outcome-Lipschitz critics. These compare intra-cell outcome laws, while cell-normalized training preserves target-design aggregation without estimating density ratios. Third, we develop minimax theory for this objective. Under finite conditioning, we statewise-extend the Wasserstein GAN (WGAN) construction of [SAL24] using [TY23]'s lower-bound mechanism. Under continuous conditioning, we combine finite-resolution localization [Nob96, Sar17, LNB22, BFR23], anisotropic Wasserstein regularity, and Besov control of discontinuous critics [SN21]. Table 1 contrasts representative causal GANs; see Appendix A for detailed related work.

Our contributions are as follows:

- We characterize the conditional marginals identified by factual-coordinate adversarial imputation and formulate an adversarial objective equal to their target-design-averaged Wasserstein risk.
- Using the one-state WGAN construction of [SAL24], we establish finite-state minimax rates under observational sampling and target-design shift.
- For continuous conditioning, we prove a Wasserstein perturbation oracle, a comparison between cell-normalized training and exact causal risk, and a matching anisotropic Assouad lower bound. A vector-valued Besov embedding characterizes the localized critics.

1.1 Basic Notation

Probability measures reside on standard Borel spaces, and conditional distributions are fixed regular versions. Write $\mathcal{P}(S)$ for Borel probability measures on S, $\mathcal{L}(Z)$ for the law of a random element Z, and $g_{\#}\xi$ for the pushforward of ξ by measurable g. For $U \sim \mathrm{Unif}([0,1]^{d_U})$, write $g_{\#}U = g_{\#}\mathcal{L}(U)$; d_U is the latent dimension. W_1 denotes Wasserstein-1 distance on $\mathcal{Y}$. We use $\|\cdot\|$, $\mathrm{diam}(S)$, δ_s, and $\mathrm{Lip}(f)$ for the Euclidean norm, diameter, Dirac measure at s, and optimal Lipschitz constant. Write $\mathbb{N}_0 = \{0, 1, \ldots\}$, $a \vee b = \max\{a, b\}$, and $\mathbf{1}\{\cdot\}$ for indicators. $A \lesssim B$ means $A \leq CB$ for a model-parameter-dependent constant C, $A \asymp B$ means both comparisons hold, and $\tilde{O}(\cdot)$ hides logarithmic factors.

Let $X \in \mathcal{X}$ be covariates with law P_X, $T \in \mathcal{T}$ treatment, and $Y(t) \in \mathcal{Y} \subset \mathbb{R}^p$ the potential outcome under t. For fixed $K_0 > 0$, the theoretical outcome space is the closed Euclidean ball $\mathcal{Y} = B^p(0, 2K_0)$. For an integer $n \geq 2$, the i.i.d. observations $(W_i, Y_i^{\mathrm{obs}})$, $i = 1, \ldots, n$, are drawn from P_{obs}, with state $W = (X, T) \in \mathcal{W}$ and observed outcome Y^{obs}. The conditioning space is finite $\mathcal{W} = \{1, \ldots, M\}$ or continuous $\mathcal{W} = [0,1]^{d_W}$; p, d_W are the outcome and conditioning dimensions. Let Q_{obs}, Q_ρ be the observational and fixed known target marginals (designs) of W, and $w_\rho = \mathrm{d}Q_\rho/\mathrm{d}Q_{\mathrm{obs}}$ their density ratio when it exists. For $w = (x, t)$ and Borel $A \subset \mathcal{Y}$, the conditional interventional law is $\mu_w^*(A) = \mathbb{P}(Y(t) \in A \mid X = x)$. P_{obs} and P^* denote the observational and target interventional joint laws. For a generator $g : \mathcal{W} \times [0,1]^{d_U} \to \mathcal{Y}$, the generated conditional law is $\nu_{g,w} = g(w, \cdot)_{\#}U$, yielding $P_g(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_{g,w}(\mathrm{d}y)$.

For a partition Π of $\mathcal{W}$ with cells $C \in \Pi$, its target and observational masses are $q_C = Q_\rho(C)$ and $\pi_C = Q_{\mathrm{obs}}(C)$. For finite $\mathcal{W}$, we abbreviate $q_j = Q_\rho(\{j\})$, $\pi_j = Q_{\mathrm{obs}}(\{j\})$, and $\mu_j^* = \mu_{w=j}^*$. For an anisotropic dyadic partition $\Pi_{\boldsymbol{m}}$, $C_{\boldsymbol{m}}(w)$ is the cell containing w. Appendix B summarizes notation.

2 Problem Setup

We work in the potential-outcomes framework [Ney23, Rub74]. Our goal is to estimate the target interventional law P^*, equivalently the family $\{\mu_w^*\}_{w \in \mathcal{W}}$ under target weighting Q_ρ. All equalities involving conditional laws are interpreted up to the relevant design marginal.

Assumption 1 (Consistency). $Y^{\mathrm{obs}} = Y(T)$ almost surely (a.s.).

Assumption 2 (Ignorability). There exist a standard Borel random element V and a jointly measurable map F such that $Y(t) = F(X, t, V)$ simultaneously for all t almost surely, and $V \perp\!\!\!\perp T \mid X$.

Use the version $\mu_{(x,t)}^* = F(x, t, \cdot)_{\#}\mathcal{L}(V \mid X = x)$. Under Assumptions 1–2, the observed conditional distribution of Y^{obs} given $W = w$ is μ_w^*; see Appendix C. Hence, $P_{\mathrm{obs}}(\mathrm{d}w, \mathrm{d}y) = Q_{\mathrm{obs}}(\mathrm{d}w)\mu_w^*(\mathrm{d}y)$ and $P^*(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\mu_w^*(\mathrm{d}y)$.

Assumption 3 (Overlap). $Q_\rho \ll Q_{\mathrm{obs}}$, and $0 \leq w_\rho(w) \leq \kappa^{-1}$ Q_{obs}-a.s. for some $\kappa \in (0, 1]$.

For finite $\mathcal{W} = \{1, \ldots, M\}$, overlap reduces to $q_j \leq \kappa^{-1}\pi_j$ for every state j. Assumption 3 implies $\mathbb{E}_{P_{\mathrm{obs}}}[w_\rho(W)\varphi(W, Y^{\mathrm{obs}})] = \mathbb{E}_{P^*}[\varphi(W, Y)]$ for every bounded measurable $\varphi : \mathcal{W} \times \mathcal{Y} \to \mathbb{R}$. This identity is the causal identification step; the remaining analysis is statistical estimation of the conditional distribution.

2.1 What Existing Counterfactual GANs Identify

Many adversarial causal generators form a completed potential-outcome vector $\bar{\boldsymbol{Y}} = (\bar{Y}(1), \ldots, \bar{Y}(A))$ for finite treatments $\mathcal{T} = \{1, \ldots, A\}$ by imputing missing coordinates alongside the factual outcome, then train a discriminator to identify the factual treatment [YJvdS18b, GHF+20, HSS24]. Given a generator, the population discriminator estimates $r_a(x, \bar{y}) = \mathbb{P}(T = a \mid X = x, \bar{\boldsymbol{Y}} = \bar{y})$. Optimizing the generator drives the conditional distribution of $\bar{\boldsymbol{Y}}$ toward T-invariance given X. Under exact factual reconstruction ($\bar{Y}(T) = Y^{\mathrm{obs}}$ a.s.), this invariance yields $\mathcal{L}(\bar{Y}(a) \mid X = x) = \mathcal{L}(Y(a) \mid X = x)\ P_X$-almost surely for any treatment a with positive propensity. Thus, this approach identifies treatment-specific conditional marginals but not their conditional joint law: any measurable coupling of these marginals is compatible with the observed law, factual reconstruction, and invariance (see Appendix C for a formal proof).

3 Proposed Method

3.1 Preparation: Extended Wasserstein Distance

This subsection defines the distributional loss used to train the generator. Since the identifiable target is the family of conditional interventional distributions, the loss must compare outcome distributions at the same treatment–covariate state.

Let μ_w, ν_w be the conditional outcome kernels of $P, R \in \mathcal{P}(\mathcal{W} \times \mathcal{Y})$. Write $\Gamma_\Delta(P, R)$ for couplings of P, R supported on $\{w = w'\}$.

Definition 1 (Extended Wasserstein distance [CHSW25]). For laws P, R with common conditioning marginal Q, the extended Wasserstein-1 distance is

$$\mathrm{eW}_1(P, R) \coloneqq \inf_{\alpha \in \Gamma_\Delta(P,R)} \int \|y - y'\|\, \alpha(\mathrm{d}w, \mathrm{d}y, \mathrm{d}w', \mathrm{d}y').$$

Formally, for a coupling $\alpha \in \Gamma_\Delta(P, R)$ on $(\mathcal{W} \times \mathcal{Y})^2$, the two state coordinates must satisfy $W' = W$ almost surely; equivalently, the (W, W')-marginal is the pushforward $\Delta_\# Q$ under the diagonal map $\Delta(w) = (w, w)$. This diagonal constraint is the causal restriction: an outcome observed at state w can be matched only to a generated outcome at the same w. On standard Borel $\mathcal{W}$ and compact $\mathcal{Y}$, it yields the statewise identity $\mathrm{eW}_1(P, R) = \int_{\mathcal{W}} W_1(\mu_w, \nu_w)\, Q(\mathrm{d}w)$. Thus, eW_1 is exactly the conditioning-marginal average of ordinary outcome-space Wasserstein distances; details are in Appendix D.1.

For the dual representation used below, fix an anchor $y_0 \in \mathcal{Y}$ and define $\mathcal{L}_1 \coloneqq \{f : \mathcal{Y} \to \mathbb{R} : f(y_0) = 0,\ \mathrm{Lip}(f) \le 1\}$. The critic class is defined as $\mathcal{F} \coloneqq \{h : \mathcal{W} \times \mathcal{Y} \to \mathbb{R} : h \text{ is jointly measurable},\ h(w, \cdot) \in \mathcal{L}_1 \text{ for } Q\text{-a.e. } w\}$. Anchoring loses nothing, since adding any $c(w)$ to h cancels under the common Q, and compactness gives $|h(w, y)| \le \mathrm{diam}(\mathcal{Y})$ for every y and Q-almost every w. For causal risks, $\mathcal{F}$ uses $Q = Q_\rho$. Extended Kantorovich–Rubinstein duality gives

$$\mathrm{eW}_1(P, R) = \sup_{h \in \mathcal{F}} \{\mathbb{E}_P[h(W, Y)] - \mathbb{E}_R[h(W, Y)]\}. \tag{1}$$

The constraint in eW_1 is stronger than ordinary joint optimal transport. If $\mathcal{W}$ has a bounded metric $d_{\mathcal{W}}$ and $W_1^{\otimes}$ denotes Wasserstein-1 distance on $\mathcal{W} \times \mathcal{Y}$ with product cost $d_{\mathcal{W}}(w, w') + \|y - y'\|$, then every diagonal coupling is an admissible joint coupling, yielding

$$W_1^{\otimes}(P, R) \le \mathrm{eW}_1(P, R). \tag{2}$$

Hence, eW_1 convergence implies joint Wasserstein convergence, while forbidding transport across distinct causal states.

3.2 GANICE: GAN for Interventional Conditional Estimation

This subsection presents the proposed method, GANICE, and introduces the population adversarial identity along with the stratified finite-resolution estimators. The population identity uses overlap reweighting to identify the exact causal risk, whereas the empirical estimators avoid estimating w_ρ by performing normalization within cells.

For a critic h, define $L(g,h) \coloneqq \mathbb{E}_{P_{\mathrm{obs}}}[w_\rho(W)h(W,Y^{\mathrm{obs}})] - \mathbb{E}[h(\tilde{W}, g(\tilde{W},U))]$, where $\tilde{W} \sim Q_\rho$ independently of U. By the identity in Section 2 and (1),

$$\sup_{h\in\mathcal{F}} L(g,h) = \mathrm{eW}_1(P^*, P_g). \tag{3}$$

Thus the exact population adversarial risk is the extended Wasserstein distance to the identifiable causal target.

The empirical estimator uses partition-based conditional measures, normalizing each cell by its sample count and weighting by its target mass, following the same principle as histogram- and partition-based estimators [Nob96, Sar17, LNB22, BFR23], with partitions defined as singletons for finite conditioning and anisotropic dyadic cells for continuous conditioning (see Appendix B.1). For a finite measurable partition Π of $\mathcal{W}$ and a cell $C \in \Pi$, let $N_C^{\mathrm{obs}} = \sum_{i=1}^n \mathbf{1}\{W_i \in C\}$ be the observed count in C. For $f : \mathcal{Y} \to \mathbb{R}$, define

$$\hat{E}_C^{\mathrm{obs}}[f] \coloneqq \begin{cases} (N_C^{\mathrm{obs}})^{-1} \sum_{i:W_i\in C} f(Y_i^{\mathrm{obs}}), & N_C^{\mathrm{obs}} > 0, \\ f(y_0), & N_C^{\mathrm{obs}} = 0. \end{cases} \tag{4}$$

This cell-normalized average is the basic object that removes the need to estimate the density ratio.

Finite conditioning. When $\mathcal{W} = \{1,\dots,M\}$, take $\Pi = \{\{1\},\dots,\{M\}\}$ and use j to index the singleton cell $\{j\}$. Let $\mathcal{G}_N$ and $\mathcal{D}_N$ be the generator and critic classes from the unconditional WGAN estimator of [SAL24], indexed by a deterministic sieve parameter $N \geq 2$. In the finite-conditioning setting, we take $N = n$ and set

$$\mathcal{G}_n^{\mathrm{disc}} = \{g(j,u) = g_j(u) : g_j \in \mathcal{G}_n,\ \forall j\}, \quad \mathcal{H}_n^{\mathrm{disc}} = \{h(j,y) = D_j(y) : D_j \in \mathcal{D}_n,\ \forall j\}.$$

Let $\{U_{\ell j} : \ell = 1,\dots,n, j = 1,\dots,M\}$ be i.i.d. copies of the noise variable U, independent of the observations. For $h(j,y) = D_j(y)$, define $\hat{L}_n^{\mathrm{disc}}(g,h)$ and the finite-state GANICE estimator $\hat{g}_n^{\mathrm{disc}}$ as follows:

$$\hat{L}_n^{\mathrm{disc}}(g,h) \coloneqq \sum_{j=1}^M q_j \left[\hat{E}_j^{\mathrm{obs}}[D_j] - \frac{1}{n}\sum_{\ell=1}^n D_j(g_j(U_{\ell j}))\right], \quad \hat{g}_n^{\mathrm{disc}} \in \operatorname*{arg\,min}_{g\in\mathcal{G}_n^{\mathrm{disc}}} \sup_{h\in\mathcal{H}_n^{\mathrm{disc}}} \hat{L}_n^{\mathrm{disc}}(g,h).$$

In finite states, stratification is exact: conditional on $W = j$, observed outcomes are sampled from μ_j^*, and q_j only weights the statewise risks.

Continuous conditioning. When $\mathcal{W} = [0,1]^{d_W}$, the exact class $\mathcal{F}$ is too rich because its Kantorovich potential may vary freely with w. We therefore localize critics on treatment–covariate cells while allowing the generator to remain a general conditional model.

Let $\boldsymbol{m} = (m_1,\dots,m_{d_W}) \in \mathbb{N}_0^{d_W}$ and $|\boldsymbol{m}|_1 = \sum_j m_j$. The anisotropic dyadic partition $\Pi_{\boldsymbol{m}}$ divides $[0,1]^{d_W}$ into axis-aligned rectangles with side length 2^{-m_j} in coordinate j. The population finite-resolution critic class is $\mathcal{F}^{(\boldsymbol{m})} \coloneqq \{h(w,y) = \sum_{C\in\Pi_{\boldsymbol{m}}} \mathbf{1}\{w \in C\} f_C(y) : f_C \in \mathcal{L}_1\}$. It defines the cell-resolution version of eW_1: one outcome-Lipschitz potential is chosen per cell, so the induced loss compares cell-averaged outcome laws and averages the resulting W_1 distances using the target masses q_C. For estimation, let $N_{\boldsymbol{m}} \coloneqq \lceil \kappa n 2^{-|\boldsymbol{m}|_1} \rceil \vee 2$ be the deterministic effective cell sample size. Using the one-state outcome-critic class $\mathcal{D}_{N_{\boldsymbol{m}}}$ from Appendix E.1, set

$$\mathcal{H}_{n,\boldsymbol{m}}^{\mathrm{cont}} \coloneqq \Big\{h(w,y) = \sum_{C\in\Pi_{\boldsymbol{m}}} \mathbf{1}\{w \in C\} D_C(y) : D_C \in \mathcal{D}_{N_{\boldsymbol{m}}}\Big\}.$$

Thus $\mathcal{F}^{(\boldsymbol{m})}$ defines the finite-resolution loss geometry, and $\mathcal{H}_{n,\boldsymbol{m}}^{\mathrm{cont}}$ is its regularized empirical critic class. Let $\mathcal{G}_n^{\mathrm{cont}}$ be the implemented conditional generator class, with the measurability conditions specified in Appendix F.4.

Draw $\tilde{W}_1,\dots,\tilde{W}_n$ i.i.d. from Q_ρ and independent copies $U_1,\dots,U_n$ of U, with both samples independent of the observations and of each other. For $C \in \Pi_{\boldsymbol{m}}$, let $M_C^\rho \coloneqq \sum_{\ell=1}^n \mathbf{1}\{\tilde{W}_\ell \in C\}$ be the target-design count in C and define the generated cell average for $f : \mathcal{Y} \to \mathbb{R}$:

$$\hat{E}_C^g[f] \coloneqq \begin{cases} (M_C^\rho)^{-1} \sum_{\ell:\tilde{W}_\ell\in C} f(g(\tilde{W}_\ell, U_\ell)), & M_C^\rho > 0, \\ f(y_0), & M_C^\rho = 0. \end{cases} \tag{5}$$

For $h \in \mathcal{H}^{\mathrm{cont}}_{n,\boldsymbol{m}}$ with cell components D_C, the continuous stratified objective $\hat{L}^{\mathrm{cont}}_{n,\boldsymbol{m}}(g,h)$ and the continuous-conditioning GANICE estimator $\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}$ are defined as follows:

$$\hat{L}^{\mathrm{cont}}_{n,\boldsymbol{m}}(g,h) \coloneqq \sum_{C\in\Pi_{\boldsymbol{m}}} q_C\big[\hat{E}^{\mathrm{obs}}_C[D_C] - \hat{E}^{g}_C[D_C]\big], \quad \hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}} \in \operatorname*{arg\,min}_{g\in\mathcal{G}^{\mathrm{cont}}_n} \sup_{h\in\mathcal{H}^{\mathrm{cont}}_{n,\boldsymbol{m}}} \hat{L}^{\mathrm{cont}}_{n,\boldsymbol{m}}(g,h). \tag{6}$$

The empirical objective learns from observed outcomes within cells and aggregates cellwise discrepancies under Q_ρ, without estimating w_ρ.

4 Minimax Optimality: Discrete Conditioning

We first consider finite conditioning. In this setting, stratification is exact: conditional on $W = j$, observed outcomes are sampled from μ^*_j, and no density-ratio estimate is needed. Using the finite-state notation from Section 1.1,

$$\mathrm{eW}_1(P^*, P_g) = \sum_{j=1}^{M} q_j W_1(\mu^*_j, \nu_{g,j}). \tag{7}$$

Thus the causal distributional risk decomposes into M local unconditional generative estimation problems, aggregated under the target design Q_ρ.

For $s > 0$, $\mathcal{H}^s_{K_0}$ is the periodic Hölder ball on the unit torus $\mathbb{T}^{d_U} = \mathbb{R}^{d_U}/\mathbb{Z}^{d_U}$: maps into $B^p(0, K_0)$ have derivatives through order $\lfloor s \rfloor$ bounded by K_0 and, for noninteger s, top-order Hölder seminorm of order $s - \lfloor s \rfloor$ at most K_0. Throughout the rate results, $\beta \geq 1$ and $1 \leq d_U \leq p$. For fixed $(M, q_{\min}, \kappa, \beta, d_U, p, K_0)$, let $\mathcal{M}_{\mathrm{disc}}$ be the class of finite-state causal models satisfying Assumptions 1–3, $q_j \geq q_{\min}$ with $0 < q_{\min} \leq 1/M$, $\mu^*_j = (g^*_j)_{\#}U$, and $g^*_j \in \mathcal{H}^{\beta+1}_{K_0}$ for all $j = 1, \ldots, M$. The generator and critic classes are obtained by applying the local WGAN construction of [SAL24] independently to each state.

Theorem 1 (Finite-state upper bound)**.** *Fix $M < \infty$ and define $a \coloneqq \min\{(\beta+1)/(2\beta+d_U), 1/2\}$. There exist constants $C, c > 0$, depending only on $(M, q_{\min}, \kappa, \beta, d_U, p, K_0)$, such that*

$$\sup_{(P_{\mathrm{obs}},P^*)\in\mathcal{M}_{\mathrm{disc}}} \mathbb{E}\big[\mathrm{eW}_1(P^*, P_{\hat{g}^{\mathrm{disc}}_n})\big] \leq C(\log n)^c n^{-a}.$$

The exponent a is the local one-state WGAN exponent: $(\beta + 1)/(2\beta + d_U)$ is the nonparametric latent-pushforward rate determined by smoothness and latent dimension, capped by the parametric barrier $1/2$. Because M is fixed, finite conditioning does not change the exponent; overlap only ensures enough observations in each target-relevant state.

Define the minimax risk as $\mathcal{R}^{\mathrm{disc}}_n = \inf_{\hat{P}_n} \sup_{(P_{\mathrm{obs}},P^*)\in\mathcal{M}_{\mathrm{disc}}} \mathbb{E}[\mathrm{eW}_1(P^*, \hat{P}_n)]$, where the infimum is taken over all estimators based on the observational sample that output laws with conditioning marginal Q_ρ.

Corollary 1 (Finite-state minimax optimality)**.** *There exist constants $0 < c_L < C_U < \infty$ and $c > 0$ such that, for all n, $c_L n^{-a} \leq \mathcal{R}^{\mathrm{disc}}_n \leq C_U(\log n)^c n^{-a}$.*

The rate is obtained in the causal loss eW_1, hence it controls the target-weighted statewise error $\sum_j q_j W_1(\mu^*_j, \nu_{\hat{g}^{\mathrm{disc}}_n,j})$ and, by (2), ordinary joint Wasserstein distance. This is a minimax statement for the identifiable interventional distribution: no estimator can uniformly recover the causal distribution faster under the same identification and overlap conditions. By Kantorovich–Rubinstein duality, it also controls target-weighted errors of all statewise 1-Lipschitz distributional functionals; quantile and tail summaries follow under the usual functional-specific regularity conditions.

Remark 1 (Statewise and joint transport). Consider scalar outcomes with $K_0 \geq 1$, two states at distance $\varepsilon \in (0, 1)$, and $Q_\rho = (\delta_{w_1} + \delta_{w_2})/2$. Let $\mu^*_{w_1} = \delta_0$ and $\mu^*_{w_2} = \delta_1$, and exchange these conditional laws in R. Then $\mathrm{eW}_1(P^*, R) = 1$ and $W^{\otimes}_1(P^*, R) = \varepsilon$. Thus a small joint transport cost can conceal a large statewise error as state separation decreases. For a fixed finite state metric, a reverse comparison holds with a constant depending on its minimum separation (Appendix D.1); our statewise bound is independent of that metric.

4.1 Proof Sketch

The proof reduces the risk to statewise WGAN risks via (7). The first key point is that finite-state stratification preserves the target: conditioning on $W = j$ yields samples from μ_j^*, while q_j weights the resulting risks. Thus, overlap lower-bounds local sample sizes ($\pi_j \geq \kappa q_j$) rather than building inverse-probability-weighted empirical processes.

The second key point is combining random state counts with [SAL24]. Given these counts, each state is an unconditional WGAN; binomial negative-moment bounds convert them into deterministic rates. The restricted critic construction yields the latent-pushforward rate n^{-a} by controlling the empirical process through the latent dimension. The lower bound restricts to a submodel with $Q_{\mathrm{obs}} = Q_\rho$ in which only one state's conditional law varies, directly transferring the unconditional lower bound of [TY23, SAL24] to eW_1. Full details are in Appendix E.

5 Minimax Optimality: Continuous Conditioning

We next analyze the continuous-conditioning estimator (6). It remains stratified and cell-normalized: w_ρ is never estimated, and overlap only controls the number of observations in target-relevant cells.

Assumption 4 (Design regularity). In the continuous-conditioning setting, Q_ρ has a density q_ρ with respect to Lebesgue measure satisfying $0 < \underline{q} \leq q_\rho(w) \leq \bar{q} < \infty$ for Lebesgue-almost every w.

Assumption 4 is a standard covariate-design condition in local nonparametric conditional estimation [FYT96, HWY99, CCJM24]. It is imposed on the target design, not on the outcome law: the conditional outcome distributions may still be implicit or singular. Together with Assumption 3, it gives expected observational cell counts of order at least $\kappa n 2^{-|\boldsymbol{m}|_1}$ in target-relevant cells.

Assumption 5 (Pointwise latent pushforward model). For every $w \in \mathcal{W}$, $\mu_w^* = (g_w^*)_\# U$ for some $g_w^* \in \mathcal{H}_{K_0}^{\beta+1}$.

This is a density-free local model for the outcome distribution and allows μ_w^* to be singular with respect to Lebesgue measure.

Assumption 6 (Anisotropic conditional regularity). There exist $\alpha_1, \ldots, \alpha_{d_W} \in (0, 1]$ and $0 < L_* < \infty$ such that $W_1(\mu_w^*, \mu_{w'}^*) \leq L_* \sum_{j=1}^{d_W} |w_j - w'_j|^{\alpha_j}$ for all $w, w' \in \mathcal{W}$.

Assumption 6 is the Wasserstein analogue of covariate-smoothness conditions used in conditional density and distribution estimation [LNB22, TLY25, HWL$^+$25]. It is stated directly in W_1, so it controls finite-resolution localization bias without requiring an outcome density or score.

For $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{d_W})$, the localization bias scale, effective anisotropic dimension, and rate exponent are $b_{\boldsymbol{m}} \coloneqq \sum_{j=1}^{d_W} 2^{-\alpha_j m_j}$, $\bar{d} \coloneqq \sum_{j=1}^{d_W} \alpha_j^{-1}$, $r \coloneqq (a^{-1} + \bar{d})^{-1}$. The optimized resolution $\boldsymbol{m}_n = (m_{n,j})_{j=1}^{d_W}$ has $m_{n,j} \coloneqq \lfloor r \log_2(\kappa n)/\alpha_j \rfloor \vee 0$. The exponent r captures a trade-off between the local difficulty of outcome generation and the cost of conditioning. The local difficulty is quantified by a^{-1}, inherited from the one-state WGAN estimation setting, while the conditioning cost is given by $\bar{d} = \sum_j \alpha_j^{-1}$. Consequently, smaller α_j or an increase in the number of conditioning coordinates lead to a slower convergence rate. The effect of weak overlap appears as the effective sample size factor κn.

For fixed $(\kappa, \beta, d_U, p, d_W, K_0, \boldsymbol{\alpha}, L_*, \underline{q}, \bar{q})$, let $\mathcal{M}_{\mathrm{cont}}$ denote the model class satisfying Assumptions 1–3 and Assumptions 4–6, with the specified known Q_ρ and $d_W \geq 1$.

For the continuous-conditioning risk bounds, write $\hat{P}_{n,\boldsymbol{m}}^{\mathrm{cont}} \coloneqq P^{(\boldsymbol{m})}_{\hat{g}_{n,\boldsymbol{m}}^{\mathrm{cont}}}$, where $P_g^{(\boldsymbol{m})}$ replaces $\nu_{g,w}$ on each cell $C \in \Pi_{\boldsymbol{m}}$ by its Q_ρ-average (Appendix F.1).

Theorem 2 (Continuous-conditioning upper bound)**.** *Assume the implemented estimator satisfies the finite-resolution transfer condition* (24) *uniformly over $\mathcal{M}_{\mathrm{cont}}$ at $\boldsymbol{m}_n$, with error at most $C_0(\log n)^{c_0}(\kappa n)^{-r}$ for fixed $C_0 > 0, c_0 \geq 0$. Then there exist constants $C, c > 0$, depending only on fixed model parameters and C_0, c_0, such that*

$$\sup_{(P_{\mathrm{obs}}, P^*) \in \mathcal{M}_{\mathrm{cont}}} \mathbb{E}\big[\mathrm{eW}_1(P^*, \hat{P}_{n,\boldsymbol{m}_n}^{\mathrm{cont}})\big] \leq C(\log n)^c (\kappa n)^{-r}. \tag{8}$$

The finite-resolution inequality underlying Theorem 2 is proved in Appendix F.2. It decomposes the risk into localization bias and a local WGAN stochastic term with effective cell sample size

$\kappa n 2^{-|\boldsymbol{m}|_1}$. Thus weak overlap reduces the effective sample size, but does not require estimating the density ratio. The bound compares the cell-averaged generated law with the exact target P^*. The transfer condition bounds the gap between the cell laws of $\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}$ and those of the routed estimator, which fits one generator per cell; for the latter it is zero and the bound holds before cell averaging (Corollary 2).

Define the minimax risk $\mathcal{R}^{\mathrm{cont}}_n := \inf_{\hat{P}_n} \sup_{(P_{\mathrm{obs}},P^*)\in\mathcal{M}_{\mathrm{cont}}} \mathbb{E}_{P_{\mathrm{obs}}^{\otimes n}}[\mathrm{eW}_1(P^*,\hat{P}_n)]$, where the infimum is over all estimators of the form $\hat{P}_n(\mathrm{d}w,\mathrm{d}y) = Q_\rho(\mathrm{d}w)\hat{\mu}_{n,w}(\mathrm{d}y)$.

Theorem 3 (Continuous-conditioning lower bound)**.** *There exists a constant $c_L > 0$, depending only on fixed model parameters, such that for all sufficiently large n, $\mathcal{R}^{\mathrm{cont}}_n \geq c_L(\kappa n)^{-r}$.*

Theorems 2–3 give minimax optimality up to logarithmic factors for the routed estimator, with an additive implementation-transfer error under the statewise causal risk eW_1.

Remark 2 (Comparison with conditional diffusion rates). For covariate and response dimensions D_X, D_Y with Hölder smoothness α_X, α_Y, [TLY25] obtain the TV rate $\tilde{O}(n^{-1/(2+D_X/\alpha_X+D_Y/\alpha_Y)})$ and a Wasserstein-1 rate involving the response density term D_Y/α_Y; their manifold result replaces (D_X, D_Y) by intrinsic dimensions and is stated in W_1 because the relevant laws may be mutually singular. [HWL+25] obtain TV-type rates for conditional diffusion transformers under Hölder density/score assumptions. Our rate replaces response density complexity by the latent-pushforward WGAN exponent a, separates the conditioning cost as $\bar{d}$, and measures the causal statewise risk eW_1. Hence no outcome density or conditional score is required.

5.1 Proof Sketch

The upper bound relies on a finite-resolution reduction rather than inverse-probability-weighted empirical processes. Within a cell C, observations follow $\mu^{\mathrm{obs}}_C = \pi_C^{-1}\int_C \mu^*_w Q_{\mathrm{obs}}(\mathrm{d}w)$, while the target averages under Q_ρ. The first key point is the proxy-to-exact inequality: anisotropic W_1 regularity ensures replacing the Q_ρ-cell mixture with μ^{obs}_C incurs only localization bias. This avoids estimating w_ρ while controlling the exact causal eW_1 risk.

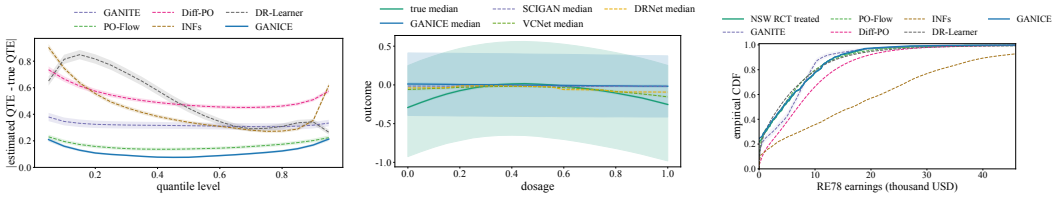
The second key point is statistical localization combined with Besov control of the critic geometry. Conditional on N^{obs}_C and M^ρ_C, each routed cell estimator is an unconditional WGAN targeting μ^{obs}_C, which lies within $L_* b_{\boldsymbol{m}}$ of a cell-representative law $\mu^*_{w_C}$ and is controlled by the perturbation oracle. Overlap and regularity yield $n\pi_C \gtrsim \kappa n 2^{-|\boldsymbol{m}|_1}$; summing one-state oracle inequalities over cells gives the stochastic error. The vector-valued Besov–Nikolskii embedding characterizes the jumps of the localized critics at partition boundaries [Tri83, Suz19, SN21]. A single shared network adds the transfer error (Appendix F.4), which vanishes when the cellwise fits are used directly. Optimizing anisotropic resolution balances bias and stochastic error, yielding (8).

The lower bound uses an anisotropic Assouad construction. Boxes have side lengths proportional to target separation raised to $1/\alpha_j$, each containing a one-state WGAN hard family. Smooth transport perturbations preserve the latent-generator class and anisotropic W_1 regularity, with Hellinger bounds uniform over the smooth cutoff transitions and local sample sizes scaling as κn times box volume. Balancing local WGAN difficulty with anisotropic conditioning yields $r = (a^{-1} + \bar{d})^{-1}$. Full details are in Appendix F.

Remark 3 (Causal role of anisotropic Besov critics). Finite-resolution critics are piecewise constant in w and Lipschitz in y. Appendix F.1 embeds the maps $w \mapsto h(w,\cdot)$ into anisotropic Besov–Nikolskii balls $B^{\boldsymbol{s}}_{p_B,\infty}$ with smoothness $s_j < 1/p_B$, the range in which hyperplane jumps remain L^{p_B}-Besov admissible [Tri83, NvS97, KLP01, HL02, Suz19, SN21]. The embedding bounds the critic norm by $C\max_j 2^{m_j s_j}$ and accommodates jumps at partition boundaries while preserving the outcome Lipschitz constraint in the conditional Wasserstein dual.

6 Extension to IPM GANs

The finite-resolution argument also applies to other adversarial losses. Let $\mathcal{V}$ be a symmetric, uniformly bounded class of measurable $f : \mathcal{Y} \to \mathbb{R}$ containing zero, with a countable subclass attaining its suprema, and local integral probability metric (IPM) $d_{\mathcal{V}}(\mu,\nu) := \sup_{f\in\mathcal{V}} \int f \mathrm{d}(\mu-\nu)$. For P, R sharing Q_ρ, let $\mu^\rho_C = q_C^{-1}\int_C \mu_w Q_\rho(\mathrm{d}w)$ and ν^ρ_C denote their target-design cell mixtures. The



(a) IHDP quantile-effect error (b) TCGA dose-wise quantiles (c) Jobs randomized CDF

Figure 1: Distributional diagnostics. (a) Absolute quantile treatment-effect error as a function of quantile level on IHDP. (b) Dose-indexed medians and predictive quantile bands on TCGA. (c) Model-implied treated-arm CDF against the RCT treated-arm CDF on Jobs.

finite-resolution conditional IPM is

$$d_{\mathcal{V}}^{(\boldsymbol{m})}(P,R) := \sup_{h(w,y)=\sum_{C\in\Pi_{\boldsymbol{m}}} \mathbf{1}\{w\in C\}f_C(y),\ f_C\in\mathcal{V}} \{\mathbb{E}_P[h]-\mathbb{E}_R[h]\}.$$

This equals $\sum_{C\in\Pi_{\boldsymbol{m}}} q_C d_{\mathcal{V}}(\mu_C^\rho,\nu_C^\rho)$. For the routed IPM estimator $\hat{P}^{\mathcal{V}}_{n,\boldsymbol{m}}$ defined in Appendix G, a one-state perturbation oracle with exponent $a_{\mathcal{V}}>0$ and anisotropic Hölder regularity of $w\mapsto\mu_w^*$ under $d_{\mathcal{V}}$ with exponents $\boldsymbol{\alpha}$ give

$$\mathbb{E}[d_{\mathcal{V}}^{(\boldsymbol{m})}(P^*,\hat{P}^{\mathcal{V}}_{n,\boldsymbol{m}})] \lesssim b_{\boldsymbol{m}} + (\log n)^c(\kappa n 2^{-|\boldsymbol{m}|_1})^{-a_{\mathcal{V}}}.$$

Optimizing yields exponent $r_{\mathcal{V}}=(a_{\mathcal{V}}^{-1}+\bar{d})^{-1}$. Thus the construction covers Wasserstein critics, Hölder IPMs, MMD critics, and outcome-Besov IPMs whenever the corresponding one-state oracle inequality holds. Appendix G gives the cell-mixture definitions and formal assumptions.

7 Experiments

We evaluate whether GANICE estimates the full conditional interventional law, rather than only conditional means. We use two semi-synthetic benchmarks with known conditional outcome laws and one real benchmark with randomized-trial validation: the Infant Health and Development Program (IHDP) for binary treatment, The Cancer Genome Atlas (TCGA) for continuous dosage, and Jobs for real-world validation. IHDP is built from the covariates and treatment assignments used in counterfactual representation learning [SJS17, Hil11]; TCGA follows the high-dimensional gene-expression benchmark used for continuous-dose causal learning [BJvdS20, SLB$^+$20]; Jobs combines the National Supported Work (NSW) randomized job-training experiment with Panel Study of Income Dynamics (PSID) controls [LaL86, DW99]. Detailed data construction, preprocessing, target designs, and hyperparameters are given in Appendix H.

Baselines and metrics. For IHDP and Jobs, we compare with GANITE [YJvdS18b], PO-Flow [WIX25], DiffPO [MMSF24], individualized normalizing flows (INFs) [MFF23], DR-Learner [Ken23], CoDiTE [PSSM21], DNet [WSL$^+$23], Causal-DRF [NPS26], and SKCD (adapted) [JL26]. Appendix H.2 describes the predictive adaptation used for SKCD. For TCGA, we compare with SCIGAN [BJvdS20], DRNet [SLB$^+$20], and VCNet [NYLN21]. For the two semi-synthetic benchmarks, the primary metric is empirical extended Wasserstein error against the known interventional law. This metric is the empirical analogue of the causal risk analyzed in our theory: it averages statewise Wasserstein distances between the true and generated outcome laws at the same treatment–covariate state under the target design. We therefore use it as the headline criterion instead of point-estimation errors, which only evaluate scalar functionals and can miss discrepancies in spread, tails, quantiles, or multimodality. For Jobs, individual counterfactual distributions are not observed; we instead compare model-implied interventional cumulative distribution functions (CDFs) with arm-level CDFs from the held-out NSW randomized controlled trial (RCT) sample.

Results. Table 2 shows GANICE achieves the lowest distributional error on TCGA and Jobs, while Causal-DRF and CoDiTE achieve lower errors on IHDP. This pattern is consistent with the strengths of local kernel and forest averaging [PSSM21, NPS26]: our IHDP construction combines binary treatment, 25 covariates, and smoothly varying conditional laws. Pooling information across similar

Table 2: Main distributional results over 100 repetitions, with standard errors in parentheses (lower is better). IHDP and TCGA use empirical extended Wasserstein error, and Jobs uses RCT-assisted arm-level Wasserstein error on the asinh-transformed earnings scale.

Method	IHDP	TCGA	Jobs
GANITE	0.762 (0.039)	–	0.973 (0.044)
PO-Flow	0.433 (0.007)	–	0.299 (0.011)
DiffPO	0.323 (0.005)	–	0.704 (0.015)
INFs	0.355 (0.005)	–	1.309 (0.011)
DR-Learner	1.056 (0.011)	–	0.365 (0.007)
CoDiTE	<u>0.221</u> (0.003)	–	<u>0.293</u> (0.009)
DNet	0.637 (0.009)	–	0.356 (0.008)
Causal-DRF	**0.203** (0.002)	–	0.553 (0.010)
SKCD (adapted)	0.287 (0.003)	–	0.376 (0.011)
SCIGAN	–	<u>0.401</u> (0.002)	–
DRNet	–	0.426 (0.001)	–
VCNet	–	0.574 (0.005)	–
GANICE (ours)	0.286 (0.004)	**0.378** (0.005)	**0.209** (0.007)

covariate profiles can stabilize conditional distribution estimates in this setting. These benchmarks illustrate settings where scalar summaries fall short: IHDP features heterogeneous non-Gaussian outcomes, TCGA requires a treatment–dosage-indexed law, and Jobs evaluates agreement with randomized arm-level distributions. Figure 1 illustrates these effects. On IHDP, GANICE yields the smallest absolute average quantile-treatment-effect error among displayed methods across nearly all quantiles, including tails, capturing distributional heterogeneity beyond mean contrasts. On TCGA, baselines track the central trajectory but miss the true law's broad uncertainty band; GANICE instead generates a quantile band covering the central mass across dosages (despite residual bias at the dosage boundaries), giving the best integrated error (Table 2). On Jobs, the treated-arm CDF generated by GANICE closely tracks the randomized NSW CDF over the earnings distribution's main body and upper tail, whereas the displayed baselines are shifted or more widely dispersed. Additional metrics, calibration diagnostics, and objective ablations in Appendix H.5 indicate that these improvements are not specific to the extended Wasserstein metric. Implementation diagnostics and sensitivity analyses are reported in Appendices H.6 and H.7.

8 Conclusion

We proposed GANICE, an extended Wasserstein GAN for causal distribution learning that targets the identifiable family of conditional interventional distributions. We showed that factual-coordinate adversarial games identify treatment-specific conditional marginals, and introduced a stratified extended-Wasserstein objective whose population value is the causal distributional risk under overlap reweighting. Its finite-resolution implementation learns from cell-normalized conditional samples and aggregates cellwise discrepancies under the target design, avoiding density-ratio estimation. We established minimax-optimal rates, up to logarithmic factors and an additive implementation-transfer error, for finite and anisotropic continuous conditioning, using extended Wasserstein duality, WGAN perturbation inequalities, stratified finite-resolution reductions, and matching lower bounds; Besov theory characterizes the discontinuous critics.

A limitation of this work is that the theoretical guarantees for continuous conditioning rely on strong regularity assumptions, including target-design regularity, anisotropic W_1 smoothness, and a pointwise latent pushforward structure. Moreover, as in many minimax analyses of adversarial estimators, guarantees for single-network implementations without cell averaging require both the transfer condition and control of how the generated law varies within cells. The resulting rates therefore characterize approximate population or empirical minimizers, rather than the dynamics or convergence of practical nonconvex optimization. Practical issues (e.g., cell resolution, weak overlap, target-design shift, and output truncation) are discussed in Appendices F.5 and H.7.

Acknowledgments and Disclosure of Funding

We are grateful to the reviewers for their insightful comments. Shu Tamano was supported by JSPS KAKENHI Grant Number 25K24203. Masaaki Imaizumi was supported by JSPS KAKENHI (24K02904), JST CREST (JPMJCR21D2), JST FOREST (JPMJFR216I), and JST BOOST (JPMJBY24A9).

References

[Aba02] A. Abadie. Bootstrap tests for distributional treatment effects in instrumental variable models. *Journal of the American Statistical Association*, 97(457):284–292, 2002.

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR, 2017.

[ATW19] S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

[AZQ26] A. G. Abdisa, Y. Zhou, and Y. Qiu. Individualized treatment effect estimation with compromised adversarial nets. *Computational Statistics*, 41, 2026.

[BCL10] E. Brunel, F. Comte, and C. Lacour. Minimax estimation of the conditional cumulative distribution function. *Sankhya A*, 72(2):293–330, 2010.

[BFR23] B. Bilodeau, D. J. Foster, and D. M. Roy. Minimax rates for conditional density estimation via empirical entropy. *Annals of Statistics*, 51(2):762–790, 2023.

[BHOY25] U. Byambadalai, T. Hirata, T. Oka, and S. Yasui. On efficient estimation of distributional treatment effects under covariate-adaptive randomization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pages 6102–6125. PMLR, 2025.

[BJvdS20] I. Bica, J. Jordon, and M. van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. *Advances in Neural Information Processing Systems*, 33:16434–16445, 2020.

[BLR16] K. Bertin, C. Lacour, and V. Rivoirard. Adaptive pointwise estimation of conditional density function. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 52(2):939–980, 2016.

[BOY24] U. Byambadalai, T. Oka, and S. Yasui. Estimating distributional treatment effects in randomized experiments: Machine learning for variance reduction. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 5082–5113. PMLR, 2024.

[CCJM24] M. D. Cattaneo, R. Chandak, M. Jansson, and X. Ma. Boundary adaptive local polynomial conditional density estimators. *Bernoulli*, 30(4):3193–3223, 2024.

[CFVM13] V. Chernozhukov, I. Fernández-Val, and B. Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.

[Cha91] P. Chaudhuri. Nonparametric estimates of regression quantiles and their local Bahadur representation. *The Annals of Statistics*, 19(2):760–777, 1991.

[CHSW25] J. Chemseddine, P. Hagemann, G. Steidl, and C. Wald. Conditional Wasserstein distances with applications in Bayesian OT flow matching. *Journal of Machine Learning Research*, 26(141):1–47, 2025.

[DW99] R. H. Dehejia and S. Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448):1053–1062, 1999.

[DZ24] C. Dombry and A. Zaoui. Distributional regression: CRPS-error bounds for model fitting, model selection and convex aggregation. *Advances in Neural Information Processing Systems*, 37:32678–32704, 2024.

[EDCMD22] K. Elie-Dit-Cosaque and V. Maume-Deschamps. Random forest estimation of conditional distribution functions and conditional quantiles. *Electronic Journal of Statistics*, 16(2):6553–6583, 2022.

[Efr07] S. Efromovich. Conditional density estimation in a regression setting. *The Annals of Statistics*, 35(6):2504–2535, 2007.

[Efr10] S. Efromovich. Oracle inequality for conditional density estimation and an actuarial example. *Annals of the Institute of Statistical Mathematics*, 62(2):249–275, 2010.

[FFG98] J. Fan, M. Farmen, and I. Gijbels. Local maximum likelihood estimation and inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 60(3):591–608, 1998.

[Fir07] S. Firpo. Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, 75(1):259–276, 2007.

[FM25] J. Fan and H.-G. Müller. Conditional Wasserstein barycenters and interpolation/extrapolation of distributions. *IEEE Transactions on Information Theory*, 71(5):3835–3853, 2025.

[FYT96] J. Fan, Q. Yao, and H. Tong. Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika*, 83(1):189–206, 1996.

[GHF+20] Q. Ge, X. Huang, S. Fang, S. Guo, Y. Liu, W. Lin, and M. Xiong. Conditional generative adversarial networks for individualized treatment effect estimation and treatment selection. *Frontiers in Genetics*, 11:585804, 2020.

[GPAM+14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

[GR07] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

[HHT25] B. Hosseini, A. W. Hsu, and A. Taghvaei. Conditional optimal transport on function spaces. *SIAM/ASA Journal on Uncertainty Quantification*, 13(1):304–338, 2025.

[Hil11] J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

[HKB14] T. Hothorn, T. Kneib, and P. Bühlmann. Conditional transformation models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):3–27, 2014.

[HL02] M. Hoffmann and O. Lepski. Random rates in anisotropic regression (with a discussion and a rejoinder by the authors). *The Annals of Statistics*, 30(2):325–396, 2002.

[HRL04] P. Hall, J. Racine, and Q. Li. Cross-validation and the estimation of conditional probability densities. *Journal of the American Statistical Association*, 99(468):1015–1026, 2004.

[HSS24] C. Huan, R. Sun, and X. Song. Conditional generative adversarial networks for individualized causal mediation analysis. *Journal of Causal Inference*, 12(1):20220069, 2024.

[HWL+25] J. Y.-C. Hu, W. Wu, Y.-C. Lee, Y.-C. Huang, M. Chen, and H. Liu. On statistical rates of conditional diffusion transformers: Approximation, estimation and minimax optimality. In *The 13th International Conference on Learning Representations*, 2025.

[HWY99] P. Hall, R. C. Wolff, and Q. Yao. Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94(445):154–163, 1999.

[HY02] R. J. Hyndman and Q. Yao. Nonparametric estimation and symmetry tests for conditional density functions. *Journal of Nonparametric Statistics*, 14(3):259–278, 2002.

[IL17] R. Izbicki and A. B. Lee. Converting high-dimensional regression to high-dimensional conditional density estimation. *Electronic Journal of Statistics*, 11(2):2800–2831, 2017.

[JH26] K. J. Jang and G. Hwang. VE-cGAN: Improved generalization analysis of conditional generative adversarial networks using vicinal estimation. *Machine Learning*, 115, 2026.

[JL26] S. Jain and A. Luedtke. Conditional distributional treatment effects: Doubly robust estimation and testing. *arXiv preprint arXiv:2603.16829*, 2026.

[KB78] R. Koenker and G. Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978.

[KBW23] E. H. Kennedy, S. Balakrishnan, and L. A. Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

[Ken23] E. H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023.

[KLC+23] Y.-g. Kim, K. Lee, Y. Choi, J.-H. Won, and M. C. Paik. Wasserstein geodesic generator for conditional distributions. *arXiv preprint arXiv:2308.10145*, 2023.

[KLP01] G. Kerkyacharian, O. Lepski, and D. Picard. Nonlinear estimation in anisotropic multi-index denoising. *Probability Theory and Related Fields*, 121(2):137–170, 2001.

[KLP23] Y.-g. Kim, K. Lee, and M. C. Paik. Conditional Wasserstein generator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7208–7219, 2023.

[KMS24] G. Kerrigan, G. Migliorini, and P. Smyth. Dynamic conditional optimal transport through simulation-free flows. *Advances in Neural Information Processing Systems*, 37:93602–93642, 2024.

[KO23] N. Kallus and M. Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 6037–6060. PMLR, 2023.

[KYL24] S. Kumar, Y. Yang, and L. Lin. A likelihood based approach to distribution regression using conditional deep generative models. *arXiv preprint arXiv:2410.02025*, 2024.

[LaL86] R. J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review*, 76(4):604–620, 1986.

[LF25] A. Luedtke and K. Fukumizu. DoubleGen: Debiased generative modeling of counterfactuals. *arXiv preprint arXiv:2509.16842*, 2025.

[LKL+20] Y. Li, K. Kuang, B. Li, P. Cui, J. Tao, H. Yang, and F. Wu. Continuous treatment effect estimation via generative adversarial de-confounding. In *Proceedings of the 2020 KDD Workshop on Causal Discovery*, volume 127, pages 4–22. PMLR, 2020.

[LNB22] M. Li, M. Neykov, and S. Balakrishnan. Minimax optimal conditional density estimation under total variation smoothness. *Electronic Journal of Statistics*, 16(2):3937–3972, 2022.

[LR08] Q. Li and J. S. Racine. Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data. *Journal of Business & Economic Statistics*, 26(4):423–434, 2008.

[Mar21] J. Martin. About exchanging expectation and supremum for conditional Wasserstein GANs. *arXiv preprint arXiv:2103.13906*, 2021.

[Mei06] N. Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006.

[MFF23] V. Melnychuk, D. Frauen, and S. Feuerriegel. Normalizing flows for interventional density estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pages 24361–24397. PMLR, 2023.

[MMSF24] Y. Ma, V. Melnychuk, J. Schweisthal, and S. Feuerriegel. DiffPO: A causal diffusion model for learning distributions of potential outcomes. *Advances in Neural Information Processing Systems*, 37:43663–43692, 2024.

[MO14] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[NEBM23] J. Näf, C. Emmenegger, P. Bühlmann, and N. Meinshausen. Confidence and uncertainty assessment for distributional random forests. *Journal of Machine Learning Research*, 24(366):1–77, 2023.

[Ney23] J. Neyman. Sur les applications de la théorie des probabilités aux expériences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10(1):1–51, 1923.

[NI25] Y. Norimatsu and M. Imaizumi. Encode-Decoder-based GAN for estimating counterfactual outcomes under sequential selection bias and combinatorial explosion. In *Proceedings of the Fourth Conference on Causal Learning and Reasoning*, volume 275, pages 451–489. PMLR, 2025.

[Nob96] A. Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.

[NP14] A. Norets and J. Pelenis. Posterior consistency in conditional density estimation by covariate dependent mixtures. *Econometric Theory*, 30(3):606–646, 2014.

[NP17] A. Norets and D. Pati. Adaptive Bayesian estimation of conditional densities. *Econometric Theory*, 33(4):980–1012, 2017.

[NPS26] J. Näf, J. Park, and H. Susmann. Causal-DRF: Conditional kernel treatment effect estimation using distributional random forest. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300. PMLR, 2026.

[NvS97] M. H. Neumann and R. von Sachs. Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *The Annals of Statistics*, 25(1):38–76, 1997.

[NYLN21] L. Nie, M. Ye, Q. Liu, and D. Nicolae. VCNet and functional targeted regularization for learning causal effects of continuous treatments. In *The 9th International Conference on Learning Representations*, 2021.

[OYHB26] T. Oka, S. Yasui, Y. Hayakawa, and U. Byambadalai. Regression adjustment for estimating distributional treatment effects in randomized controlled trials. *Econometric Reviews*, 45(1):2–17, 2026.

[PDT13] D. Pati, D. B. Dunson, and S. T. Tokdar. Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, 116:456–472, 2013.

[PSSM21] J. Park, U. Shalit, B. Schölkopf, and K. Muandet. Conditional distributional treatment effect with kernel conditional mean embeddings and U-statistic regression. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8401–8412. PMLR, 2021.

[Rot10] C. Rothe. Nonparametric estimation of distributional policy effects. *Journal of Econometrics*, 155(1):56–70, 2010.

[Rot12] C. Rothe. Partial distributional policy effects. *Econometrica*, 80(5):2269–2301, 2012.

[Rub74] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

[RZLL16] Y. Ren, J. Zhu, J. Li, and Y. Luo. Conditional generative moment-matching networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[SAL24] A. Stéphanovitch, E. Aamari, and C. Levrard. Wasserstein Generative Adversarial Networks are Minimax Optimal Distribution Estimators. *The Annals of Statistics*, 52(5):2167–2193, 2024.

[Sar17] M. Sart. Estimating the conditional density by histogram type estimators and model selection. *ESAIM: Probability and Statistics*, 21:34–55, 2017.

[SG16] W. Shen and S. Ghosal. Adaptive Bayesian density regression for high-dimensional data. *Bernoulli*, 22(1):396–420, 2016.

[SJS17] U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3076–3085. PMLR, 2017.

[SLB+20] P. Schwab, L. Linhardt, S. Bauer, J. M. Buhmann, and W. Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5612–5619, 2020.

[SN21] T. Suzuki and A. Nitanda. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic Besov space. *Advances in Neural Information Processing Systems*, 34:3609–3621, 2021.

[SR13] G. J. Székely and M. L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*, 143(8):1249–1272, 2013.

[STS+10] M. Sugiyama, I. Takeuchi, T. Suzuki, T. Kanamori, H. Hachiya, and D. Okanohara. Least-squares conditional density estimation. *IEICE Transactions on Information and Systems*, 93(3):583–594, 2010.

[Suz19] T. Suzuki. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality. In *The 7th International Conference on Learning Representations*, 2019.

[SWS+26] S. Song, T. Wang, G. Shen, Y. Lin, and J. Huang. Wasserstein generative regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 88(1):330–351, 2026.

[TLY25] R. Tang, L. Lin, and Y. Yang. Conditional diffusion models are minimax-optimal and manifold-adaptive for conditional distribution estimation. In *The 13th International Conference on Learning Representations*, 2025.

[Tri83] H. Triebel. *Theory of Function Spaces*. Birkhäuser, 1983.

[TTZ20] E. G. Tabak, G. Trigila, and W. Zhao. Conditional density estimation and simulation through optimal transport. *Machine Learning*, 109(4):665–688, 2020.

[TTZ21] E. G. Tabak, G. Trigila, and W. Zhao. Data driven conditional optimal transport. *Machine Learning*, 110(11):3135–3155, 2021.

[TY23] R. Tang and Y. Yang. Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282–1308, 2023.

[WCM+13] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, 2013.

[WIX25] D. Wu, D. I. Inouye, and Y. Xie. Flow-based generative modeling of potential outcomes and counterfactuals. *arXiv preprint arXiv:2505.16051*, 2025.

[WLZ+26] Y. Wang, H. Li, M. Zhu, A. Wu, B. Li, K. Yin, R. Xiong, F. Wu, and K. Kuang. Causal inference with complex treatments: A survey. *ACM Computing Surveys*, 58(9):1–36, 2026.

[WQX26] D. Wu, F. Qiu, and Y. Xie. DoFlow: Flow-based generative models for interventional and counterfactual forecasting on time series. In *The 14th International Conference on Learning Representations*, 2026.

[WSL+23] G. Wu, G. Song, X. Lv, S. Luo, C. Shi, and H. Zhu. DNet: Distributional network for distributional individualized treatment effects. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5215–5224, 2023.

[WZCZ24] S. Wu, W. Zhou, M. Chen, and S. Zhu. Counterfactual generative models for time-varying treatments. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3402–3413, 2024.

[YJvdS18a] J. Yoon, J. Jordon, and M. van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5689–5698. PMLR, 2018.

[YJvdS18b] J. Yoon, J. Jordon, and M. van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *The 6th International Conference on Learning Representations*, 2018.

[ZLZ+26] J. Zhang, Y. Lin, R. Zhang, X. Song, and H. Ning. GAHMN: A generative approach for high-dimensional mediation analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 28247–28255, 2026.

A Detailed Related Work

Adversarial and implicit generative models. The adversarial formulation of generative modeling was introduced by [GPAM$^+$14]. Conditional GANs [MO14] and conditional moment-matching networks [RZLL16] extend implicit generation to settings with covariates, labels, or side information, while WGANs replace the Jensen–Shannon objective with a Wasserstein objective with a metric geometry that remains informative for singular distributions [ACB17]. This feature is important for causal generative modeling: outcome distributions may be low-dimensional, non-Gaussian, or supported on structured sets, and density-based divergences may be poorly aligned with the target distributional discrepancy. Our work uses Wasserstein geometry as the statistical risk for the identifiable conditional interventional distribution, and not merely as a training heuristic.

Adversarial causal inference. GANITE [YJvdS18b] introduced adversarial counterfactual imputation for individualized treatment-effect estimation with discrete treatments, building on ideas from adversarial missing-data imputation [YJvdS18a]. SCIGAN [BJvdS20] extended this architecture to continuous-valued interventions through hierarchical discriminators over treatments and dosages. Subsequent adversarial causal methods include conditional GANs for treatment selection [GHF$^+$20], generative adversarial deconfounding for continuous treatments [LKL$^+$20], time-varying and sequential treatment generators [WZCZ24, NI25], compromised adversarial nets for individualized effects [AZQ26], and mediation-oriented adversarial generators [HSS24, ZLZ$^+$26]. Recent flow-based methods such as PO-Flow and DoFlow further illustrate the shift toward generative causal targets [WIX25, WQX26], and [LF25] study debiased generative counterfactual modeling. A complementary survey perspective on complex treatments is given by [WLZ$^+$26]. Instead of proposing another counterfactual imputation architecture, we analyze the population objective of factual-coordinate adversarial games, show that they identify conditional interventional marginals rather than a joint potential-outcome distribution, and replace the factual-coordinate discriminator with an extended Wasserstein objective for the identifiable causal target.

Counterfactual distributions and distributional treatment effects. Classical work on distributional causal parameters includes counterfactual distribution inference [CFVM13], quantile treatment-effect estimation [Fir07], bootstrap tests for distributional effects [Aba02], and distributional policy effects [Rot10, Rot12]. Recent work studies semiparametric counterfactual density estimation [KBW23], conditional distributional treatment effects through kernel embeddings and robust procedures [PSSM21, KO23], distributional individualized treatment effects [WSL$^+$23], distributional random forests for causal targets [NPS26], randomized-experiment distributional effects with machine-learning variance reduction [BOY24, BHOY25, OYHB26], and doubly robust conditional distributional effect estimation [JL26]. These works typically target distribution functions, densities, quantiles, distributional functionals, or kernel mean embeddings. We instead target an implicit generative estimate of the full conditional interventional distribution under extended Wasserstein risk.

Conditional distribution and density estimation. Conditional CDF and quantile estimation have been studied using kernel, local-polynomial, transformation, and random-forest methods [KB78, Cha91, HWY99, Mei06, LR08, BCL10, HKB14, ATW19, EDCMD22, NEBM23]. Conditional density estimation has a large nonparametric literature, including local likelihood and kernel estimators [FYT96, FFG98, HY02, HRL04], oracle and minimax theory [Efr07, Efr10, BLR16, Sar17, LNB22, BFR23, CCJM24], least-squares and high-dimensional reductions [STS$^+$10, IL17], Bayesian conditional density models [PDT13, NP14, NP17, SG16], and likelihood-based conditional deep generative models [KYL24]. These approaches are foundational for conditional distribution learning, but many of them rely on distributions with densities or distribution-function representations. The implicit generative formulation studied here permits conditional outcome distributions that may be singular with respect to Lebesgue measure, while the causal target is handled through overlap-based reweighting.

Conditional Wasserstein distances and conditional optimal transport. Several Wasserstein-type constructions have been proposed for conditional distributions. Data-driven conditional optimal transport estimates transport maps between conditional distributions indexed by covariates [TTZ20, TTZ21]. Conditional Wasserstein barycenters use Wasserstein geometry for interpolation, extrapolation, and distribution-on-predictor regression [FM25]. Conditional optimal transport

on function spaces studies constrained triangular transport maps and Kantorovich relaxations for conditional measures, with applications to amortized Bayesian inference [HHT25]. Dynamic conditional optimal transport uses conditional Wasserstein geometry to construct simulation-free conditional flows [KMS24]. In conditional generative modeling, conditional Wasserstein generators and Wasserstein geodesic generators use optimal-transport geometry to learn conditional distributions and geodesic interpolation between observed domains [KLP23, KLC$^+$23]; [Mar21] study the exchange between expectation and supremum in conditional WGAN objectives.

Our use of extended Wasserstein distance follows [CHSW25] for a specific reason. Their distance is defined through couplings that preserve the conditioning coordinate, equals the target-design average of pointwise Wasserstein distances between conditional distributions, and admits a conditional Kantorovich–Rubinstein dual whose critic is Lipschitz in the outcome coordinate conditional on the state. This dual is precisely the form needed for a conditional WGAN objective. This choice is essential in causal inference: a joint Wasserstein distance on the treatment–covariate–outcome space can reduce discrepancy by moving treatment–covariate states, whereas the causal estimand compares outcome distributions at the same treatment–covariate state. The extended Wasserstein distance therefore matches the identifiable causal risk, while still yielding an adversarial training objective.

Deep generative distribution estimation theory. For unconditional generative distribution estimation, [TY23] derive minimax rates under adversarial losses on unknown submanifolds, and [SAL24] prove minimax optimality of WGAN estimators using a neural generator sieve and an optimal-dual-on-a-net critic construction. Conditional diffusion models have recently been analyzed by [TLY25] and [HWL$^+$25]; their rates quantify the statistical cost of conditioning under smooth density or score assumptions. Wasserstein generative regression [SWS$^+$26], Wasserstein geodesic generators [KLC$^+$23], and vicinal analyses of conditional GANs [JH26] provide related conditional generative perspectives. Our theory differs in four aspects: the target is causal, the loss is extended Wasserstein risk, observational sampling is handled by stratified cell normalization under overlap, and the local generative difficulty is inherited from minimax-optimal WGAN estimation under a latent pushforward model.

Anisotropic Besov structures and non-smooth critics. Anisotropic Besov spaces are classical tools for describing functions whose regularity differs across coordinates. They have been used in wavelet thresholding and adaptive nonparametric estimation over anisotropic smoothness classes [NvS97, KLP01, HL02], and their approximation-theoretic foundations are developed in the function-space literature [Tri83]. In deep learning theory, [Suz19] shows adaptivity of deep ReLU networks over Besov and mixed-smooth Besov spaces, while [SN21] develop approximation, estimation, and minimax analyses for anisotropic Besov spaces and intrinsic smoothness structures. These works use Besov regularity primarily to characterize target regression functions or neural approximation classes.

Our use of Besov structure is different. The Besov object is not the conditional mean or density, but the critic map from treatment–covariate states to outcome-Lipschitz Kantorovich potentials. This distinction matters because extended Wasserstein critics must remain Lipschitz in the outcome coordinate while being allowed to change sharply across treatment–covariate regimes. Embedding finite-resolution step critics into anisotropic Besov balls below the continuity threshold gives a norm bound of $C \max_j 2^{m_j s_j}$ for their partition-boundary jumps. The stochastic risk bounds follow from the restricted one-state critic classes and stability under cell averaging.

B Notation

This appendix summarizes the notation used throughout the paper in Tables 3 and 4. All probability measures are defined on standard Borel spaces, and conditional distributions are fixed regular versions. Equalities involving conditional distributions are interpreted up to the relevant conditioning marginal. Write $\Gamma(\eta, \zeta)$ for the set of couplings with marginals η, ζ. We use $\|f\|_\infty = \sup_x \|f(x)\|$ for the uniform norm, e_j for the jth coordinate vector in the relevant Euclidean space. We also write $\mathbf{1}_A$ for the indicator of a measurable set A.

Table 3: Summary of basic causal and generative notation

Symbol	Description
$\mathcal{P}(S)$	Borel probability measures on a standard Borel space S.
$g_{\#}\xi$	Pushforward of a probability law ξ by a measurable map g; $g_{\#}U$ abbreviates $g_{\#}\mathrm{Unif}([0,1]^{d_U})$.
W_1	Ordinary Wasserstein-1 distance on the outcome space $\mathcal{Y}$; $W_1^{\otimes}$ denotes the product-space version where applicable.
$\mathrm{Lip}(f)$	Optimal Lipschitz constant of f with respect to the relevant metric.
$\mathcal{X}, \mathcal{T}, \mathcal{Y}$	Covariate, treatment, and outcome spaces; in the theoretical analysis, $\mathcal{Y} = B^p(0, 2K_0)$ is a closed Euclidean ball and true generator images lie in $B^p(0, K_0)$.
$X, T, Y(t)$	Covariates, treatment, and potential outcome under treatment t.
Y^{obs}	Observed outcome, equal to $Y(T)$ under consistency.
$W = (X, T)$	Treatment–covariate state used as the conditioning variable.
$\mathcal{W}$	Conditioning space; either finite $\{1, \dots, M\}$ or continuous $[0,1]^{d_W}$.
n	Total observational sample size; $(W_i, Y_i^{\mathrm{obs}})$, $i = 1, \dots, n$, are i.i.d. observations from P_{obs}.
Q_{obs}, Q_ρ	Observational and target design marginals of W.
w_ρ	Density ratio $\mathrm{d}Q_\rho/\mathrm{d}Q_{\mathrm{obs}}$; overlap assumes $0 \le w_\rho \le \kappa^{-1}$.
κ	Overlap constant in $(0, 1]$; smaller κ means weaker overlap.
μ_w^*	Conditional interventional outcome law at state $w = (x,t)$: $\mu^*_{(x,t)}(A) = \mathbb{P}(Y(t) \in A \mid X = x)$ for Borel $A \subset \mathcal{Y}$.
P_{obs}	Observational law $P_{\mathrm{obs}}(\mathrm{d}w, \mathrm{d}y) = Q_{\mathrm{obs}}(\mathrm{d}w)\mu_w^*(\mathrm{d}y)$ under consistency and ignorability.
P^*	Target interventional law $P^*(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\mu_w^*(\mathrm{d}y)$.
U, d_U	Exogenous noise $U \sim \mathrm{Unif}([0,1]^{d_U})$ and its latent dimension d_U.
g	Conditional generator $g : \mathcal{W} \times [0,1]^{d_U} \to \mathcal{Y}$.
$\nu_{g,w}$	Generated conditional law $g(w, \cdot)_{\#}U$; generic ν_w or ν denotes a comparison or generated probability law.
P_g	Raw generated law $P_g(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_{g,w}(\mathrm{d}y)$.

B.1 Dyadic Partitions

For $\boldsymbol{m} = (m_1, \dots, m_{d_W}) \in \mathbb{N}_0^{d_W}$, the anisotropic dyadic partition $\Pi_{\boldsymbol{m}}$ consists of rectangles

$$\prod_{j=1}^{d_W} I_{k_j, m_j}, \quad k_j \in \{0, \dots, 2^{m_j} - 1\},$$

where $I_{k,m} = [k2^{-m}, (k+1)2^{-m})$ for $k < 2^m - 1$ and $I_{2^m-1,m} = [(2^m - 1)2^{-m}, 1]$. Thus, each cell has side length 2^{-m_j} in coordinate j and Lebesgue volume $2^{-|\boldsymbol{m}|_1}$. Under Assumption 4, $q_C = Q_\rho(C) \asymp 2^{-|\boldsymbol{m}|_1}$ uniformly over $C \in \Pi_{\boldsymbol{m}}$.

C Proofs of Section 2

Let K_x be a regular conditional law of V given $X = x$. For Borel $A \subset \mathcal{Y}$, write

$$\mu^*_{(x,t)}(A) = \int \mathbf{1}\{F(x,t,v) \in A\} K_x(\mathrm{d}v).$$

Lemma 1 (Identification of the conditional outcome kernel)**.** *Under Assumptions 1–2, $w \mapsto \mu_w^*$ is a Borel probability kernel, and for every bounded measurable $\varphi : \mathcal{W} \times \mathcal{Y} \to \mathbb{R}$,*

$$\mathbb{E}[\varphi(W, Y^{\mathrm{obs}})] = \int\int \varphi(w,y)\mu_w^*(\mathrm{d}y)Q_{\mathrm{obs}}(\mathrm{d}w).$$

Proof of Lemma 1. Conditional independence gives $\mathcal{L}(V \mid X, T) = K_X$ almost surely. Joint measurability of F and consistency yield

$$\mathbb{E}[\varphi(X, T, F(X, T, V)) \mid X, T] = \int \varphi(X, T, F(X, T, v)) K_X(\mathrm{d}v) = \int \varphi(W, y) \mu^*_W(\mathrm{d}y),$$

where the second equality is the definition of $\mu^*_{(x,t)}$ as the pushforward of K_x by $F(x, t, \cdot)$. Taking expectations proves the identity. Measurability of parameterized kernel integrals gives the Borel kernel property. For each fixed t, the same definition is a version of $\mathcal{L}(Y(t) \mid X = x)$. □

This section formalizes the claim that factual-coordinate adversarial imputation identifies conditional marginals but not a joint potential-outcome law. We state the result for finite treatments $\mathcal{T} = \{1, \ldots, A\}$ and standard Borel spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\bar{\boldsymbol{Y}} = (\bar{Y}(1), \ldots, \bar{Y}(A)) \in \mathcal{Y}^A$ be the completed outcome vector produced by an imputation generator after inserting the factual outcome. All conditional probabilities are fixed regular versions, and $0 \log 0 = 0$.

Write G for the imputation generator that determines the completed vector $\bar{\boldsymbol{Y}}$, and $d = (d_1, \ldots, d_A)$ for a discriminator with coordinate functions $d_a : \mathcal{X} \times \mathcal{Y}^A \to [0, 1]$. For a fixed G, define

$$r_a(x, \bar{y}) = \mathbb{P}(T = a \mid X = x, \bar{\boldsymbol{Y}} = \bar{y}), \quad \ell(u) = u \log u + (1 - u) \log(1 - u), \quad u \in [0, 1].$$

The discriminator objective is

$$V(G, d) \coloneqq \mathbb{E}\left[\sum_{a=1}^{A} \{\mathbf{1}\{T = a\} \log d_a(X, \bar{\boldsymbol{Y}}) + \mathbf{1}\{T \neq a\} \log(1 - d_a(X, \bar{\boldsymbol{Y}}))\}\right].$$

Proposition 1 (Population factual-coordinate discriminator)**.** *For every fixed G, maximization over measurable discriminators d has the following properties:*

(i) the pointwise maximizer is $d^*_a(x, \bar{y}) = r_a(x, \bar{y})$;

(ii) the maximized value is $\sup_d V(G, d) = \mathbb{E}[\sum_{a=1}^{A} \ell(r_a(X, \bar{\boldsymbol{Y}}))]$;

(iii)

$$\sup_d V(G, d) \geq \mathbb{E}\left[\sum_{a=1}^{A} \ell(\mathbb{P}(T = a \mid X))\right], \tag{9}$$

with equality if and only if $T \perp\!\!\!\perp \bar{\boldsymbol{Y}} \mid X$, up to null sets.

Proof of Proposition 1. Fix a and condition on $(X, \bar{\boldsymbol{Y}}) = (x, \bar{y})$. The conditional contribution is $r_a(x, \bar{y}) \log q + (1 - r_a(x, \bar{y})) \log(1 - q)$ with $q = d_a(x, \bar{y})$. This Bernoulli log-likelihood is concave in q and has derivative $r_a/q - (1 - r_a)/(1 - q)$. The unique interior maximizer is $q = r_a$ when $r_a \in (0, 1)$; when $r_a \in \{0, 1\}$, the same value is attained in the closure under the convention $0 \log 0 = 0$. This proves (i) and (ii).

The function ℓ is convex on $[0, 1]$ and strictly convex on $(0, 1)$. Jensen's inequality conditional on X gives

$$\mathbb{E}[\ell(r_a(X, \bar{\boldsymbol{Y}})) \mid X] \geq \ell(\mathbb{E}[r_a(X, \bar{\boldsymbol{Y}}) \mid X]).$$

By the tower property, $\mathbb{E}[r_a(X, \bar{\boldsymbol{Y}}) \mid X] = \mathbb{P}(T = a \mid X)$. Summing over a and taking expectations proves (9). Equality holds exactly when $r_a(X, \bar{\boldsymbol{Y}}) = \mathbb{P}(T = a \mid X)$ almost surely for every a on nondegenerate conditional supports, which is equivalent to $T \perp\!\!\!\perp \bar{\boldsymbol{Y}} \mid X$. □

Let $p_a(x) = \mathbb{P}(T = a \mid X = x)$ be a fixed regular version of the propensity for treatment a.

Theorem 4 (Identified object of factual-coordinate adversarial imputation)**.** *Assume Assumptions 1 and 2. Suppose that a completed vector $\bar{\boldsymbol{Y}}$ satisfies exact factual reconstruction, $\bar{Y}(T) = Y^{\mathrm{obs}}$ almost surely, and population invariance, $T \perp\!\!\!\perp \bar{\boldsymbol{Y}} \mid X$. Then, for every treatment a,*

$$\mathcal{L}(\bar{Y}(a) \mid X = x) = \mathcal{L}(Y(a) \mid X = x) \tag{10}$$

for P_X-almost every x such that $p_a(x) > 0$. Conversely, let $x \mapsto \kappa_x$ be any measurable family of couplings of the marginals $\{\mathcal{L}(Y(a) \mid X = x) : a = 1, \ldots, A\}$. There exists a randomized imputation generator satisfying exact factual reconstruction and population invariance such that $\mathcal{L}(\bar{\boldsymbol{Y}} \mid X = x) = \kappa_x$ for P_X-almost every x. Therefore, factual-coordinate adversarial objectives identify treatment-specific conditional marginals but not the conditional joint law of all potential outcomes.

Proof of Theorem 4. We first prove marginal identification. Fix a and let x be a point at which the relevant regular conditional distributions are defined and $p_a(x) > 0$. By $T \perp\!\!\!\perp \bar{\boldsymbol{Y}} \mid X$, the conditional law of $\bar{\boldsymbol{Y}}$ given $(X = x, T = t)$ is independent of t over the support of $T \mid X = x$, and in particular coincides with $\mathcal{L}(\bar{\boldsymbol{Y}} \mid X = x)$ at $t = a$. Projecting onto coordinate a, using exact factual reconstruction $\bar{Y}(a) = Y^{\mathrm{obs}}$ on $\{T = a\}$, and then applying consistency ($Y^{\mathrm{obs}} = Y(a)$ on $\{T = a\}$) and ignorability ($Y(a) \perp\!\!\!\perp T \mid X$) yields

$$\begin{aligned}\mathcal{L}(\bar{Y}(a) \mid X = x) &= \mathcal{L}(\bar{Y}(a) \mid X = x, T = a) \\ &= \mathcal{L}(Y^{\mathrm{obs}} \mid X = x, T = a) \\ &= \mathcal{L}(Y(a) \mid X = x, T = a) \\ &= \mathcal{L}(Y(a) \mid X = x),\end{aligned}$$

which proves (10).

For the converse, write y_{-a} for the outcome vector with coordinate a omitted, and disintegrate $P_X(\mathrm{d}x)\kappa_x(\mathrm{d}\boldsymbol{y})$ with respect to (x, y_a), separately for each of the finitely many treatments a. Since the ath marginal of κ_x is $\mathcal{L}(Y(a) \mid X = x)$, standard Borel disintegration supplies a jointly measurable kernel $K_{x,a}(\mathrm{d}y_{-a} \mid y_a)$ such that

$$\kappa_x(\mathrm{d}y_a, \mathrm{d}y_{-a}) = \mathcal{L}(Y(a) \mid X = x)(\mathrm{d}y_a)\, K_{x,a}(\mathrm{d}y_{-a} \mid y_a)$$

for P_X-almost every x. By the randomization lemma for kernels on standard Borel spaces, there exist an auxiliary $Z \sim \mathrm{Unif}([0, 1])$, independent of (X, T, Y^{obs}), and a measurable map R such that $R(x, a, y, Z)$ has law $K_{x,a}(\cdot \mid y)$. Given $(X, T, Y^{\mathrm{obs}}, Z) = (x, a, y, z)$, define $\bar{Y}(a) = y$ and $\bar{\boldsymbol{Y}}_{-a} = R(x, a, y, z)$. Exact factual reconstruction is immediate. Conditional on $(X = x, T = a)$, the factual outcome has law $\mathcal{L}(Y(a) \mid X = x)$ by consistency and ignorability; sampling the remaining coordinates from $K_{x,a}$ reconstructs κ_x. Therefore $\mathcal{L}(\bar{\boldsymbol{Y}} \mid X = x, T = a) = \kappa_x$ for each treatment in the support of $T \mid X = x$, which is equivalent to $T \perp\!\!\!\perp \bar{\boldsymbol{Y}} \mid X$. □

D Proofs of Section 3

D.1 Technical Conventions and Extended Wasserstein Duality

Throughout this section, $\mathcal{Y}$ is compact and metrized by the Euclidean distance inherited from $\mathbb{R}^p$, and $\mathcal{W}$ is standard Borel. Every probability measure on $\mathcal{W} \times \mathcal{Y}$ admits a regular conditional distribution given W. For $P(\mathrm{d}w, \mathrm{d}y) = Q(\mathrm{d}w)\mu_w(\mathrm{d}y)$ and $R(\mathrm{d}w, \mathrm{d}y) = Q(\mathrm{d}w)\nu_w(\mathrm{d}y)$, the kernels $w \mapsto \mu_w$ and $w \mapsto \nu_w$ are fixed versions, and $\Delta(w) = (w, w)$.

Lemma 2 (Disintegration of diagonal couplings)**.** *Let P and R share the conditioning marginal Q. A coupling α of P and R belongs to $\Gamma_\Delta(P, R)$ if and only if there exists a Q-measurable kernel $w \mapsto \gamma_w$ from $\mathcal{W}$ to $\mathcal{Y} \times \mathcal{Y}$ such that $\gamma_w \in \Gamma(\mu_w, \nu_w)$ for Q-almost every w and*

$$\alpha(\mathrm{d}w, \mathrm{d}y, \mathrm{d}w', \mathrm{d}y') = Q(\mathrm{d}w)\delta_w(\mathrm{d}w')\gamma_w(\mathrm{d}y, \mathrm{d}y'). \tag{11}$$

Consequently,

$$\inf_{\alpha \in \Gamma_\Delta(P,R)} \int \|y - y'\| \mathrm{d}\alpha = \int_{\mathcal{W}} W_1(\mu_w, \nu_w) Q(\mathrm{d}w). \tag{12}$$

Proof of Lemma 2. Let $\alpha \in \Gamma_\Delta(P, R)$. Since its (W, W')-marginal is $\Delta_\# Q$, α is concentrated on $\{w = w'\}$. Disintegrating α with respect to the first w coordinate gives $\alpha(\mathrm{d}w, \mathrm{d}y, \mathrm{d}w', \mathrm{d}y') = Q(\mathrm{d}w)K_w(\mathrm{d}y, \mathrm{d}w', \mathrm{d}y')$. The diagonal constraint implies that K_w is concentrated on $\mathcal{Y} \times \{w\} \times \mathcal{Y}$ for Q-almost every w, so there is a kernel γ_w satisfying (11). The marginal constraints give $\gamma_w \in \Gamma(\mu_w, \nu_w)$ almost surely.

Conversely, any representation (11) with $\gamma_w \in \Gamma(\mu_w, \nu_w)$ has marginals P and R and diagonal (W, W')-marginal. The value in (12) therefore reduces to an infimum of $\int\int \|y - y'\| \gamma_w(\mathrm{d}y, \mathrm{d}y')Q(\mathrm{d}w)$ over measurable selections $w \mapsto \gamma_w \in \Gamma(\mu_w, \nu_w)$. Every selection has cost at least $\int W_1(\mu_w, \nu_w)Q(\mathrm{d}w)$, because the inner integral is at least $W_1(\mu_w, \nu_w)$ for each w. For the reverse inequality, compactness of $\mathcal{Y}$ and continuity of the cost imply that the set of optimal couplings of (μ_w, ν_w) is nonempty and compact for each w, and $w \mapsto (\mu_w, \nu_w)$ is measurable; the measurable selection theorem for compact-valued measurable correspondences then yields a measurable kernel $w \mapsto \gamma_w^*$ of optimal couplings, whose cost equals $\int W_1(\mu_w, \nu_w)Q(\mathrm{d}w)$. □

Lemma 3 (Extended Kantorovich–Rubinstein duality)**.** *Let P and R share the marginal Q on $\mathcal{W}$. Then*

$$\mathrm{eW}_1(P,R) = \sup_{h\in\mathcal{F}}\{\mathbb{E}_P[h(W,Y)] - \mathbb{E}_R[h(W,Y)]\}.$$

Moreover, for fixed Borel kernels $w \mapsto \mu_w$ and $w \mapsto \nu_w$, the supremum is attained.

Proof of Lemma 3. For every $h \in \mathcal{F}$, ordinary Kantorovich–Rubinstein duality on compact $\mathcal{Y}$ gives

$$\int h(w,y)\mu_w(\mathrm{d}y) - \int h(w,y)\nu_w(\mathrm{d}y) \le W_1(\mu_w,\nu_w)$$

for Q-almost every w. Integrating and using Lemma 2 gives the upper bound.

Let $C(\mathcal{Y})$ denote the space of continuous real-valued functions on $\mathcal{Y}$, equipped with the uniform norm. For the reverse direction, $\mathcal{L}_1$ is compact in $C(\mathcal{Y})$ by Arzelà–Ascoli. Let $\Phi(w,f) = \int f\mathrm{d}\mu_w - \int f\mathrm{d}\nu_w$. The map $w \mapsto \Phi(w,f)$ is measurable for each f, and $f \mapsto \Phi(w,f)$ is continuous for each w. The measurable maximum theorem yields a measurable selector $f_w^* \in \mathcal{L}_1$ maximizing $\Phi(w,\cdot)$. Ordinary duality gives $\Phi(w,f_w^*) = W_1(\mu_w,\nu_w)$. Since evaluation is continuous on $C(\mathcal{Y}) \times \mathcal{Y}$, $h^*(w,y) = f_w^*(y)$ is jointly measurable and belongs to $\mathcal{F}$. Integrating $\Phi(w,f_w^*) = W_1(\mu_w,\nu_w)$ over Q gives $\mathbb{E}_P[h^*] - \mathbb{E}_R[h^*] = \int W_1(\mu_w,\nu_w)Q(\mathrm{d}w) = \mathrm{eW}_1(P,R)$ by Lemma 2, which proves equality and attainment. □

For a finite metric conditioning space with at least two points, denote its minimum separation by $\delta_{\mathcal{W}} = \min_{w\neq w'} d_{\mathcal{W}}(w,w')$.

Proposition 2 (Comparison on a finite conditioning space)**.** *Suppose $(\mathcal{W}, d_{\mathcal{W}})$ is a finite metric space with at least two points. For any P, R with the same conditioning marginal,*

$$W_1^{\otimes}(P,R) \le \mathrm{eW}_1(P,R) \le \max\{1, \mathrm{diam}(\mathcal{Y})/\delta_{\mathcal{W}}\}\, W_1^{\otimes}(P,R).$$

Proof of Proposition 2. The first inequality follows by restricting the admissible couplings to diagonal couplings. For the second, fix any coupling γ of P and R and let $m = \gamma\{W \neq W'\}$. Retain the restriction of γ to $\{W = W'\}$. For each state $w \in \mathcal{W}$, since P and R share the state marginal Q, we have

$$\begin{aligned}\gamma(\{W = w, W' \neq w\}) &= Q(\{w\}) - \gamma(\{W = w, W' = w\}) \\ &= \gamma(\{W \neq w, W' = w\}) =: m_w,\end{aligned}$$

with $\sum_{w\in\mathcal{W}} m_w = \gamma\{W \neq W'\} = m$. Thus, at each state w, the remaining first and second outcome marginals have equal total mass m_w; when $m_w > 0$, we couple these remaining measures within state w (for instance, via their normalized product measure). The resulting diagonal coupling has outcome cost at most

$$\int_{\{W=W'\}} \|y - y'\|\mathrm{d}\gamma + \mathrm{diam}(\mathcal{Y})\, m.$$

The original product transport cost is at least

$$\int_{\{W=W'\}} \|y - y'\|\mathrm{d}\gamma + \delta_{\mathcal{W}} m.$$

Hence

$$\begin{aligned}\mathrm{eW}_1(P,R) &\le \int_{\{W=W'\}} \|y - y'\|\mathrm{d}\gamma + \mathrm{diam}(\mathcal{Y})\, m \\ &\le \max\{1, \mathrm{diam}(\mathcal{Y})/\delta_{\mathcal{W}}\}\Big(\int_{\{W=W'\}} \|y - y'\|\mathrm{d}\gamma + \delta_{\mathcal{W}} m\Big),\end{aligned}$$

and taking the infimum over γ proves the claim. □

D.2 Population Identity

Proof of (3). Let g be measurable and let $P_g(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_{g,w}(\mathrm{d}y)$. For every bounded measurable $h \in \mathcal{F}$, the identity in the main text

$$\mathbb{E}_{P_{\mathrm{obs}}}[w_\rho(W)\varphi(W, Y^{\mathrm{obs}})] = \mathbb{E}_{P^*}[\varphi(W, Y)]$$

gives

$$\mathbb{E}_{P_{\mathrm{obs}}}[w_\rho(W)h(W, Y^{\mathrm{obs}})] = \mathbb{E}_{P^*}[h(W, Y)].$$

Since $\tilde{W} \sim Q_\rho$ independently of U,

$$\mathbb{E}[h(\tilde{W}, g(\tilde{W}, U))] = \mathbb{E}_{P_g}[h(W, Y)].$$

Hence $L(g, h) = \mathbb{E}_{P^*}[h] - \mathbb{E}_{P_g}[h]$. The laws P^* and P_g share the marginal Q_ρ, so Lemma 3 proves the identity. □

E Proofs of Section 4

E.1 Imported One-State WGAN Primitive

Let $U \sim \mathrm{Unif}([0,1]^{d_U})$ and set

$$a \coloneqq \min\left\{\frac{\beta+1}{2\beta+d_U}, \frac{1}{2}\right\}.$$

The one-state construction uses the generator approximation and covering bounds of [SAL24] for their general latent-pushforward model. For each integer $N \geq 2$, take a finite N^{-1}-net of their bounded neural generator sieve in the uniform norm, and include the constant generator $g(u) = y_0$. Denote the resulting class by $\mathcal{G}_N$, and fix an enumeration with this constant generator first. The sieve and its wavelet approximation precision can be chosen so that every generator takes values in $\mathcal{Y}$ and

$$\sup_{g \in \mathcal{G}_N} \mathrm{Lip}(g) \leq B_0, \quad \log(2|\mathcal{G}_N|) \leq B_0 N^{d_U/(2\beta+d_U)}(\log N)^{c_1}, \tag{13}$$

$$A_N \coloneqq \sup_{g^* \in \mathcal{H}^{\beta+1}_{K_0}} \min_{g \in \mathcal{G}_N} W_1((g^*)_\# U, g_\# U) \leq B_0(\log N)^{c_1} N^{-(\beta+1)/(2\beta+d_U)}. \tag{14}$$

The constants $B_0, c_1 > 0$, like all constants below, may depend on the fixed dimensions and model parameters. To obtain the bounded sieve, restrict the wavelet generator class to maps with uniform norm at most $2K_0$. The truncated wavelet expansion approximates each target uniformly at order $N^{-(\beta+1)/(2\beta+d_U)}$, and its neural basis functions can be chosen with arbitrarily small C^1 approximation error. Consequently, for all sufficiently large N, these approximants remain in the bounded sieve. The wavelet coefficient bounds give a uniform first-derivative bound, since $\beta \geq 1$; choose the neural approximation precision so that the additional first-derivative error is at most one. Passage to a finite net preserves the displayed bounds after changing their constants. For the finitely many smaller N, take $\mathcal{G}_N = \{g : g(u) = y_0\}$ and enlarge the constants. These approximation and covering arguments apply coordinatewise for every $p \geq 1$.

For $p = 1$, the generator construction can equivalently be applied in $\mathbb{R}^2$ to $(g^*, 0)$ and projected onto its first coordinate; the approximation, covering, and Lipschitz bounds are preserved.

For each pair $g, g' \in \mathcal{G}_N$, choose an anchored Kantorovich potential $D_{g,g'} \in \mathcal{L}_1$ such that

$$\int D_{g,g'} \mathrm{d}(g_\# U - g'_\# U) = W_1(g_\# U, g'_\# U),$$

and set

$$\mathcal{D}_N = \{0\} \cup \{\pm D_{g,g'} : g, g' \in \mathcal{G}_N\}.$$

These finitely many potentials exist by compactness of $\mathcal{L}_1$ in the uniform norm. Thus $\mathcal{D}_N$ is symmetric, is contained in $\mathcal{L}_1$, and satisfies

$$d_{\mathcal{D}_N}(g_\# U, g'_\# U) = W_1(g_\# U, g'_\# U), \quad g, g' \in \mathcal{G}_N, \tag{15}$$

where $d_{\mathcal{D}_N}(\eta,\nu) = \max_{D\in\mathcal{D}_N} \int D\mathrm{d}(\eta-\nu)$. This is the finite optimal-potential critic construction for the general pushforward model in [SAL24].

For $\eta \in \mathcal{P}(\mathcal{Y})$, let $Z_1,\dots,Z_k$ be i.i.d. with law η, and let $U_1,\dots,U_\ell$ be independent copies of U, independent of the observations. Risk expectations average over both samples and any independent estimator randomization; the subscript in $\mathbb{E}_\eta$ specifies the observational outcome law. For $k,\ell \geq 1$, define

$$J_{k,\ell,N}(g) = \max_{D\in\mathcal{D}_N} \left\{ \frac{1}{k}\sum_{i=1}^{k} D(Z_i) - \frac{1}{\ell}\sum_{i=1}^{\ell} D(g(U_i)) \right\}. \tag{16}$$

The first minimizer in the fixed enumeration of $\mathcal{G}_N$ is a measurable estimator. For a deterministic tolerance $\tau \geq 0$, a τ-approximate minimizer has criterion value at most $\inf_{g\in\mathcal{G}_N} J_{k,\ell,N}(g) + \tau$. For a measurable map $g_0 : [0,1]^{d_U} \to \mathcal{Y}$, write $\eta_0 = (g_0)_{\#}U$.

Lemma 4 (One-state WGAN oracle primitive)**.** *Suppose $g_0 \in \mathcal{H}^{\beta+1}_{K_0}$ and $\hat{g} \in \mathcal{G}_N$ is a measurable τ-approximate minimizer of* (16) *based on observations from η_0. For every fixed $c_* > 0$, if $\min\{k,\ell\} \geq c_* N$, then*

$$\sup_{g_0\in\mathcal{H}^{\beta+1}_{K_0}} \mathbb{E}[W_1(\eta_0, \hat{g}_{\#}U)] \leq C(\log N)^c N^{-a} + \tau.$$

The constants may also depend on c_*.

Proof of Lemma 4. Choose $\bar{g} \in \mathcal{G}_N$ with $W_1(\eta_0, \bar{g}_{\#}U) \leq A_N$. Since $\mathcal{D}_N \subset \mathcal{L}_1$ is symmetric, $d_{\mathcal{D}_N}$ is a symmetric pseudometric with $d_{\mathcal{D}_N} \leq W_1$ by Kantorovich–Rubinstein duality. For $g \in \mathcal{G}_N$, the triangle inequality for W_1, (15), and the triangle inequality for $d_{\mathcal{D}_N}$ therefore give

$$\begin{aligned} W_1(\eta_0, g_{\#}U) &\leq A_N + W_1(\bar{g}_{\#}U, g_{\#}U) = A_N + d_{\mathcal{D}_N}(\bar{g}_{\#}U, g_{\#}U) \\ &\leq A_N + d_{\mathcal{D}_N}(\bar{g}_{\#}U, \eta_0) + d_{\mathcal{D}_N}(\eta_0, g_{\#}U) \\ &\leq 2A_N + d_{\mathcal{D}_N}(\eta_0, g_{\#}U). \end{aligned}$$

Write

$$\begin{aligned} S_k &= \max_{D\in\mathcal{D}_N} \left| \frac{1}{k}\sum_{i=1}^{k} D(Z_i) - \int D\mathrm{d}\eta_0 \right|, \\ T_\ell &= \max_{g\in\mathcal{G}_N,\, D\in\mathcal{D}_N} \left| \frac{1}{\ell}\sum_{i=1}^{\ell} D(g(U_i)) - \mathbb{E}[D(g(U))] \right|. \end{aligned}$$

Since $\mathcal{D}_N$ is finite, $|J_{k,\ell,N}(g) - d_{\mathcal{D}_N}(\eta_0, g_{\#}U)| \leq S_k + T_\ell$ for every $g \in \mathcal{G}_N$. The approximate minimizing property of $\hat{g}$ therefore gives

$$\begin{aligned} d_{\mathcal{D}_N}(\eta_0, \hat{g}_{\#}U) &\leq J_{k,\ell,N}(\hat{g}) + S_k + T_\ell \\ &\leq J_{k,\ell,N}(\bar{g}) + \tau + S_k + T_\ell \\ &\leq d_{\mathcal{D}_N}(\eta_0, \bar{g}_{\#}U) + 2S_k + 2T_\ell + \tau \\ &\leq A_N + 2S_k + 2T_\ell + \tau. \end{aligned}$$

Combined with the previous display, this is the deterministic basic inequality

$$W_1(\eta_0, \hat{g}_{\#}U) \leq 3A_N + 2S_k + 2T_\ell + \tau. \tag{17}$$

Since every $D \in \mathcal{D}_N \subset \mathcal{L}_1$ satisfies $\|D\|_\infty \leq \mathrm{diam}(\mathcal{Y})$ and $\mathrm{Lip}(D) \leq 1$, while $\sup_{g\in\mathcal{G}_N} \mathrm{Lip}(g) \leq B_0$ by (13) and $\mathrm{Lip}(g_0) \leq C$, each of the finite classes $\{D\circ g_0 : D \in \mathcal{D}_N\}$ and $\{D\circ g : D \in \mathcal{D}_N, g \in \mathcal{G}_N\}$ consists of functions on $[0,1]^{d_U}$ uniformly bounded by $\mathrm{diam}(\mathcal{Y})$ and uniformly Lipschitz with constant at most $\max\{B_0, C\}$. Let $\mathcal{A}$ denote either class, and write $Pf = \mathbb{E}[f(U)]$ and $P_s f = s^{-1}\sum_{i=1}^{s} f(V_i)$ for independent copies $V_1,\dots,V_s$ of U. Let $\mathcal{N}(\epsilon, \mathcal{A}, \|\cdot\|_\infty)$ be the minimum number of uniform-norm balls of radius ϵ needed to cover $\mathcal{A}$. By (13) and $|\mathcal{D}_N| \leq 2|\mathcal{G}_N|^2 + 1$, counting the finite class gives $|\mathcal{A}| \leq |\mathcal{G}_N|(2|\mathcal{G}_N|^2 + 1)$ and hence the first bound below, while the classical Kolmogorov–Tikhomirov entropy bound for uniformly bounded Lipschitz functions on $[0,1]^{d_U}$ gives the second:

$$\log \mathcal{N}(\epsilon, \mathcal{A}, \|\cdot\|_\infty) \leq C \min\left\{ N^{d_U/(2\beta+d_U)} (\log N)^{c_1}, \epsilon^{-d_U} \right\}, \quad 0 < \epsilon < 1.$$

Apply Dudley's entropy-integral bound

$$\mathbb{E}[\sup_{f\in\mathcal{A}}|(P_s-P)f|] \le Cs^{-1/2}\int_0^{\mathrm{diam}(\mathcal{Y})}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{A},\|\cdot\|_\infty)}\mathrm{d}\epsilon,$$

splitting the integral over $(0,t_N]$, $(t_N,1]$, and $(1,\mathrm{diam}(\mathcal{Y})]$ with $t_N = N^{-1/(2\beta+d_U)}$, and using the cardinality bound on $(0,t_N]$ and the Lipschitz entropy bound on $(t_N,1]$. For a sample size $s \ge c_*N$, this yields

$$\mathbb{E}\big[\sup_{f\in\mathcal{A}}|(P_s-P)f|\big] \le \frac{C}{\sqrt{s}}\left\{t_N N^{d_U/(2(2\beta+d_U))}(\log N)^{c_1/2} + \int_{t_N}^1 \epsilon^{-d_U/2}\mathrm{d}\epsilon + 1\right\}.$$

Here $t_N N^{d_U/(2(2\beta+d_U))} = N^{(d_U-2)/(2(2\beta+d_U))}$, and $\int_{t_N}^1 \epsilon^{-d_U/2}\mathrm{d}\epsilon$ is bounded by a constant for $d_U < 2$, by $\log(1/t_N) \lesssim \log N$ for $d_U = 2$, and by $Ct_N^{1-d_U/2} = CN^{(d_U-2)/(2(2\beta+d_U))}$ for $d_U > 2$. Since $s \ge c_*N$, the right-hand side is at most $C(\log N)^c N^{-1/2}$ when $d_U \le 2$ and at most $C(\log N)^c N^{-1/2+(d_U-2)/(2(2\beta+d_U))} = C(\log N)^c N^{-(\beta+1)/(2\beta+d_U)}$ when $d_U > 2$, that is, at most $C(\log N)^c N^{-a}$ in both cases. For S_k, represent the observations as $g_0(V_i)$ with independent uniform V_i; for T_ℓ, apply the same bound directly to the noise sample. Substituting these bounds and (14) into (17) proves the result. □

In the following minimax risk, $\hat{\eta}_k$ ranges over all measurable, possibly randomized estimators based on k i.i.d. observations and taking values in $\mathcal{P}(\mathcal{Y})$.

Lemma 5 (One-state minimax lower bound)**.** *Under the stated dimension and smoothness conditions, for all integers $k \ge 1$,*

$$\inf_{\hat{\eta}_k}\sup_{g^*\in\mathcal{H}^{\beta+1}_{K_0}}\mathbb{E}[W_1((g^*)_\#U,\hat{\eta}_k)] \ge c(k\vee 1)^{-a}.$$

Proof of Lemma 5. For sufficiently large k, apply Lemma 13 with $N = k$ and $\zeta \le 1/16$, and write $\eta_\theta = \eta_\theta^k$. With the Hellinger normalization specified in Appendix F.3, the tensorization bound $H^2(P^{\otimes k},Q^{\otimes k}) \le kH^2(P,Q)$ and Lemma 13(ii) give $H^2(\eta_\theta^{\otimes k},\eta_{\theta'}^{\otimes k}) \le \zeta \le 1/16$ for every pair (θ,θ') differing in one bit, hence $\mathrm{TV}(\eta_\theta^{\otimes k},\eta_{\theta'}^{\otimes k}) \le \sqrt{H^2(\eta_\theta^{\otimes k},\eta_{\theta'}^{\otimes k})} \le 1/4$. Given any estimate $\hat{\eta}_k$, let $\hat{\theta}$ index its closest vertex in W_1, breaking ties lexicographically. The triangle inequality and the choice of $\hat{\theta}$ give $W_1(\eta_\theta,\eta_{\hat{\theta}}) \le W_1(\eta_\theta,\hat{\eta}_k) + W_1(\hat{\eta}_k,\eta_{\hat{\theta}}) \le 2W_1(\eta_\theta,\hat{\eta}_k)$, so Lemma 13(iii) yields

$$W_1(\eta_\theta,\hat{\eta}_k) \ge \frac{c\Delta_k}{2}d_H(\theta,\hat{\theta}).$$

Under the uniform prior on $\theta\in\{0,1\}^{K_k}$, Assouad's lemma bounds the error probability of each bit below by $(1-1/4)/2 = 3/8$, so the prior-averaged risk satisfies $\mathbb{E}[d_H(\theta,\hat{\theta})] \ge 3K_k/8$. Hence the maximal risk is at least $cK_k\Delta_k \asymp k^{-a}$. The construction in Appendix F.3 proves this family directly using smooth transport perturbations, following the testing approach of [TY23]. For the finitely many smaller k, monotonicity of minimax risk in sample size extends the bound after decreasing c. Independent estimator randomization preserves the testing bound. □

Lemma 6 (Binomial negative moment)**.** *Let $K \sim \mathrm{Binomial}(n,p)$, $a > 0$, and $q \ge 0$. There exists $C < \infty$, depending only on (a,q), such that*

$$\mathbb{E}[(\log(K\vee 2))^q(K\vee 1)^{-a}] \le C(\log(n\vee 2))^q(np\vee 1)^{-a}.$$

Proof of Lemma 6. If $np < 2$, the left side is bounded by a constant and $(np\vee 1)^{-a} \ge 2^{-a}$. If $np \ge 2$, split on $A = \{K \ge np/2\}$. On A, $(\log(K\vee 2))^q(K\vee 1)^{-a} \le C(\log(n\vee 2))^q(np)^{-a}$. On A^c, the expression is at most $(\log(n\vee 2))^q$, while Chernoff's bound gives $\mathbb{P}(A^c) \le \exp(-np/8)$. Since $\exp(-x/8)(\log(n\vee 2))^q \le Cx^{-a}(\log(n\vee 2))^q$ for $x = np \ge 2$, the claim follows. □

E.2 Finite-State Upper Bound

For conditional outcome laws $\mu_j, \nu_j \in \mathcal{P}(\mathcal{Y})$, $j = 1, \dots, M$, write $P(\mathrm{d}w, \mathrm{d}y) = \sum_{j=1}^{M} q_j \delta_j(\mathrm{d}w)\mu_j(\mathrm{d}y)$ and $R(\mathrm{d}w, \mathrm{d}y) = \sum_{j=1}^{M} q_j \delta_j(\mathrm{d}w)\nu_j(\mathrm{d}y)$.

Lemma 7 (Finite-state extended-Wasserstein identity)**.** *For P and R with the same finite conditioning marginal,*

$$\mathrm{eW}_1(P, R) = \sum_{j=1}^{M} q_j W_1(\mu_j, \nu_j).$$

Proof of Lemma 7. This is Lemma 2 specialized to finite $\mathcal{W}$: the diagonal constraint forces the coupling to decompose independently over states. □

Proof of Theorem 1. Since critic coordinates are independent, $\sup_{h \in \mathcal{H}_n^{\mathrm{disc}}} \hat{L}_n^{\mathrm{disc}}(g, h)$ is the sum of statewise suprema. The minimization over $g = (g_1, \dots, g_M)$ therefore separates across states; in each state, choose the first minimizer in the fixed enumeration of $\mathcal{G}_n$. Conditional on the state sequence, the outcomes with $W_i = j$ are i.i.d. from μ_j^*. By overlap and $q_j \geq q_{\min}$, $\pi_j \geq \kappa q_j \geq \kappa q_{\min}$. On the event $A_j = \{N_j^{\mathrm{obs}} \geq n\pi_j/2\}$, $\min\{N_j^{\mathrm{obs}}, n\} \geq (\kappa q_{\min}/2)n$. Conditional on N_j^{obs}, the state-j estimator is the one-state estimator of Appendix E.1 with $N = n$, $k = N_j^{\mathrm{obs}}$, and $\ell = n$, so on A_j Lemma 4 with $c_* = \kappa q_{\min}/2$ gives

$$\mathbb{E}[W_1(\mu_j^*, \hat{\mu}_j) \mid N_j^{\mathrm{obs}}] \leq C(\log n)^c n^{-a},$$

where $\hat{\mu}_j$ is the generated law in state j. On A_j^c, the error is at most $\mathrm{diam}(\mathcal{Y})$, and Chernoff's bound gives $\mathbb{P}(A_j^c) \leq \exp(-n\pi_j/8) \leq \exp(-\kappa q_{\min} n/8)$. Hence

$$\mathbb{E}[W_1(\mu_j^*, \hat{\mu}_j)] \leq C(\log n)^c n^{-a} + \mathrm{diam}(\mathcal{Y}) \exp(-\kappa q_{\min} n/8) \leq C(\log n)^c n^{-a}.$$

By Lemma 7,

$$\mathbb{E}[\mathrm{eW}_1(P^*, P_{\hat{g}_n^{\mathrm{disc}}})] = \sum_{j=1}^{M} q_j \mathbb{E}[W_1(\mu_j^*, \hat{\mu}_j)] \leq C(\log n)^c n^{-a},$$

because M is fixed and $\sum_j q_j = 1$. □

Proof of Corollary 1. The upper bound is Theorem 1. For the lower bound, restrict to the submodel with $Q_{\mathrm{obs}} = Q_\rho$, conditional laws δ_0 in states $j \geq 2$, and only $\mu_1^* = (g^*)_{\#}U$ varying over $\mathcal{H}_{K_0}^{\beta+1}$. For any estimator $\bar{P}_n(\mathrm{d}j, \mathrm{d}y) = \sum_j q_j \delta_j(\mathrm{d}j)\bar{\mu}_{n,j}(\mathrm{d}y)$,

$$\mathrm{eW}_1(P^*, \bar{P}_n) \geq q_1 W_1(\mu_1^*, \bar{\mu}_{n,1}) \geq q_{\min} W_1(\mu_1^*, \bar{\mu}_{n,1}).$$

This is a causal submodel: draw an independent uniform variable and use the corresponding statewise generators for the potential outcomes. From n i.i.d. observations $Z_1, \dots, Z_n \sim \mu_1^*$, independently draw $J_1, \dots, J_n \sim Q_\rho$. Set $Y_i = Z_i$ when $J_i = 1$ and $Y_i = 0$ otherwise. The pairs (J_i, Y_i) are exactly an observational sample from this finite-state submodel. Applying any finite-state estimator and retaining its first conditional law therefore gives a randomized one-state estimator based on n observations. Lemma 5 and the first display then give

$$\sup_{\mathcal{M}_{\mathrm{disc}}} \mathbb{E}[\mathrm{eW}_1(P^*, \bar{P}_n)] \geq c q_1 n^{-a} \geq c q_{\min} n^{-a}.$$

Taking the infimum over estimators proves the claim. □

F Proofs of Section 5

F.1 Continuous Finite-Resolution Geometry and Besov Embedding

Let $\mathcal{L} = \{f : \mathcal{Y} \to \mathbb{R} : f(y_0) = 0,\ \mathrm{Lip}(f) < \infty\}$ with norm $\|f\|_{\mathcal{L}} = \mathrm{Lip}(f)$. For $p_B \in [1, \infty)$ and $\boldsymbol{s} = (s_1, \dots, s_{d_W}) \in (0, 1)^{d_W}$, define the vector-valued anisotropic Besov–Nikolskii norm of a Bochner-measurable map $H : \mathcal{W} \to \mathcal{L}$ by

$$\|H\|_{B^{\boldsymbol{s}}_{p_B,\infty}(\mathcal{W};\mathcal{L})} := \|H\|_{L^{p_B}(\mathcal{W};\mathcal{L})} + \max_{1 \leq j \leq d_W} \sup_{0 < \tau \leq 1} \tau^{-s_j} \|H(\cdot + \tau e_j) - H(\cdot)\|_{L^{p_B}(\mathcal{W}_{j,\tau};\mathcal{L})},$$

where $\mathcal{W}_{j,\tau} = \{w \in \mathcal{W} : w + \tau e_j \in \mathcal{W}\}$. For measurable $S \subset \mathcal{W}$, $\|H\|_{L^{p_B}(S;\mathcal{L})} = (\int_S \|H(w)\|_{\mathcal{L}}^{p_B} \, \mathrm{d}w)^{1/p_B}$; all such norms in this subsection use Lebesgue measure. We use Borel, strongly measurable representatives of H, setting them equal to zero on exceptional null sets. Under Assumption 4, these changes also occur on Q_ρ-null sets. The anchored Lipschitz space $\mathcal{L}$ is a Banach space, and $|f(y) - g(y)| \le \operatorname{diam}(\mathcal{Y})\|f - g\|_{\mathcal{L}}$, so evaluation is continuous. In particular, finite-range step maps are strongly measurable and yield jointly measurable critics. This is the anisotropic Besov–Nikolskii structure used in classical nonparametric theory and modern deep approximation theory [Tri83, NvS97, KLP01, HL02, Suz19, SN21]. For a radius $L > 0$, define

$$\mathcal{H}_{\boldsymbol{s},p_B}(L) \coloneqq \{h(w,y) = H(w)(y) : \|H\|_{B^{\boldsymbol{s}}_{p_B,\infty}(\mathcal{W};\mathcal{L})} \le L,\ H(w) \in \mathcal{L}_1 \text{ for a.e. } w\}.$$

For laws $P(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\mu_w(\mathrm{d}y)$ and $R(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_w(\mathrm{d}y)$, define, for $q_C > 0$,

$$\mu_C^\rho = q_C^{-1} \int_C \mu_w Q_\rho(\mathrm{d}w), \quad \nu_C^\rho = q_C^{-1} \int_C \nu_w Q_\rho(\mathrm{d}w). \tag{18}$$

Under Assumption 4, every dyadic cell has positive q_C.

Define the finite-resolution discrepancy by

$$\mathrm{eW}_1^{(\boldsymbol{m})}(P,R) \coloneqq \sup_{h \in \mathcal{F}^{(\boldsymbol{m})}} \{\mathbb{E}_P[h] - \mathbb{E}_R[h]\}.$$

Proposition 3 (Coarsened conditional Wasserstein identity)**.** *For P and R sharing Q_ρ,*

$$\mathrm{eW}_1^{(\boldsymbol{m})}(P,R) = \sum_{C \in \Pi_{\boldsymbol{m}}} q_C W_1(\mu_C^\rho, \nu_C^\rho).$$

Proof of Proposition 3. Each $h \in \mathcal{F}^{(\boldsymbol{m})}$ has form $h(w,y) = \sum_C \mathbf{1}\{w \in C\} f_C(y)$ with $f_C \in \mathcal{L}_1$. Expanding expectations gives

$$\mathbb{E}_P[h] - \mathbb{E}_R[h] = \sum_C q_C \left\{ \int f_C \mathrm{d}\mu_C^\rho - \int f_C \mathrm{d}\nu_C^\rho \right\}.$$

The choices of f_C are independent across cells, so the supremum separates. Ordinary Kantorovich–Rubinstein duality on $\mathcal{Y}$ proves the claim. □

Proposition 4 (Stepwise critics are anisotropic Besov–Nikolskii critics)**.** *Fix $p_B \in [1,\infty)$ and $\boldsymbol{s}$ satisfying $0 < s_j < 1/p_B$ for all j. There exists $C < \infty$, depending only on $(p_B, \boldsymbol{s}, d_W)$, such that, for every $\boldsymbol{m}$,*

$$\mathcal{F}^{(\boldsymbol{m})} \subset \mathcal{H}_{\boldsymbol{s},p_B}\big(C \max_j 2^{m_j s_j}\big). \tag{19}$$

Proof of Proposition 4. Let $h(w,y) = \sum_C \mathbf{1}\{w \in C\} f_C(y)$ and set $H(w) = h(w,\cdot)$. Since $\mathrm{Lip}(f_C) \le 1$, $\|H\|_{L^{p_B}(\mathcal{W};\mathcal{L})} \le 1$. Fix j and $\tau \in (0,1]$. Since w and $w + \tau e_j$ lie in different cells only if one of the $2^{m_j} - 1$ internal grid hyperplanes orthogonal to e_j separates them, the set $A_{j,\tau} = \{w \in \mathcal{W}_{j,\tau} : H(w + \tau e_j) \ne H(w)\}$ is contained in the union of the slabs of width τ below these hyperplanes, and its Lebesgue measure is bounded by $\min\{1, (2^{m_j} - 1)\tau\} \le C \min\{1, 2^{m_j}\tau\}$. On $A_{j,\tau}$, $\|H(w + \tau e_j) - H(w)\|_{\mathcal{L}} \le 2$. Therefore

$$\|H(\cdot + \tau e_j) - H(\cdot)\|_{L^{p_B}(\mathcal{W}_{j,\tau};\mathcal{L})} \le C \min\{1, (2^{m_j}\tau)^{1/p_B}\}.$$

If $\tau \le 2^{-m_j}$, since $1/p_B - s_j > 0$, multiplying by τ^{-s_j} gives $C 2^{m_j/p_B} \tau^{1/p_B - s_j} \le C 2^{m_j/p_B} (2^{-m_j})^{1/p_B - s_j} = C 2^{m_j s_j}$. If $\tau > 2^{-m_j}$, the minimum is 1, so multiplying by τ^{-s_j} gives $C\tau^{-s_j} \le C(2^{-m_j})^{-s_j} = C 2^{m_j s_j}$. Taking suprema over $\tau \in (0,1]$ and $1 \le j \le d_W$ proves (19). □

For a generator g and $q_C > 0$, define the target-design generated cell law

$$\nu_{g,C}^\rho = q_C^{-1} \int_C \nu_{g,w} Q_\rho(\mathrm{d}w),$$

with arbitrary values on cells of zero target mass, and define the cell-resolution generated law $P_g^{(\boldsymbol{m})}(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_{g,C_{\boldsymbol{m}}(w)}^\rho(\mathrm{d}y)$.

Proposition 5 (Finite-resolution sandwich and raw-law oscillation)**.** *Under Assumption 4, with $p_B, \boldsymbol{s}, C$ as in Proposition 4, for any P, R sharing Q_ρ,*

$$\mathrm{eW}_1^{(\boldsymbol{m})}(P,R) \le \sup_{h\in\mathcal{H}_{\boldsymbol{s},p_B}(C\max_j 2^{m_j s_j})} \{\mathbb{E}_P[h]-\mathbb{E}_R[h]\} \le \mathrm{eW}_1(P,R). \tag{20}$$

If $R(\mathrm{d}w,\mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_C(\mathrm{d}y)$ for $w \in C$ and $P = P^*$ satisfies Assumption 6, then

$$\mathrm{eW}_1(P^*,R) \le \mathrm{eW}_1^{(\boldsymbol{m})}(P^*,R) + L_* b_{\boldsymbol{m}}. \tag{21}$$

Finally, for any generator g,

$$\mathrm{eW}_1(P^*,P_g) \le \mathrm{eW}_1(P^*,P_g^{(\boldsymbol{m})}) + \int W_1(\nu_{g,w},\nu^\rho_{g,C_{\boldsymbol{m}}(w)})Q_\rho(\mathrm{d}w). \tag{22}$$

Proof of Proposition 5. The first inequality in (20) follows from Proposition 4. The second follows from extended Kantorovich–Rubinstein duality, because every admissible Besov critic is sectionwise 1-Lipschitz in y.

For (21), use the target cell mixture μ_C^ρ from (18) with $P = P^*$. Since R is cellwise constant,

$$\mathrm{eW}_1(P^*,R) = \sum_C \int_C W_1(\mu_w^*,\nu_C)Q_\rho(\mathrm{d}w).$$

The triangle inequality gives

$$W_1(\mu_w^*,\nu_C) \le W_1(\mu_w^*,\mu_C^\rho) + W_1(\mu_C^\rho,\nu_C).$$

For $w, w' \in C$, Assumption 6 yields $W_1(\mu_w^*,\mu_{w'}^*) \le L_* b_{\boldsymbol{m}}$. Convexity of W_1 in its second argument gives $W_1(\mu_w^*,\mu_C^\rho) \le L_* b_{\boldsymbol{m}}$. Summing over cells and using Proposition 3 proves (21).

Equation (22) is the triangle inequality for eW_1 and the disintegration formula:

$$\mathrm{eW}_1(P_g,P_g^{(\boldsymbol{m})}) = \int W_1(\nu_{g,w},\nu^\rho_{g,C_{\boldsymbol{m}}(w)})Q_\rho(\mathrm{d}w).$$

□

F.2 Stratified Finite-Resolution Upper Bound

For theoretical calibration, use $\mathcal{D}_{N_{\boldsymbol{m}}}$, where $\mathcal{D}_N$ is the one-state WGAN critic class from Appendix E.1. Throughout this subsection, μ_C^ρ denotes the target cell mixture in (18) with $\mu_w = \mu_w^*$. For $C \in \Pi_{\boldsymbol{m}}$, define

$$\mu_C^{\mathrm{obs}} = \pi_C^{-1}\int_C \mu_w^* Q_{\mathrm{obs}}(\mathrm{d}w).$$

Overlap gives $\pi_C \ge \kappa q_C$.

Lemma 8 (Cell-normalized sampling)**.** *Conditional on N_C^{obs}, the outcomes $\{Y_i^{\mathrm{obs}} : W_i \in C\}$ are i.i.d. from μ_C^{obs}. Moreover, $N_C^{\mathrm{obs}} \sim \mathrm{Binomial}(n,\pi_C)$ and $\pi_C \ge \kappa q_C$.*

Proof of Lemma 8. By consistency and ignorability, the conditional law of Y^{obs} given $W = w$ is μ_w^*. Conditioning on $W \in C$ averages this kernel with respect to $Q_{\mathrm{obs}}(\mathrm{d}w \mid C)$, giving μ_C^{obs}. The count is binomial by i.i.d. sampling. Finally, $q_C = \int_C w_\rho(w)Q_{\mathrm{obs}}(\mathrm{d}w) \le \kappa^{-1}\pi_C$. □

Lemma 9 (Observed-target cell bias)**.** *Under Assumption 6,*

$$W_1(\mu_C^{\mathrm{obs}},\mu_C^\rho) \le L_* b_{\boldsymbol{m}}$$

for every $C \in \Pi_{\boldsymbol{m}}$.

Proof of Lemma 9. By convexity of W_1 in both arguments,

$$W_1(\mu_C^{\mathrm{obs}},\mu_C^\rho) \le \int_C\int_C W_1(\mu_w^*,\mu_{w'}^*)Q_{\mathrm{obs}}(\mathrm{d}w \mid C)Q_\rho(\mathrm{d}w' \mid C).$$

For $w, w' \in C$, $|w_j - w_j'| \le 2^{-m_j}$. Assumption 6 proves the claim. □

Lemma 10 (Proxy-to-exact inequality for the finite-resolution generator)**.** *Under Assumptions 3, 4, and 6, for any measurable conditional generator g,*

$$\mathrm{eW}_1(P^*, P_g^{(\boldsymbol{m})}) \le \sum_{C \in \Pi_{\boldsymbol{m}}} q_C W_1(\mu_C^{\mathrm{obs}}, \nu_{g,C}^{\rho}) + L_* b_{\boldsymbol{m}}.$$

Proof of Lemma 10. For $w \in C$, convexity and Assumption 6 give

$$W_1(\mu_w^*, \mu_C^{\mathrm{obs}}) \le \int_C W_1(\mu_w^*, \mu_{w'}^*) Q_{\mathrm{obs}}(\mathrm{d}w' \mid C) \le L_* b_{\boldsymbol{m}}.$$

Thus

$$W_1(\mu_w^*, \nu_{g,C}^{\rho}) \le L_* b_{\boldsymbol{m}} + W_1(\mu_C^{\mathrm{obs}}, \nu_{g,C}^{\rho}).$$

Integrating over C under Q_ρ and summing over cells proves the result. □

Use $\eta_0 = (g_0)_{\#}U$ as in Appendix E.1.

Lemma 11 (Structural one-state stability)**.** *Suppose $\eta \in \mathcal{P}(\mathcal{Y})$, $g_0 \in \mathcal{H}_{K_0}^{\beta+1}$, and $W_1(\eta, \eta_0) \le \delta$ for some $\delta \ge 0$. Suppose $\hat{g} \in \mathcal{G}_N$ is a measurable τ-approximate minimizer of* (16)*, based on k i.i.d. observations from η and ℓ independent copies of U, with the two samples independent. For each fixed $c_* > 0$, if $\min\{k, \ell\} \ge c_* N$, then*

$$\mathbb{E}_\eta[W_1(\eta, \hat{g}_{\#}U)] \le 3\delta + C(\log N)^c N^{-a} + \tau.$$

Proof of Lemma 11. On an auxiliary probability space, draw i.i.d. pairs (Z_i, Z_i^0) from an optimal coupling of η and η_0, independently of the generated-noise sample. Set $\xi_k = k^{-1}\sum_{i=1}^k \|Z_i - Z_i^0\|$. Since every $D \in \mathcal{D}_N$ is 1-Lipschitz, the empirical criteria computed with (Z_i) and (Z_i^0), denoted by J and J^0, satisfy

$$\max_{g \in \mathcal{G}_N} |J(g) - J^0(g)| \le \xi_k.$$

Consequently, for every $g \in \mathcal{G}_N$,

$$J^0(\hat{g}) \le J(\hat{g}) + \xi_k \le J(g) + \tau + \xi_k \le J^0(g) + \tau + 2\xi_k,$$

so $J^0(\hat{g}) \le \min_{g \in \mathcal{G}_N} J^0(g) + \tau + 2\xi_k$. Applying the samplewise inequality (17) to the clean sample (Z_i^0) with random tolerance $\tau + 2\xi_k$ yields $W_1(\eta_0, \hat{g}_{\#}U) \le 3A_N + 2S_k + 2T_\ell + \tau + 2\xi_k$. The empirical-process terms are bounded as in Lemma 4, and $\mathbb{E}[\xi_k] = W_1(\eta, \eta_0) \le \delta$, so

$$\mathbb{E}[W_1(\eta_0, \hat{g}_{\#}U)] \le 3A_N + 2\mathbb{E}[S_k] + 2\mathbb{E}[T_\ell] + \tau + 2\mathbb{E}[\xi_k] \le C(\log N)^c N^{-a} + \tau + 2\delta.$$

Finally, the triangle inequality

$$W_1(\eta, \hat{g}_{\#}U) \le W_1(\eta, \eta_0) + W_1(\eta_0, \hat{g}_{\#}U) \le \delta + W_1(\eta_0, \hat{g}_{\#}U)$$

proves the result. □

At resolution $\boldsymbol{m}$, consider the finite product class of generators satisfying $g(w, u) = g_C(u)$ for $w \in C$, where $g_C \in \mathcal{G}_{N_{\boldsymbol{m}}}$ for every $C \in \Pi_{\boldsymbol{m}}$. Using the observational sample and the same target-design and noise samples as in (6), choose in each cell the first minimizer of its empirical criterion over $\mathcal{G}_{N_{\boldsymbol{m}}}$ and denote its law by $\hat{\nu}_C$. This defines a measurable minimizer over the product class, because both the generator and critic choices separate across cells. The zero-count conventions in (4)–(5) and the constant first generator specify the estimator also in empty cells.

Lemma 12 (Uniform cell oracle)**.** *Uniformly over cells $C \in \Pi_{\boldsymbol{m}}$ and models in $\mathcal{M}_{\mathrm{cont}}$,*

$$\mathbb{E}[W_1(\mu_C^{\mathrm{obs}}, \hat{\nu}_C)] \le C b_{\boldsymbol{m}} + C(\log n)^c (\kappa n 2^{-|\boldsymbol{m}|_1})^{-a}.$$

Proof of Lemma 12. Choose any representative $w_C \in C$. Assumption 5 gives $\mu_{w_C}^* = (g_{w_C}^*)_{\#}U$ with $g_{w_C}^* \in \mathcal{H}_{K_0}^{\beta+1}$. By Assumption 6 and convexity,

$$W_1(\mu_C^{\mathrm{obs}}, \mu_{w_C}^*) \le L_* b_{\boldsymbol{m}}.$$

Put $t = \kappa n 2^{-|\boldsymbol{m}|_1}$. If $t < 2$, the stated bound follows from the diameter bound after increasing its constant. Suppose $t \geq 2$. The two cell counts satisfy

$$N_C^{\mathrm{obs}} \sim \mathrm{Binomial}(n, \pi_C), \quad M_C^{\rho} \sim \mathrm{Binomial}(n, q_C),$$

and the two counts are independent. Moreover, Assumption 4 gives $q_C \geq \underline{q}\, 2^{-|\boldsymbol{m}|_1}$, so $nq_C \geq \underline{q}\, n 2^{-|\boldsymbol{m}|_1} \geq \underline{q}\, t$ because $\kappa \leq 1$, and overlap gives $n\pi_C \geq \kappa n q_C \geq \underline{q}\, t$. Define

$$A_C = \{N_C^{\mathrm{obs}} \geq n\pi_C/2,\ M_C^{\rho} \geq nq_C/2\}.$$

On A_C, since $N_{\boldsymbol{m}} \leq 2t$,

$$\min\{N_C^{\mathrm{obs}}, M_C^{\rho}\} \geq (\underline{q}/4) N_{\boldsymbol{m}}.$$

Conditional on both counts, the observed outcomes in the cell are i.i.d. from μ_C^{obs} and the selected noise draws are independent copies of U. The routed generator is constant in the conditioning coordinate within the cell. Therefore, Lemma 11, with $k = N_C^{\mathrm{obs}}$, $\ell = M_C^{\rho}$, and $\delta = L_* b_{\boldsymbol{m}}$, gives on A_C

$$\mathbb{E}[W_1(\mu_C^{\mathrm{obs}}, \hat{\nu}_C) \mid N_C^{\mathrm{obs}}, M_C^{\rho}] \leq 3L_* b_{\boldsymbol{m}} + C(\log N_{\boldsymbol{m}})^c N_{\boldsymbol{m}}^{-a}.$$

Chernoff's inequality gives $\mathbb{P}(A_C^c) \leq 2\exp(-\underline{q}\, t/8)$, and on A_C^c we use $W_1 \leq \mathrm{diam}(\mathcal{Y})$. Taking expectations over the counts,

$$\mathbb{E}[W_1(\mu_C^{\mathrm{obs}}, \hat{\nu}_C)] \leq 3L_* b_{\boldsymbol{m}} + C(\log N_{\boldsymbol{m}})^c N_{\boldsymbol{m}}^{-a} + 2\,\mathrm{diam}(\mathcal{Y}) \exp(-\underline{q}\, t/8).$$

Since $t \leq N_{\boldsymbol{m}} \leq 2t$, $\log N_{\boldsymbol{m}} \lesssim \log n$, and $\exp(-\underline{q}\, t/8) \leq Ct^{-a}$, the desired bound follows, uniformly over cells and over the stated model class. □

Proposition 6 (Finite-resolution continuous upper bound)**.** *Assume that the implemented estimator satisfies, uniformly over $\mathcal{M}_{\mathrm{cont}}$, the finite-resolution transfer condition* (24) *with error $\varepsilon_{n,\boldsymbol{m}}$. There exist constants $C, c > 0$, depending only on fixed model parameters, such that for every resolution $\boldsymbol{m}$,*

$$\sup_{(P_{\mathrm{obs}}, P^*) \in \mathcal{M}_{\mathrm{cont}}} \mathbb{E}\big[\mathrm{eW}_1(P^*, \hat{P}_{n,\boldsymbol{m}}^{\mathrm{cont}})\big] \leq C\{b_{\boldsymbol{m}} + (\log n)^c (\kappa n 2^{-|\boldsymbol{m}|_1})^{-a}\} + \varepsilon_{n,\boldsymbol{m}}. \tag{23}$$

Proof of Proposition 6. First consider the ideal routed finite-resolution estimator with cell laws $\hat{\nu}_C$. By Lemma 12,

$$\mathbb{E}\left[\sum_C q_C W_1(\mu_C^{\mathrm{obs}}, \hat{\nu}_C)\right] \leq Cb_{\boldsymbol{m}} + C(\log n)^c (\kappa n 2^{-|\boldsymbol{m}|_1})^{-a},$$

because $\sum_C q_C = 1$. Lemma 10 transfers this proxy risk to the extended-Wasserstein risk of the routed estimator, adding only another constant multiple of $b_{\boldsymbol{m}}$.

For the implemented estimator, the transfer condition (24) gives

$$\mathbb{E}\left[\sum_C q_C W_1(\nu^{\rho}_{\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}, C}, \hat{\nu}_C)\right] \leq \varepsilon_{n,\boldsymbol{m}}.$$

Applying Lemma 10 to $\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}$ and the triangle inequality $W_1(\mu_C^{\mathrm{obs}}, \nu^{\rho}_{\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}, C}) \leq W_1(\mu_C^{\mathrm{obs}}, \hat{\nu}_C) + W_1(\hat{\nu}_C, \nu^{\rho}_{\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}, C})$ in each cell,

$$\mathbb{E}\big[\mathrm{eW}_1(P^*, \hat{P}^{\mathrm{cont}}_{n,\boldsymbol{m}})\big] \leq \mathbb{E}\left[\sum_C q_C W_1(\mu_C^{\mathrm{obs}}, \hat{\nu}_C)\right] + \mathbb{E}\left[\sum_C q_C W_1(\hat{\nu}_C, \nu^{\rho}_{\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}, C})\right] + L_* b_{\boldsymbol{m}},$$

and the two displays above bound the first two terms, which gives (23). □

Proof of Theorem 2. The finite-resolution inequality is Proposition 6. If $\kappa n < 1$, the claim follows from $\mathrm{eW}_1(P^*, \hat{P}^{\mathrm{cont}}_{n,\boldsymbol{m}_n}) \leq \mathrm{diam}(\mathcal{Y})$ after increasing C. Assume henceforth that $\kappa n \geq 1$. For $m_{n,j} = \lfloor r \log_2(\kappa n)/\alpha_j \rfloor \vee 0$, we have $2^{-\alpha_j m_{n,j}} \leq 2^{\alpha_j} (\kappa n)^{-r} \leq 2(\kappa n)^{-r}$, hence $b_{\boldsymbol{m}_n} \leq 2d_W(\kappa n)^{-r}$, and $2^{-|\boldsymbol{m}_n|_1}$ lies between $(\kappa n)^{-r\bar{d}}$ and $2^{d_W}(\kappa n)^{-r\bar{d}}$, hence $\kappa n 2^{-|\boldsymbol{m}_n|_1} \asymp (\kappa n)^{1-r\bar{d}} = (\kappa n)^{1/(1+a\bar{d})}$ because $r = a/(1+a\bar{d})$. Hence

$$(\kappa n 2^{-|\boldsymbol{m}_n|_1})^{-a} \asymp (\kappa n)^{-a/(1+a\bar{d})} = (\kappa n)^{-r}.$$

The assumed transfer bound at $\boldsymbol{m}_n$ yields (8). The raw-law statement, when needed, follows from (22). □

Write $\hat{g} = \hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}_n}$ for the following corollary.

Corollary 2 (Raw conditional law)**.** *Under the hypotheses of Theorem 2, suppose that, uniformly over $\mathcal{M}_{\mathrm{cont}}$ and almost surely, either $\nu_{\hat{g},w}$ is constant within each cell of $\Pi_{\boldsymbol{m}_n}$ or*

$$W_1(\nu_{\hat{g},w}, \nu_{\hat{g},w'}) \le L_G \sum_{j=1}^{d_W} |w_j - w'_j|, \quad w, w' \in \mathcal{W},$$

for a fixed $0 \le L_G < \infty$ independent of n. Then (8) *also holds with $\hat{P}^{\mathrm{cont}}_{n,\boldsymbol{m}_n}$ replaced by $P_{\hat{g}}$, with constants possibly depending on L_G.*

Proof of Corollary 2. For a cellwise constant law, the within-cell term in (22) is zero. Otherwise, for $\kappa n \ge 1$, convexity of W_1 in its second argument, the cell side lengths $2^{-m_{n,j}}$, and the bound $2^{-m_{n,j}} \le 2^{-\alpha_j m_{n,j}} \le 2(\kappa n)^{-r}$ from the proof of Theorem 2, which uses $\alpha_j \le 1$, give

$$\int W_1(\nu_{\hat{g},w}, \nu^{\rho}_{\hat{g},C_{\boldsymbol{m}_n}(w)}) Q_\rho(\mathrm{d}w) \le L_G \sum_{j=1}^{d_W} 2^{-m_{n,j}} \le 2L_G d_W (\kappa n)^{-r}.$$

Combine this bound with (22) and Theorem 2; the case $\kappa n < 1$ follows from compactness as above. □

For $t > 0$, Markov's inequality also gives

$$\mathbb{E}\Big[Q_\rho\big(\{w \in \mathcal{W} : W_1(\mu^*_w, \nu_{\hat{g},w}) > t\}\big)\Big] \le \frac{C(\log n)^c (\kappa n)^{-r}}{t}.$$

This bounds the expected target mass of states with conditional Wasserstein error above t.

F.3 Continuous-Conditioning Minimax Lower Bound

The lower bound uses an anisotropic Assouad construction. We state the one-state lower-bound primitive in the form needed below and write $d = d_U$. For probability measures P, Q, use the squared Hellinger distance $H^2(P,Q) = \int(\sqrt{\mathrm{d}P/\mathrm{d}\lambda} - \sqrt{\mathrm{d}Q/\mathrm{d}\lambda})^2 \mathrm{d}\lambda$, where λ dominates both measures, and the total variation distance $\mathrm{TV}(P,Q) = \sup_A |P(A) - Q(A)|$, where the supremum is over measurable sets. Let d_H denote the Hamming distance, the number of coordinates on which two binary vectors of the same length differ. For an indexed generator family, write $\eta^N_{\theta,t} = (g^N_{\theta,t})_{\#}U$ and $\eta^N_\theta = \eta^N_{\theta,1}$. For an open set S, $C^\infty_c(S)$ denotes the smooth functions with compact support in S; these functions are extended by zero outside S.

Lemma 13 (One-state Assouad primitive)**.** *Suppose that $1 \le d \le p$, $\beta \ge 1$, $K_0 > 0$, and $\mathcal{Y} = B^p(0, 2K_0)$. For any fixed $\zeta > 0$ and every sufficiently large integer N, there exist an integer K_N, a number $\Delta_N > 0$, a fixed generator g_0, and generators $\{g^N_{\theta,t} : \theta \in \{0,1\}^{K_N},\ t \in [0,1]\}$ in $\mathcal{H}^{\beta+1}_{K_0}$ with the following properties.*

(i) $g^N_{\theta,0} = g_0$ and $g^N_{\theta,t} = g_0 + t(g^N_{\theta,1} - g_0)$, with $\|g^N_{\theta,1} - g_0\|_\infty \le N^{-a}$.

(ii) For every pair differing in exactly one coordinate,

$$\sup_{t \in [0,1]} H^2(\eta^N_{\theta,t}, \eta^N_{\theta',t}) \le \frac{\zeta}{N}.$$

(iii) For all θ, θ',

$$W_1(\eta^N_\theta, \eta^N_{\theta'}) \ge c\Delta_N d_H(\theta, \theta'), \quad K_N \Delta_N \asymp N^{-a}.$$

Constants may depend on ζ and the fixed model parameters.

Proof of Lemma 13. Choose $\sigma \in (0,1)$ sufficiently small and define

$$g_0(u) = \sigma\big(\sin(2\pi u_1), \dots, \sin(2\pi u_d), 0, \dots, 0\big).$$

This periodic map has image in $B^p(0, K_0/2)$ and belongs to $\mathcal{H}^{\beta+1}_{K_0/2}$ after decreasing σ if necessary. Identifying the first d coordinates with $\mathbb{R}^d$, the law of $g_0(U)$ has density

$$f_0(z) = \begin{cases} \prod_{i=1}^{d} \dfrac{1}{\pi\sqrt{\sigma^2 - z_i^2}}, & z \in (-\sigma, \sigma)^d, \\ 0, & \text{otherwise.} \end{cases}$$

This density is smooth and bounded above and below by positive constants on $[-\sigma/2, \sigma/2]^d$. Fix a nonzero, nonnegative $b \in C_c^\infty((-1/8, 1/8)^d)$ and $\varphi \in C_c^\infty((-1/3, 1/3)^d)$ such that $\varphi(z) = z_1$ on $[-1/4, 1/4]^d$. If $d > 2$, let

$$J_N = \lceil N^{1/(2\beta+d)} \rceil, \quad \ell_N = \frac{\sigma}{2J_N}, \quad K_N = J_N^d, \quad A_N = c_0 \ell_N^{\beta+1}.$$

If $d \le 2$, let

$$J_N = K_N = 1, \quad \ell_N = \sigma/2, \quad A_N = c_0 N^{-1/2}.$$

Here $c_0 > 0$ is a fixed constant chosen below. Divide $[-\sigma/4, \sigma/4]^d$ into K_N cubes Q_k of side length ℓ_N, and denote their centers by z_k. For $z \in \mathbb{R}^d$, define

$$S^N_{\theta,t}(z) = z + tA_N \sum_{k=1}^{K_N} \theta_k b\big((z - z_k)/\ell_N\big) e_1.$$

For sufficiently small c_0, $\|DS^N_{\theta,t} - I\|_\infty \le 1/4$ uniformly in (N, θ, t). Here $DS^N_{\theta,t}$ is the Jacobian matrix, I is the $d \times d$ identity matrix, and the uniform matrix norm uses the Euclidean operator norm. Each map is a diffeomorphism, equals the identity near the boundary of every Q_k, and maps Q_k onto itself. Indeed, all coordinates except the first are unchanged, and its first coordinate is strictly increasing on every line parallel to e_1. Define $g^N_{\theta,t}$ by applying $S^N_{\theta,t}$ to the first d coordinates of g_0 and leaving the remaining coordinates zero.

The scaled bumps have disjoint supports separated by distances proportional to ℓ_N, and consequently

$$\left\| A_N \sum_k \theta_k b\big((\cdot - z_k)/\ell_N\big) e_1 \right\|_{C^{\beta+1}(\mathbb{R}^d; \mathbb{R}^d)} \le C A_N \ell_N^{-(\beta+1)}.$$

This is the usual Hölder norm; at integer smoothness the bound applies to all derivatives through that order. Composition with the fixed smooth map g_0 preserves this bound up to a fixed constant. The right-hand side is at most Cc_0 when $d > 2$ and at most $Cc_0\ell_N^{-(\beta+1)}$ when $d \le 2$. Decreasing c_0 therefore ensures $g^N_{\theta,t} \in \mathcal{H}^{\beta+1}_{K_0}$ uniformly. It also ensures $\|g^N_{\theta,1} - g_0\|_\infty \le N^{-a}$. The image of every generator is contained in the image of g_0, since $S^N_{\theta,t}$ maps $[-\sigma, \sigma]^d$ onto itself. This proves (i) and the required support condition.

To prove (ii), consider an edge changing bit k. The two pushforward densities coincide outside Q_k. On Q_k, one density is f_0 and the other is the pushforward by $S(z) = z + tA_N b((z - z_k)/\ell_N)e_1$. Its density is

$$f_t(x) = \frac{f_0(S^{-1}(x))}{J_t(x)}, \quad J_t(x) := 1 + (tA_N/\ell_N)\partial_1 b((S^{-1}(x) - z_k)/\ell_N).$$

Since $\|S^{-1}(x) - x\| \le A_N\|b\|_\infty$ and $|J_t(x) - 1| \le (A_N/\ell_N)\|\partial_1 b\|_\infty \le 1/4$, the decomposition $|f_t(x) - f_0(x)| \le J_t(x)^{-1}\{|f_0(S^{-1}(x)) - f_0(x)| + f_0(x)|1 - J_t(x)|\}$ together with uniform bounds on f_0 and ∇f_0 on Q_k and $\ell_N \le \sigma/2 < 1$ gives $f_t, f_0 \ge c > 0$ on Q_k and $\|f_t - f_0\|_{L^\infty(Q_k)} \le C(A_N + A_N/\ell_N) \le CA_N/\ell_N$. Since $(\sqrt{f_t} - \sqrt{f_0})^2 = (f_t - f_0)^2/(\sqrt{f_t} + \sqrt{f_0})^2 \le (f_t - f_0)^2/(4c)$ on Q_k, integrating over the cube of volume ℓ_N^d gives

$$H^2(\eta^N_{\theta,t}, \eta^N_{\theta',t}) \le CA_N^2\ell_N^{d-2} \le Cc_0^2 N^{-1},$$

where the last inequality uses $A_N^2\ell_N^{d-2} = c_0^2\ell_N^{2\beta+d} \asymp c_0^2 N^{-1}$ when $d > 2$ and $A_N^2\ell_N^{d-2} = c_0^2(\sigma/2)^{d-2}N^{-1}$ when $d \le 2$. For $d < p$, these densities are taken with respect to Lebesgue

measure on the common d-dimensional coordinate subspace. Choosing c_0 small enough gives (ii) for the prescribed ζ.

For (iii), write $y_{1:d} = (y_1, \dots, y_d)$ and put $\varphi_k(y) = \ell_N \varphi((y_{1:d} - z_k)/\ell_N)$ for $y \in \mathbb{R}^p$. The kth bump is supported in $z_k + (-\ell_N/8, \ell_N/8)^d$ and displaces points by at most $A_N \|b\|_\infty \le \ell_N/8$ for c_0 sufficiently small, so every displaced point remains in $z_k + [-\ell_N/4, \ell_N/4]^d$, where $\varphi_k(y) = y_1 - z_{k,1}$. Since each $S^N_{\theta,1}$ maps Q_k onto itself, depends only on θ_k on Q_k, and φ_k vanishes outside Q_k, writing η^N_θ as the image of $f_0(z)\mathrm{d}z$ under $S^N_{\theta,1}$ gives, for any $\theta, \theta' \in \{0,1\}^{K_N}$,

$$\begin{aligned}\int \varphi_k \mathrm{d}(\eta^N_\theta - \eta^N_{\theta'}) &= \int_{Q_k} \big\{\varphi_k(S^N_{\theta,1}(z)) - \varphi_k(S^N_{\theta',1}(z))\big\} f_0(z)\mathrm{d}z \\ &= A_N(\theta_k - \theta'_k) \int_{Q_k} b((z - z_k)/\ell_N) f_0(z)\mathrm{d}z.\end{aligned}$$

By the change of variables $u = (z - z_k)/\ell_N$ and $f_0 \ge c_f > 0$ on Q_k, the last integral is bounded below by $c\ell_N^d$. Because the supports of the φ_k are disjoint,

$$F_{\theta,\theta'}(y) = \frac{1}{\max\{1, \|\nabla\varphi\|_\infty\}} \sum_k \mathrm{sgn}(\theta_k - \theta'_k)\varphi_k(y)$$

is 1-Lipschitz on $\mathbb{R}^p$. Since $\mathrm{sgn}(\theta_k - \theta'_k)(\theta_k - \theta'_k) = |\theta_k - \theta'_k|$, Kantorovich–Rubinstein duality gives

$$W_1(\eta^N_\theta, \eta^N_{\theta'}) \ge \int F_{\theta,\theta'} \mathrm{d}(\eta^N_\theta - \eta^N_{\theta'}) \ge c A_N \ell_N^d \sum_{k=1}^{K_N} |\theta_k - \theta'_k| = c A_N \ell_N^d d_H(\theta, \theta').$$

Taking $\Delta_N = A_N \ell_N^d$ yields $K_N \Delta_N = (\sigma/2)^d A_N \asymp N^{-a}$. □

Proof of Theorem 3. Fix the known target design Q_ρ satisfying Assumption 4, and put $D = d_W \ge 1$ and $\bar{d} = \sum_{j=1}^D \alpha_j^{-1}$. Fix $\chi \in C_c^\infty((0,1))$ with $0 \le \chi \le 1$ and $\chi = 1$ on $[1/4, 3/4]$, and put $C_\chi = \max\{1, \|\chi'\|_\infty\}$ and $C_1 = \max\{2, C_\chi\}$. Choose

$$B \ge \max\{1, C_1/L_*\}, \quad 0 < \zeta \le (16\bar{q}B^{\bar{d}})^{-1}.$$

These constants are fixed independently of n and κ. Write

$$N = \left\lceil (\kappa n)^{1/(1+a\bar{d})} \right\rceil, \quad \delta = N^{-a}, \quad h_j = (B\delta)^{1/\alpha_j}.$$

Then $\delta \asymp (\kappa n)^{-r}$ and $v := \prod_j h_j = B^{\bar{d}} N^{-a\bar{d}}$. All following constructions apply once κn is sufficiently large.

Choose a fixed $\tau \in (0, \min\{1/2, (4\bar{q})^{-1}\}]$. Set $L_1 = \tau$, $L_j = 1$ for $j \ge 2$, and $J_j = \lfloor L_j/(2h_j) \rfloor$. For each multi-index $k \in \prod_j \{0, \dots, J_j - 1\}$, take the rectangle

$$R_k = \prod_{j=1}^D [2k_j h_j, (2k_j + 1)h_j]$$

and its central subrectangle B_k, whose side lengths are $h_j/2$. Relabel the rectangles by $\ell = 1, \dots, M$, where $M = \prod_{j=1}^D J_j$, and set $A = \bigcup_{\ell=1}^M R_\ell$. Their closures are pairwise disjoint, and

$$M \asymp v^{-1}, \quad c \le \sum_{\ell=1}^M Q_\rho(B_\ell), \quad Q_\rho(A) \le 1/4.$$

Indeed, $L_j/(4h_j) \le J_j \le L_j/(2h_j)$ for large n, so $M = \prod_j J_j$ lies between $4^{-D}\tau/v$ and $2^{-D}\tau/v$; the density bounds then give $\sum_\ell Q_\rho(B_\ell) \ge \underline{q} M \prod_j (h_j/2) \ge \underline{q} 8^{-D}\tau$ and $Q_\rho(A) \le \bar{q} M \prod_j h_j \le \bar{q} 2^{-D}\tau \le 1/4$ by the choice of τ. If $a_{\ell,j}$ is the lower endpoint of R_ℓ in coordinate j, put

$$\psi_\ell(w) = \prod_{j=1}^D \chi((w_j - a_{\ell,j})/h_j).$$

Then $\psi_\ell = 1$ on B_ℓ, $\mathrm{supp}(\psi_\ell) \subset R_\ell$, and $\|\partial_j \psi_\ell\|_\infty \le C_\chi / h_j$.

Define the observational design by

$$\frac{\mathrm{d}Q_{\mathrm{obs}}}{\mathrm{d}Q_\rho}(w) = \kappa \mathbf{1}_A(w) + \frac{1 - \kappa Q_\rho(A)}{1 - Q_\rho(A)} \mathbf{1}_{A^c}(w).$$

This defines a probability measure, since its total mass is $\kappa Q_\rho(A) + (1 - \kappa Q_\rho(A)) = 1$, and its density relative to Q_ρ is at least κ because $\kappa \le 1$ implies $(1 - \kappa Q_\rho(A))/(1 - Q_\rho(A)) \ge \kappa$. Hence $w_\rho = \mathrm{d}Q_\rho / \mathrm{d}Q_{\mathrm{obs}} \le \kappa^{-1}$, which is Assumption 3.

Apply Lemma 13 at N with the prescribed ζ. For $\Theta = (\theta_1, \dots, \theta_M) \in (\{0,1\}^{K_N})^M$, define

$$g_\Theta(w, u) = g_0(u) + \sum_{\ell=1}^{M} \psi_\ell(w) (g^N_{\theta_\ell, 1}(u) - g_0(u)), \quad \mu_{\Theta, w} = (g_\Theta(w, \cdot))_{\#} U.$$

Let $P^*_\Theta(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w) \mu_{\Theta, w}(\mathrm{d}y)$. At most one summand is nonzero for every w. Consequently each section is either g_0 or $g^N_{\theta_\ell, \psi_\ell(w)}$, and belongs to $\mathcal{H}^{\beta+1}_{K_0}$. The construction is jointly measurable. It is also a causal submodel: independently of the assigned $(X, T) \sim Q_{\mathrm{obs}}$, draw U and set $Y_\Theta(t) = g_\Theta((X, t), U)$ for every t, with $Y^{\mathrm{obs}}_\Theta = Y_\Theta(T)$. This gives consistency and conditional ignorability and induces the specified observational kernel.

For points differing only in coordinate j, the displacement bound $\|g^N_{\theta,1} - g_0\|_\infty \le \delta$, the derivative bound $\|\partial_j \psi_\ell\|_\infty \le C_\chi / h_j$, and the inequality $\min\{x, y\} \le x^{1-\alpha_j} y^{\alpha_j}$ for $x, y \ge 0$ give

$$\|g_\Theta(w, \cdot) - g_\Theta(w', \cdot)\|_\infty \le \min\{2\delta, C_\chi \delta h_j^{-1} |w_j - w'_j|\} \le C_1 \delta h_j^{-\alpha_j} |w_j - w'_j|^{\alpha_j} = \frac{C_1}{B} |w_j - w'_j|^{\alpha_j},$$

where the last equality uses $h_j^{\alpha_j} = B\delta$. Coupling $\mu_{\Theta, w}$ and $\mu_{\Theta, w'}$ through the same U bounds W_1 by the uniform distance of the generators, and moving one coordinate at a time along a path from w to w' gives

$$W_1(\mu_{\Theta, w}, \mu_{\Theta, w'}) \le \frac{C_1}{B} \sum_{j=1}^{D} |w_j - w'_j|^{\alpha_j}.$$

The choice of B therefore gives Assumption 6.

Consider global vertices differing in a single bit in rectangle R_ℓ. Write $P_{\Theta, \mathrm{obs}}(\mathrm{d}w, \mathrm{d}y) = Q_{\mathrm{obs}}(\mathrm{d}w) \mu_{\Theta, w}(\mathrm{d}y)$. Let $\mathbb{E}_\Theta$ denote expectation under $P^{\otimes n}_{\Theta, \mathrm{obs}}$, including any independent estimator randomization. Since their design marginals agree, the conditional Hellinger identity applies; the two conditional laws coincide outside R_ℓ, on R_ℓ they equal $\eta^N_{\theta_\ell, \psi_\ell(w)}$ and $\eta^N_{\theta'_\ell, \psi_\ell(w)}$, and $Q_{\mathrm{obs}}(R_\ell) = \kappa Q_\rho(R_\ell) \le \kappa \bar{q} v$, so Lemma 13(ii) gives

$$H^2(P_{\Theta, \mathrm{obs}}, P_{\Theta', \mathrm{obs}}) = \int H^2(\mu_{\Theta, w}, \mu_{\Theta', w}) Q_{\mathrm{obs}}(\mathrm{d}w) \le \frac{\zeta}{N} \kappa Q_\rho(R_\ell) \le \frac{\zeta}{N} \kappa \bar{q} v.$$

By the tensorization bound $H^2(P^{\otimes n}, Q^{\otimes n}) \le n H^2(P, Q)$, the identity $v = B^{\bar{d}} N^{-a\bar{d}}$, and $N \ge (\kappa n)^{1/(1+a\bar{d})}$,

$$H^2(P^{\otimes n}_{\Theta, \mathrm{obs}}, P^{\otimes n}_{\Theta', \mathrm{obs}}) \le \zeta \bar{q} B^{\bar{d}} (\kappa n) N^{-(1+a\bar{d})} \le \zeta \bar{q} B^{\bar{d}} \le \frac{1}{16}.$$

Hence every such pair has n-sample total variation distance at most $\sqrt{1/16} = 1/4$.

To reduce arbitrary estimators to binary decisions, write $\hat{P}_n(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w) \hat{\mu}_{n, w}(\mathrm{d}y)$ and, for each ℓ, let $\hat{\theta}_\ell$ minimize

$$\int_{B_\ell} W_1(\hat{\mu}_{n, w}, \eta^N_\theta) Q_\rho(\mathrm{d}w), \quad \theta \in \{0,1\}^{K_N},$$

with lexicographic tie breaking. The minimization is over a finite set and its integrands are measurable, so these decisions are measurable. On B_ℓ, the true conditional law is $\eta^N_{\theta_\ell}$, because $\psi_\ell = 1$ there. By the triangle inequality and the minimizing property of $\hat{\theta}_\ell$,

$$\begin{aligned} Q_\rho(B_\ell) W_1(\eta^N_{\theta_\ell}, \eta^N_{\hat{\theta}_\ell}) &\le \int_{B_\ell} W_1(\eta^N_{\theta_\ell}, \hat{\mu}_{n, w}) Q_\rho(\mathrm{d}w) + \int_{B_\ell} W_1(\hat{\mu}_{n, w}, \eta^N_{\hat{\theta}_\ell}) Q_\rho(\mathrm{d}w) \\ &\le 2 \int_{B_\ell} W_1(\mu_{\Theta, w}, \hat{\mu}_{n, w}) Q_\rho(\mathrm{d}w), \end{aligned}$$

so Lemma 13(iii) gives

$$\int_{B_\ell} W_1(\mu_{\Theta,w}, \hat{\mu}_{n,w}) Q_\rho(\mathrm{d}w) \geq \frac{Q_\rho(B_\ell)}{2} W_1(\eta^N_{\theta_\ell}, \eta^N_{\hat{\theta}_\ell}) \geq c Q_\rho(B_\ell) \Delta_N d_H(\theta_\ell, \hat{\theta}_\ell).$$

Under the uniform prior on the MK_N bits, the testing bound implies an error probability at least $3/8$ for every bit: the average error in each neighboring pair is at least one half of one minus the total variation distance between its two n-sample observational laws. Restricting the integral defining $\mathrm{eW}_1(P^*_\Theta, \hat{P}_n)$ to $\bigcup_\ell B_\ell$, applying the previous display in each B_ℓ, and averaging over the uniform prior, which bounds the maximum over Θ from below, we obtain

$$\begin{aligned}
\sup_\Theta \mathbb{E}_\Theta[\mathrm{eW}_1(P^*_\Theta, \hat{P}_n)] &\geq c \sum_{\ell=1}^M Q_\rho(B_\ell) \Delta_N \mathbb{E}[d_H(\theta_\ell, \hat{\theta}_\ell)] \\
&\geq \frac{3c}{8} \sum_{\ell=1}^M Q_\rho(B_\ell) K_N \Delta_N \\
&\geq c N^{-a} \\
&\geq c(\kappa n)^{-r},
\end{aligned}$$

where the last two inequalities use $\sum_\ell Q_\rho(B_\ell) \geq c$, $K_N \Delta_N \asymp N^{-a}$, and $N^{-a} = \delta \asymp (\kappa n)^{-r}$. Independent randomization of an estimator leaves the testing bound unchanged. Taking the infimum over all estimators proves the theorem. □

F.4 Transfer from Routed Sieves to Single-Network Implementations

The proof uses an ideal routed estimator with one expert per cell. Write the implemented class as $\mathcal{G}^{\mathrm{cont}}_n = \{g_\theta : \theta \in \Theta_n\}$, where Θ_n is a compact metric space, $(\theta, w, u) \mapsto g_\theta(w, u)$ is jointly measurable, and $\theta \mapsto g_\theta(w, u)$ is continuous for every (w, u). Each empirical criterion is a maximum of finitely many continuous functions of θ. The measurable maximum theorem therefore gives a measurable minimizer; fix one such selection throughout. All risk expectations include both the observational sample and the independent simulation draws used in fitting. For a general implementation, the required condition is integrated finite-resolution transfer, not uniform approximation of discontinuous indicators. Let $\hat{\nu}_C$ denote the routed cell law from the proof of Proposition 6, and let $\nu^\rho_{\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}, C}$ be the generated cell law of the implemented estimator in (6). Let $\varepsilon_{n,\boldsymbol{m}} \geq 0$ be a deterministic bound on the integrated implementation error. We assume, uniformly over $(P_{\mathrm{obs}}, P^*) \in \mathcal{M}_{\mathrm{cont}}$,

$$\mathbb{E}\left[\sum_{C \in \Pi_{\boldsymbol{m}}} q_C W_1(\nu^\rho_{\hat{g}^{\mathrm{cont}}_{n,\boldsymbol{m}}, C}, \hat{\nu}_C)\right] \leq \varepsilon_{n,\boldsymbol{m}}. \tag{24}$$

Hard routing of the same fitted cell experts $\hat{\nu}_C$, with the same samples and minimizer selection, gives $\varepsilon_{n,\boldsymbol{m}} = 0$. The within-cell term in (22) is also zero, giving the raw-law guarantee in Corollary 2. For approximate empirical minimizers, the optimization tolerances in Lemma 11 enter the bound through their expected target-weighted average. The following proposition bounds the transfer error when soft stochastic gates combine the same fitted cell experts.

Write $\Pi_{\boldsymbol{m}} = \{C_1, \ldots, C_K\}$, with $K = |\Pi_{\boldsymbol{m}}|$, and set $r_k(w) = \mathbf{1}\{w \in C_k\}$. For expert measures ν_k and weight functions γ_k, use the notation

$$\eta = \int \sum_{k=1}^K |\gamma_k(w) - r_k(w)| Q_\rho(\mathrm{d}w), \quad \nu_w = \sum_{k=1}^K \gamma_k(w) \nu_k.$$

Let $\nu^\rho_{C_k}$ be the cell mixture of ν_w from (18), with arbitrary probability laws on cells of zero target mass.

Proposition 7 (Integrated soft-gate transfer)**.** *Suppose that $\nu_k \in \mathcal{P}(\mathcal{Y})$ and that the functions $\gamma_k : \mathcal{W} \to [0,1]$ are measurable and satisfy $\sum_{k=1}^K \gamma_k(w) = 1$ for every $w \in \mathcal{W}$. Then*

$$\sum_{k=1}^K q_{C_k} W_1(\nu^\rho_{C_k}, \nu_k) \leq \mathrm{diam}(\mathcal{Y}) \eta. \tag{25}$$

Moreover,

$$\int W_1(\nu_w, \nu^{\rho}_{C_{\boldsymbol{m}}(w)}) Q_\rho(\mathrm{d}w) \le 2\,\mathrm{diam}(\mathcal{Y})\eta. \tag{26}$$

Proof of Proposition 7. Fix $w \in C_k$. Since $r_k(w) = 1$ and $r_\ell(w) = 0$ for $\ell \neq k$, we have $\nu_k = \sum_{\ell=1}^K r_\ell(w)\nu_\ell$, and every $f \in \mathcal{L}_1$ satisfies $\|f\|_\infty \le \mathrm{diam}(\mathcal{Y})$, so Kantorovich–Rubinstein duality gives

$$W_1\left(\sum_{\ell=1}^K \gamma_\ell(w)\nu_\ell, \nu_k\right) = \sup_{f\in\mathcal{L}_1} \sum_{\ell=1}^K (\gamma_\ell(w) - r_\ell(w)) \int f \mathrm{d}\nu_\ell \le \mathrm{diam}(\mathcal{Y}) \sum_{\ell=1}^K |\gamma_\ell(w) - r_\ell(w)|.$$

For $q_{C_k} > 0$, convexity of W_1 in its first argument gives

$$W_1(\nu^{\rho}_{C_k}, \nu_k) \le q_{C_k}^{-1} \int_{C_k} W_1(\nu_w, \nu_k) Q_\rho(\mathrm{d}w).$$

Multiplying by q_{C_k} and summing over the positive-mass cells proves (25); zero-mass cells contribute nothing.

For $w \in C_k$,

$$W_1(\nu_w, \nu^{\rho}_{C_k}) \le W_1(\nu_w, \nu_k) + W_1(\nu_k, \nu^{\rho}_{C_k}).$$

Integrating and using the first part twice proves (26). □

For $\delta > 0$, define the boundary layer

$$\mathcal{B}_\delta = \left\{ w \in \mathcal{W} : \mathrm{dist}\left(w, \bigcup_{C\in\Pi_{\boldsymbol{m}}} \partial C\right) \le \delta \right\}.$$

Here dist is Euclidean point-to-set distance and ∂C is the boundary of C.

Lemma 14 (Dyadic boundary-layer control)**.** *Suppose Assumption 4 holds. For* $0 < \delta \le 1$*,*

$$Q_\rho(\mathcal{B}_\delta) \le C\delta \sum_{j=1}^{d_W} 2^{m_j}. \tag{27}$$

If, for constants $A, c_0 > 0$ and gate temperature $\tau > 0$, gates satisfying the conditions of Proposition 7 also satisfy

$$\sum_k |\gamma_k(w) - r_k(w)| \le A\exp(-c_0\delta/\tau) \quad \textit{for all } w \in \mathcal{W} \setminus \mathcal{B}_\delta,$$

then

$$\eta \le 2Q_\rho(\mathcal{B}_\delta) + A\exp(-c_0\delta/\tau). \tag{28}$$

Proof of Lemma 14. In coordinate j, the dyadic partition has $2^{m_j}+1$ boundary hyperplanes, including the two outer faces. The δ-neighborhood of these hyperplanes has Lebesgue measure at most $C\delta 2^{m_j}$. A union bound over coordinates and the density upper bound $q_\rho \le \bar{q}$ prove (27). On $\mathcal{B}_\delta$, the quantity $\sum_k |\gamma_k - r_k|$ is at most 2 because the soft and hard gate vectors are probability vectors. Outside $\mathcal{B}_\delta$, use the assumed exponential bound and integrate. □

Choosing δ and the gate temperature τ so that the right-hand side of (28) is $o((\kappa n)^{-r})$ makes the transfer term negligible. This is an integrated condition; no continuous network is required to approximate discontinuous indicators uniformly. Under the gate bound in Lemma 14, with A, c_0 independent of n, one choice at resolution $\boldsymbol{m}_n$ for $\kappa n > 1$ is

$$\delta_n = \frac{(\kappa n)^{-(r+1)}}{\sum_{j=1}^{d_W} 2^{m_{n,j}}}, \quad \tau_n = \frac{c_0\delta_n}{(r+1)\log(\kappa n)}.$$

Then $\eta = O((\kappa n)^{-(r+1)})$, so Proposition 7 makes both the transfer error and within-cell oscillation negligible relative to $(\kappa n)^{-r}$.

F.5 Outcome Truncation

Use the statewise integral representation of eW_1 for outcomes in $\mathbb{R}^p$. For $R > 0$, let T_R be projection onto $B^p(0, R)$ and define $P_R^*(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)((T_R)_\# \mu_w^*)(\mathrm{d}y)$. Write $(x)_+ = \max\{x, 0\}$.

Proposition 8 (Truncation bias)**.** *Suppose* $\int\int \|y\| \mu_w^*(\mathrm{d}y) Q_\rho(\mathrm{d}w) < \infty$. *For any estimated law* $\hat{P}$ *sharing* Q_ρ *and supported on* $\mathcal{W} \times B^p(0, R)$,

$$\mathrm{eW}_1(P^*, \hat{P}) \leq \mathrm{eW}_1(P_R^*, \hat{P}) + \int\int (\|y\| - R)_+ \mu_w^*(\mathrm{d}y) Q_\rho(\mathrm{d}w). \tag{29}$$

Proof of Proposition 8. Couple y with $T_R(y)$ at cost $(\|y\| - R)_+$ and apply the statewise triangle inequality before integrating over Q_ρ. □

Conditional on the fixed covariates and coefficients, the Gaussian and Student-t_5 components in the semi-synthetic benchmarks have truncation biases $O(e^{-cR^2})$ and $O(R^{-4})$, respectively; the latter follows by integrating the $O(R^{-5})$ tail probability. For Jobs, $Y = \mathrm{asinh}(re78/1000)$ compresses earnings tails but remains unbounded. A uniform bound $\mathbb{P}(|re78| > z \mid W = w) \lesssim z^{-\gamma}$ for some $\gamma > 0$ gives transformed-scale truncation bias $O(e^{-\gamma R})$. The experiments enforce a bounded output range through a bounded generator output layer (Appendix H.2). Equation (29) separates truncation bias from estimation error; extending the fixed-range minimax results to growing R additionally requires control of generator smoothness and range-dependent constants.

G Proofs of Section 6

Let $\mathcal{V}$ be a symmetric class of measurable functions containing zero and uniformly bounded by $B_\mathcal{V} < \infty$. Set $d_\mathcal{V}(\mu, \nu) = \sup_{f \in \mathcal{V}} \int f \mathrm{d}(\mu - \nu)$. Assume that a countable subclass $\mathcal{V}_0 \subset \mathcal{V}$ satisfies

$$d_\mathcal{V}(\mu, \nu) = \sup_{f \in \mathcal{V}_0} \int f \mathrm{d}(\mu - \nu)$$

for every pair of probability measures on $\mathcal{Y}$. All estimators below are measurable probability kernels. For $P(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\mu_w(\mathrm{d}y)$ and $R(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\nu_w(\mathrm{d}y)$, let μ_C^ρ and ν_C^ρ be the cell mixtures defined as in (18).

Lemma 15 (Finite-resolution IPM disintegration)**.** *For* P, R *sharing* Q_ρ,

$$d_\mathcal{V}^{(\boldsymbol{m})}(P, R) = \sum_{C \in \Pi_{\boldsymbol{m}}} q_C d_\mathcal{V}(\mu_C^\rho, \nu_C^\rho). \tag{30}$$

Proof of Lemma 15. Expanding expectations gives

$$\sum_C q_C \left\{ \int f_C \mathrm{d}\mu_C^\rho - \int f_C \mathrm{d}\nu_C^\rho \right\}.$$

Since the functions f_C are chosen independently across cells, the supremum separates and yields (30). □

Fix $a_\mathcal{V} > 0$ and assume the following one-state perturbation oracle for the local IPM. There exist constants $C, c > 0$, local generator classes, and estimators such that, whenever $\eta, \eta_0 \in \mathcal{P}(\mathcal{Y})$ satisfy $d_\mathcal{V}(\eta, \eta_0) \leq \delta$ for $\delta \geq 0$ and η_0 belongs to the pointwise smooth pushforward model, the estimator $\hat{\eta}_k^\mathcal{V}$ based on k i.i.d. samples from η satisfies

$$\mathbb{E}_\eta[d_\mathcal{V}(\eta, \hat{\eta}_k^\mathcal{V})] \leq C\{\delta + (\log(k \vee 2))^c (k \vee 1)^{-a_\mathcal{V}}\}. \tag{31}$$

Assume also that the conditional law is anisotropic Hölder under $d_\mathcal{V}$: for some $0 < L_\mathcal{V} < \infty$ and exponents $\alpha_j \in (0, 1]$,

$$d_\mathcal{V}(\mu_w^*, \mu_{w'}^*) \leq L_\mathcal{V} \sum_{j=1}^{d_W} |w_j - w_j'|^{\alpha_j}. \tag{32}$$

This inequality holds for all $w, w' \in \mathcal{W}$. Use $b_{\boldsymbol{m}}$ and $\bar{d}$ with these exponents, and assume that the constants in the local perturbation oracle are uniform over the pointwise model. Conditional on $N_C^{\mathrm{obs}} = k \geq 1$, let $\hat{\nu}_C^{\mathcal{V}}$ be the local estimator in (31), with any required simulation independent of the data. For $k = 0$, set $\hat{\nu}_C^{\mathcal{V}} = \delta_{y_0}$. Define the routed estimator by

$$\hat{P}_{n,\boldsymbol{m}}^{\mathcal{V}}(\mathrm{d}w, \mathrm{d}y) = Q_\rho(\mathrm{d}w)\hat{\nu}_{C_{\boldsymbol{m}}(w)}^{\mathcal{V}}(\mathrm{d}y).$$

For the optimized resolution in this section, set $r_{\mathcal{V}} = (a_{\mathcal{V}}^{-1} + \bar{d})^{-1}$ and $\boldsymbol{m}_n = (m_{n,1}, \ldots, m_{n,d_W})$, where

$$m_{n,j} = \left\lfloor \frac{r_{\mathcal{V}} \log_2(\kappa n)}{\alpha_j} \right\rfloor \vee 0.$$

Theorem 5 (Generic anisotropic IPM rate)**.** *Under Assumptions 1–3, 4, and 5, and under* (31)–(32)*, the routed IPM estimator defined above satisfies*

$$\mathbb{E}[d_{\mathcal{V}}^{(\boldsymbol{m})}(P^*, \hat{P}_{n,\boldsymbol{m}}^{\mathcal{V}})] \leq C\left(b_{\boldsymbol{m}} + (\log n)^c(\kappa n 2^{-|\boldsymbol{m}|_1})^{-a_{\mathcal{V}}}\right). \tag{33}$$

At the optimized resolution,

$$\mathbb{E}[d_{\mathcal{V}}^{(\boldsymbol{m}_n)}(P^*, \hat{P}_{n,\boldsymbol{m}_n}^{\mathcal{V}})] \leq C(\log n)^c(\kappa n)^{-r_{\mathcal{V}}}. \tag{34}$$

Proof of Theorem 5. The proof is the IPM analogue of Appendix F.2. For a cell C, conditional on N_C^{obs}, observed outcomes in the cell are i.i.d. from μ_C^{obs}. Choose a representative $w_C \in C$. By (32),

$$d_{\mathcal{V}}(\mu_C^{\mathrm{obs}}, \mu_{w_C}^*) \leq L_{\mathcal{V}} b_{\boldsymbol{m}}.$$

Conditional on $N_C^{\mathrm{obs}} = k \geq 1$, the local perturbation oracle gives

$$\mathbb{E}[d_{\mathcal{V}}(\mu_C^{\mathrm{obs}}, \hat{\nu}_C^{\mathcal{V}}) \mid N_C^{\mathrm{obs}} = k] \leq C\{L_{\mathcal{V}} b_{\boldsymbol{m}} + (\log(k \vee 2))^c(k \vee 1)^{-a_{\mathcal{V}}}\}.$$

For $k = 0$, uniform boundedness of $\mathcal{V}$ gives the same bound after increasing C. Apply Lemma 6 with $a = a_{\mathcal{V}}$ and use $n\pi_C \geq \underline{q}\,\kappa n 2^{-|\boldsymbol{m}|_1}$ to obtain

$$\mathbb{E}[d_{\mathcal{V}}(\mu_C^{\mathrm{obs}}, \hat{\nu}_C^{\mathcal{V}})] \leq C b_{\boldsymbol{m}} + C(\log n)^c(\kappa n 2^{-|\boldsymbol{m}|_1})^{-a_{\mathcal{V}}}.$$

Multiplying by q_C and summing over cells gives the same bound for the proxy risk because $\sum_C q_C = 1$. The target-versus-observational cell-mixture bias is controlled under (32) exactly as in Lemma 9. Lemma 15 then yields (33). If $\kappa n < 1$, (34) follows from the uniform boundedness of $\mathcal{V}$. For $\kappa n \geq 1$, optimizing the anisotropic resolution gives (34). □

H Detailed Experiments

This appendix provides the experimental protocols for the three benchmarks presented in Section 7. We begin by summarizing the datasets and data sources, followed by descriptions of the dataset-specific construction procedures, baseline methods, the primary evaluation metrics reported in Section 7, as well as additional metrics and supplementary results. The benchmark comparisons are run over 100 independent repetitions. Within each repetition, all methods use the same train/validation/test split, simulation seed, target design, and evaluation protocol; reported standard errors are standard errors over these 100 repetitions. The paired implementation diagnostics and synthetic sensitivity experiments use the seed counts specified in their respective subsections. Throughout all experiments, GANICE is trained using the stratified, cell-normalized objective introduced in Section 3.2. The main benchmark experiments use one shared conditional generator.

H.1 Dataset Information

Table 5 summarizes the three datasets. IHDP [Hil11, SJS17] and TCGA [WCM$^+$13] are semi-synthetic benchmarks: the covariates and treatment assignments are taken from standard causal datasets, while the potential-outcome laws are generated from known stochastic mechanisms. This allows direct evaluation of the full conditional interventional distribution. Jobs [LaL86, DW99] is a

real-data benchmark with randomized-trial validation: individual counterfactual distributions are not observed, so we evaluate model-implied arm-level interventional cumulative distribution functions against held-out randomized experimental arms. Table 6 reports the data sources and license or access terms.

For semi-synthetic benchmarks, the target designs are known by construction. For Jobs, the target design is the empirical covariate distribution of the held-out randomized NSW sample with equal treatment-arm weights. All continuous covariates are standardized using training-set statistics, and all train/validation/test splits are shared across methods within each repetition.

H.1.1 IHDP

Source and preprocessing. We use the IHDP covariates and treatment assignments from the benchmark popularized in counterfactual representation learning [Hil11, SJS17]. The raw [Hil11] file contains 985 randomized IHDP units. Following the standard [Hil11, SJS17] preprocessing, we remove non-white treated children and exclude race indicators, yielding 747 units, 25 covariates, and a binary treatment, with 139 treated units and 608 control units. Continuous covariates are standardized using training-set means and standard deviations, while binary covariates are kept as $0/1$ variables. We use stratified splits within treatment arms with training fraction 0.63, validation fraction 0.27, and the remaining fraction for testing. The seed of the data-generating process (DGP) and the split seed are varied across the 100 repetitions, and all methods share the resulting split within each repetition.

Stochastic potential-outcome law. The original IHDP benchmark is primarily a conditional-mean benchmark. To evaluate distributional causal learning, we retain the observed covariates and treatment assignments but generate stochastic potential outcomes from a known law. Let $\tilde{x} \in \mathbb{R}^d$ denote the standardized covariate vector, with $d = 25$. For scalar distributions, $\mathcal{N}(m, s^2)$ denotes the normal law with mean m and variance s^2, and $\mathrm{logit}^{-1}(z) = (1 + \exp(-z))^{-1}$. For fixed coefficient vectors in $\mathbb{R}^d$, sampled once per repetition, define

$$m_0(x) = \exp(0.2 b_0^\top \tilde{x}) - 1, \quad \tau(x) = 1 + 0.5 \tanh(b_\tau^\top \tilde{x}), \quad m_1(x) = m_0(x) + \tau(x).$$

For each treatment $t \in \{0, 1\}$, define

$$\pi_t(x) = \mathrm{logit}^{-1}(b_{\pi,t}^\top \tilde{x}), \quad \Delta_t(x) = 0.5 + 0.5 \tanh(b_{\Delta,t}^\top \tilde{x}),$$

$$\sigma_{t,1}(x) = 0.2 + 0.3\, \mathrm{logit}^{-1}(b_{\sigma,t}^\top \tilde{x}), \quad \sigma_{t,2}(x) = 0.3 + 0.5\, \mathrm{logit}^{-1}((b'_{\sigma,t})^\top \tilde{x}).$$

The potential-outcome law is

$$\begin{aligned} Y(t) \mid X = x \sim \pi_t(x) \mathcal{N}\big(m_t(x) + (1 - \pi_t(x))\Delta_t(x), \sigma_{t,1}^2(x)\big) \\ + (1 - \pi_t(x)) t_\nu\big(m_t(x) - \pi_t(x)\Delta_t(x), \sigma_{t,2}(x)\big), \end{aligned} \tag{35}$$

where $t_\nu(\ell, s)$ denotes a Student-t distribution with location ℓ, scale s, and degrees of freedom ν. We set $\nu = 5$. The coefficient entries of b_0 and b_τ are drawn independently from $\mathcal{N}(0, d^{-1})$. The entries of $b_{\pi,t}$, $b_{\Delta,t}$, $b_{\sigma,t}$, and $b'_{\sigma,t}$ are drawn independently from $\mathcal{N}(0, 2.5^2/d)$, separately for $t = 0, 1$. These coefficients are redrawn for each repetition and then held fixed within that repetition. This construction preserves the conditional mean $m_t(x)$ while inducing heteroskedasticity, skewness, heavy tails, and covariate-dependent distributional treatment heterogeneity.

Target design. Write n_{test} for the number of test units and $\hat{P}_{X,\mathrm{test}} = n_{\mathrm{test}}^{-1} \sum_{i=1}^{n_{\mathrm{test}}} \delta_{x_i}$ for their empirical covariate law. The target design is the empirical covariate distribution on the test set crossed with a uniform treatment intervention:

$$Q_\rho = \hat{P}_{X,\mathrm{test}} \otimes \mathrm{Unif}\{0, 1\}.$$

Thus, the evaluation averages the conditional interventional distributions across both treatment arms and over the held-out covariate distribution.

H.1.2 TCGA

Source and preprocessing. We use the processed TCGA gene-expression benchmark used in continuous-dose causal inference [BJvdS20, SLB+20]. The benchmark contains $9{,}659$ units and $4{,}000$ gene-expression features. We read RNA-seq arrays from the TCGA SQLite database, apply a $\log(1+\cdot)$ transform, min–max scale each gene, select the $4{,}000$ most variable genes, and normalize each sample by its row norm. The selected feature matrix is then standardized using training-set means and standard deviations. The train/validation/test split fractions are $0.64/0.16/0.20$. For the semi-synthetic assignment and outcome mechanisms below, we compute the principal component analysis (PCA) score vector $z(x) \in \mathbb{R}^{d_z}$ from the leading $d_z = 8$ principal components of the standardized expression matrix. All predictive models are trained on the full $4{,}000$-dimensional feature vector unless otherwise stated.

Treatment and dosage assignment. There are three treatment classes, $A \in \{1,2,3\}$, and a continuous dosage $D \in [0,1]$; the intervention variable is $T = (A, D)$. The treatment-specific coefficient vectors $v_a, r_a, \theta_a, q_a, s_a, u_{a,1}, u_{a,2}$ belong to $\mathbb{R}^{d_z}$; their sampling laws are specified below. Treatment is assigned according to

$$\mathbb{P}(A = a \mid X = x) = \frac{\exp(\gamma_A v_a^\top z(x))}{\sum_{b=1}^{3} \exp(\gamma_A v_b^\top z(x))},$$

where γ_A controls treatment-selection strength. For each treatment a, define the covariate-dependent reference dosage $d_a(x) = \mathrm{logit}^{-1}(r_a^\top z(x))$. The observed dosage is then sampled as $D \mid X = x, A = a \sim \mathrm{Beta}(1 + \alpha_D d_a(x), 1 + \alpha_D\{1 - d_a(x)\})$. We set $\gamma_A = 1.0$ and $\alpha_D = 8.0$.

Stochastic dose-response law. Let $m_a(x,d) = \iota_a + \theta_a^\top z(x) - \lambda_a\{d - d_a(x)\}^2 + \rho_a \sin(2\pi d)$ be the treatment- and dosage-specific conditional mean, where ι_a is a treatment-specific intercept. We generate outcomes from the mixture law

$$\begin{aligned} Y(a,d) \mid X = x \sim{}& \pi_a(x,d)\mathcal{N}\big(m_a(x,d) + (1-\pi_a(x,d))\Delta_a(x,d), \sigma_{a,1}^2(x,d)\big) \\ &+ \{1-\pi_a(x,d)\}\mathcal{N}\big(m_a(x,d) - \pi_a(x,d)\Delta_a(x,d), \sigma_{a,2}^2(x,d)\big), \end{aligned} \tag{36}$$

where

$$\pi_a(x,d) = \mathrm{logit}^{-1}(q_a^\top z(x) + 2d - 1), \quad \Delta_a(x,d) = 0.5 + 0.5\tanh(s_a^\top z(x) + d),$$
$$\sigma_{a,1}(x,d) = 0.1 + 0.3\,\mathrm{logit}^{-1}(u_{a,1}^\top z(x) + d), \quad \sigma_{a,2}(x,d) = 0.2 + 0.5\,\mathrm{logit}^{-1}(u_{a,2}^\top z(x) - d).$$

Let I_{d_z} denote the $d_z \times d_z$ identity matrix. For each repetition, the coefficient vectors are drawn once and then fixed within that repetition:

$$v_a \sim \mathcal{N}\big(0, (0.9^2/d_z) I_{d_z}\big), \quad r_a \sim \mathcal{N}\big(0, (1.0^2/d_z) I_{d_z}\big), \quad \theta_a \sim \mathcal{N}\big(0, (0.55^2/d_z) I_{d_z}\big),$$
$$q_a \sim \mathcal{N}\big(0, (0.75^2/d_z) I_{d_z}\big), \quad s_a \sim \mathcal{N}\big(0, (0.70^2/d_z) I_{d_z}\big), \quad u_{a,1}, u_{a,2} \sim \mathcal{N}\big(0, (0.65^2/d_z) I_{d_z}\big),$$

with independent entries and independent draws across treatment arms. We also draw $\lambda_a \sim \mathrm{Unif}(0.8, 1.5)$, $\rho_a \sim \mathrm{Unif}(-0.30, 0.30)$, and $\iota_a \sim \mathcal{N}(0, 0.12^2)$ independently across treatment arms. This setup requires methods to capture dose-dependent uncertainty and potential multimodality in addition to the mean dose-response surface. Finally, the observed outcome is given by $Y_i^{\mathrm{obs}} = Y_i(A_i, D_i)$.

Target design. The target design is the empirical test covariate distribution crossed with a uniform intervention over treatments and a fixed dosage grid:

$$Q_\rho = \hat{P}_{X,\mathrm{test}} \otimes \mathrm{Unif}\{1,2,3\} \otimes \mathrm{Unif}(\mathcal{D}), \quad \mathcal{D} = \{0, 0.05, \ldots, 1\}.$$

Equivalently, $\mathcal{D}$ consists of 21 equally spaced dosage values on $[0,1]$. Evaluation therefore averages over all treatment classes and all grid dosages.

H.1.3 Jobs

Source and sample construction. We use the National Supported Work randomized job-training experiment and Panel Study of Income Dynamics controls from the LaLonde benchmark [LaL86, DW99]. The NSW randomized sample contains 297 treated units and 425 control units. The PSID comparison group contains $2{,}490$ control units. The common feature set drops the PSID-specific $re74$ column and uses $re75$ as the pre-treatment earnings covariate.

Leakage-free splitting. To prevent data leakage between the observational training set and the randomized evaluation set, we first split each data source prior to constructing the training and evaluation sets. The NSW treated sample, NSW control sample, and PSID control sample are each split into training (0.56), validation (0.24), and test (remaining) fractions. The held-out RCT split consists solely of NSW treated and control units and is never used for training. The observational training split is formed by combining the corresponding training splits of NSW treated units, NSW control units, and PSID controls. The validation split is used exclusively for hyperparameter tuning.

Covariates and outcome. We use the common [LaL86] covariates

$$X = (\text{age}, \text{education}, \text{black}, \text{hispanic}, \text{married}, \text{nodegree}, re75).$$

Continuous covariates are standardized using training-set statistics, with age, education, and $re75$ treated as continuous variables. The outcome is post-treatment earnings $re78$. Because earnings have a point mass near zero and a heavy right tail, models are trained on the transformed outcome

$$Y = \operatorname{asinh}(re78/1000).$$

Figures and earnings-scale metrics use the inverse transform $re78 = 1000 \sinh(Y)$.

Target design. Let RCT denote the index set of the held-out randomized NSW sample, let $n_{\mathrm{RCT}} = |\mathrm{RCT}|$, and write $\hat{P}_{X,\mathrm{RCT}} = n_{\mathrm{RCT}}^{-1} \sum_{i \in \mathrm{RCT}} \delta_{x_i}$ for its empirical covariate law. The target design is the empirical covariate distribution of the held-out randomized NSW sample crossed with equal treatment-arm weights:

$$Q_\rho = \hat{P}_{X,\mathrm{RCT}} \otimes \mathrm{Unif}\{0, 1\}.$$

Since individual counterfactual laws are unavailable, evaluation compares model-implied arm-level interventional distributions with randomized NSW arm-level distributions.

H.2 Baselines

For IHDP and Jobs, we compare with GANITE [YJvdS18b], PO-Flow [WIX25], DiffPO [MMSF24], individualized normalizing flows (INFs) [MFF23], DR-Learner [Ken23], CoDiTE [PSSM21], DNet [WSL$^+$23], Causal-DRF [NPS26], and SKCD (adapted) [JL26]. GANITE is an adversarial counterfactual imputation method for binary treatments. PO-Flow and DiffPO are generative potential-outcome methods based on flow matching and diffusion models, respectively. INFs estimate interventional outcome densities with normalizing flows. DR-Learner is a doubly robust conditional mean method; it is included to assess the extent to which a strong point-estimation method can be extended to a distributional plug-in estimator. CoDiTE uses conditional kernel mean embeddings, DNet estimates noncrossing conditional quantiles, and Causal-DRF uses distributional forests to estimate conditional kernel treatment effects. [JL26] develop doubly robust estimation and testing, including the smoothed kernel conditional discrepancy (SKCD); our SKCD (adapted) row evaluates predictive distributions obtained through a sampling adapter applied to their cross-fitted outcome conditional mean embedding estimates.

For TCGA, we compare with SCIGAN [BJvdS20], DRNet [SLB$^+$20], and VCNet [NYLN21]. SCIGAN is the main adversarial baseline for continuous-valued treatments and dosages. DRNet and VCNet are strong neural dose-response baselines designed for continuous treatment-response estimation. All hyperparameters listed below are those used in the 100 repetitions.

For distributional evaluation, let B be the number of Monte Carlo samples drawn from each conditional law, with the benchmark-specific values given in Appendix H.3; generative baselines are evaluated using B such draws. For deterministic or mean-only baselines, including DR-Learner, DRNet, and VCNet, we construct a distributional plug-in baseline by adding an empirical residual distribution to the estimated conditional mean. For IHDP and Jobs, residuals are pooled within treatment arms. For TCGA, residuals are pooled within treatment–dosage strata. This construction gives mean-only methods access to a simple predictive distribution for distributional evaluation, while preserving their native mean estimates for point and policy metrics.

GANICE is trained with the stratified objective described in Section 3.2. For binary treatments, cells are treatment arms or low-dimensional treatment–covariate cells. For TCGA, cells are treatment–dosage strata at the selected dosage resolution, with covariates handled by the conditional generator. Mini-batches are sampled and normalized within cells. Cells with too few observations are

merged with nearby target-relevant cells at the resolution selected by validation. No propensity score, inverse-propensity weight, or density-ratio estimate is used by GANICE. All methods use the same train/validation/test splits within each repetition, and no method is tuned on test-set distributional metrics.

H.2.1 Hyperparameters for GANICE.

Table 7 reports the GANICE hyperparameters used in the main benchmark experiments. Here G/C hidden denotes the hidden-layer widths of the generator and critic. These runs use batch size 128, Adam optimizers with generator learning rate 2×10^{-4} and critic learning rate 10^{-4}, Adam betas $(0, 0.9)$, gradient-penalty weight 10, and outcome critics anchored at $y = 0$.

Raw generator outputs are bounded through a scaled sigmoid output layer, using intervals determined from training outcomes rather than hard clipping to the empirical training minimum and maximum. Writing $\hat{Q}_\alpha$ for the empirical training-outcome quantile on the modeling scale, the intervals are $[\hat{Q}_{0.01} - 0.75, \hat{Q}_{0.99} + 0.75]$ for IHDP, expanded symmetrically to width 1 if narrower; $[\hat{Q}_{0.002} - 0.4, \hat{Q}_{0.998} + 0.4]$ for TCGA; and $[0, \max\{1, \hat{Q}_{0.995} + 0.35\}]$ for Jobs.

The implemented generator loss consists of the generator component of the stratified adversarial objective in Section 3.2, with coefficient 1, a separate cellwise transport regularizer multiplied by the listed generator transport weight, and factual CRPS and squared-error auxiliary terms with the listed weights. The transport regularizer compares sorted observed and generated scalar outcomes within each minibatch cell, with generated outcomes evaluated at the sampled observational states. It aggregates the mean absolute differences using the cell frequencies in that minibatch; cells containing fewer than two sampled observations do not contribute. The factual auxiliary terms compare the predictive distribution with the observed outcome and the generated sample mean with the observed outcome, respectively, using 6 generated samples per observation. The factual squared-error term during adversarial training is disabled for Jobs.

Pretraining performs generator updates before adversarial training using only factual CRPS with coefficient 1 and sample-mean squared error with coefficient 2 for IHDP and TCGA, and 1.5 for Jobs. These pretraining losses also use 6 generated samples per observation.

The cell map lists the coordinates on which the dyadic partition of Section 3.2 is defined, the resolution gives the vector $\boldsymbol{m}$, and the target-mass Monte Carlo size is the number of draws from Q_ρ used to estimate the cell masses q_C.

Residual quantile calibration on IHDP is fitted separately within each active cell using training residuals relative to generated conditional sample means, with 12 generated samples per training observation and 256 quantile levels. At prediction time, calibrated and uncalibrated samples are blended with weights 0.75 and 0.25, respectively.

For Jobs, generated samples are first replaced by zero with an arm-specific probability equal to the empirical frequency of nonpositive earnings in the NSW training split, capped at 0.95. A monotone arm-level quantile map fitted on the validation RCT split using 256 quantile levels and 128 generated samples per validation covariate and treatment arm is then applied. This subsequent quantile map can change the injected zero mass, so the final zero-earnings frequency is not constrained to equal the training frequency. The Jobs output bounds and calibration operate on the transformed scale $Y = \operatorname{asinh}(re78/1000)$. Neither the IHDP residual calibration nor the Jobs quantile calibration is followed by clipping back to the raw generator bounds. Appendix F.5 discusses truncation bias associated with restricting outcome support.

H.2.2 Hyperparameters for IHDP and Jobs baselines.

Table 8 reports the baseline hyperparameters for the two binary-treatment benchmarks. All listed neural methods use batch size 128. The outcome bounds for bounded generators and heads are set from the corresponding training outcomes. The settings of CoDiTE, DNet, Causal-DRF, and SKCD (adapted) are shared across IHDP and Jobs unless stated otherwise. CoDiTE regularization and DNet training steps are selected separately within each repetition by minimizing validation extended Wasserstein error on IHDP and validation RCT-assisted Wasserstein error on Jobs. Candidate comparisons use the same initialization and evaluation random streams. This validation uses 96 Monte Carlo samples per conditional state on IHDP and 64 samples per observation and treat-

ment arm on Jobs. The LightGBM propensity model of SKCD (adapted) is tuned with Optuna using the tree-structured Parzen estimator over the following ranges: learning rate $[10^{-3}, 0.2]$ (log scale); number of leaves $\{7, \dots, 255\}$ (log scale); maximum depth $\{-1, 3, 5, 7, 10, 15\}$, where -1 denotes no depth limit; minimum child samples $\{2, \dots, 100\}$ (log scale); minimum split gain $[0, 1]$; feature and bagging fractions $[0.5, 1]$; bagging frequency $\{1, \dots, 10\}$; L1 and L2 regularization $[10^{-8}, 10]$ (log scale); and maximum bin count $\{63, 127, 255\}$. The selected propensity model is refitted on the full outer training fold using the selected boosting-round count.

H.2.3 Hyperparameters for TCGA baselines.

Table 9 reports the baseline hyperparameters for the continuous-dose TCGA benchmark. For VCNet, the discrete treatment indicator is appended to the covariates as a one-hot vector, and dosage is passed as the continuous treatment. For deterministic mean-response methods, distributional evaluation uses the residual plug-in construction described above.

H.3 Evaluation Metrics in Section 7

This subsection gives the exact definitions of the primary distributional metrics and figure diagnostics reported in Section 7. Each benchmark is evaluated over 100 repetitions. The Monte Carlo sample size is $B = 192$ per conditional law on IHDP and $B = 128$ on TCGA, with the same number of samples drawn from the true law and from the fitted model, and $B = 96$ per covariate vector and treatment arm on Jobs, where the comparison is with observed randomized outcomes. For a scalar probability law μ, write $F_\mu(y) = \mu((-\infty, y])$ and $Q_\mu(\alpha) = \inf\{y \in \mathbb{R} : F_\mu(y) \geq \alpha\}$ for its CDF and generalized quantile function, with $\alpha \in (0, 1)$. For an evaluation state s, μ_s^* and $\hat{\mu}_s$ denote the true and fitted conditional laws; their CDFs are F_s^* and $\hat{F}_s$, and their quantile functions are Q_s^* and $\hat{Q}_s$.

Empirical extended Wasserstein error for semi-synthetic benchmarks. For IHDP and TCGA, the true conditional interventional laws are known by construction. The primary metric is the empirical analogue of the extended Wasserstein risk analyzed in the theory: a target-design average of statewise Wasserstein distances, with comparisons made only at the same treatment–covariate state. Since outcomes are scalar, each W_1 distance is approximated using sorted Monte Carlo samples drawn from the true and fitted conditional laws. When writing distributional distances, we identify a univariate law with its CDF.

For IHDP, the evaluation states are (x_i, t) for $i = 1, \dots, n_{\text{test}}$ and $t \in \{0, 1\}$. For each state (x_i, t), we draw B samples from the true law in (35) and B samples from the fitted model. The empirical extended Wasserstein error is

$$\frac{1}{2n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} \sum_{t=0}^{1} W_1\big(\hat{\mu}_{x_i,t}, \mu^*_{x_i,t}\big).$$

For TCGA, the evaluation states are (x_i, a, d) for $i = 1, \dots, n_{\text{test}}$, $a \in \{1, 2, 3\}$, and $d \in \mathcal{D}$. The empirical extended Wasserstein error is

$$\frac{1}{3n_{\text{test}}|\mathcal{D}|} \sum_{i=1}^{n_{\text{test}}} \sum_{a=1}^{3} \sum_{d \in \mathcal{D}} W_1\big(\hat{\mu}_{x_i,a,d}, \mu^*_{x_i,a,d}\big).$$

Randomized-trial-assisted Wasserstein error for Jobs. For Jobs, individual counterfactual distributions are not observed. We therefore evaluate arm-level interventional distributions using the held-out randomized NSW sample. For $a \in \{0, 1\}$, let $n_{a,\text{RCT}} = \sum_{i \in \text{RCT}} \mathbf{1}\{T_i = a\}$ denote the size of its randomized treatment arm. For a fitted model, define the model-implied interventional CDF

$$\hat{F}_a(y) = \frac{1}{n_{\text{RCT}}} \sum_{i \in \text{RCT}} \hat{F}_{x_i,a}(y), \quad a \in \{0, 1\},$$

where the average is over covariates in the held-out randomized sample and each $\hat{F}_{x_i,a}$ is the empirical CDF of the B generated samples at (x_i, a). The randomized empirical arm-level CDF is

$$\hat{F}_a^{\text{RCT}}(y) = \frac{1}{n_{a,\text{RCT}}} \sum_{i \in \text{RCT}: T_i = a} \mathbf{1}\{Y_i^{\text{obs}} \leq y\}.$$

The quantile functions $\hat{Q}_a$ and $\hat{Q}_a^{\mathrm{RCT}}$ are the generalized inverses of $\hat{F}_a$ and $\hat{F}_a^{\mathrm{RCT}}$, respectively. The Jobs metric reported in Table 2 is

$$\text{RCT-}W_1 = \frac{1}{2}\sum_{a=0}^{1} W_1\Big(\hat{F}_a, \hat{F}_a^{\mathrm{RCT}}\Big).$$

This metric is not a pointwise counterfactual error; it measures whether the generated interventional distributions agree with randomized arm-level outcome distributions.

Figure 1 (a): IHDP quantile treatment-effect error. For each quantile level $\alpha \in \{0.05, 0.10, \ldots, 0.95\}$, define the true average quantile treatment effect (QTE)

$$\Delta_\alpha^* = \frac{1}{n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}}\big[Q_{x_i,1}^*(\alpha) - Q_{x_i,0}^*(\alpha)\big],$$

and the corresponding model estimate

$$\hat{\Delta}_\alpha = \frac{1}{n_{\mathrm{test}}}\sum_{i=1}^{n_{\mathrm{test}}}\big[\hat{Q}_{x_i,1}(\alpha) - \hat{Q}_{x_i,0}(\alpha)\big].$$

The quantile functions are computed from $4{,}096$ samples of the true law and 512 generated samples per state. The panel plots $|\hat{\Delta}_\alpha - \Delta_\alpha^*|$ as a function of α for the selected methods.

Figure 1 (b): TCGA dose-wise distributional diagnostics. Fix treatment class $a = 1$ and let $\mathcal{I}_{\mathrm{sub}}$ be the index set of the displayed held-out subgroup, selected as the test units with the largest predicted dose sensitivity; we compute model-implied dose-indexed conditional summaries from generated samples. The median curve is

$$d \mapsto \frac{1}{|\mathcal{I}_{\mathrm{sub}}|}\sum_{i\in\mathcal{I}_{\mathrm{sub}}} \hat{Q}_{x_i,a,d}(0.5),$$

with the corresponding true median computed from (36). The displayed predictive bands are constructed analogously from the generated $\alpha = 0.1$ and $\alpha = 0.9$ quantiles. All displayed quantiles are computed from 96 samples per state, drawn from the fitted model and from (36), respectively. This diagnostic distinguishes methods that track only the central dose-response trajectory from those that recover dose-dependent uncertainty.

Figure 1 (c): Jobs randomized CDF matching. We plot the randomized treated-arm empirical CDF $\hat{F}_1^{\mathrm{RCT}}$ and the model-implied treated-arm CDF $\hat{F}_1$ on the original earnings scale. The inverse transformation $re78 = 1000\sinh(Y)$ is applied before plotting. Analogous control-arm CDF diagnostics are reported in Appendix H.5.

H.4 Additional Evaluation Metrics

The main text focuses on distributional errors aligned with the causal target. Appendix H.5 reports additional diagnostics to verify that the empirical conclusions are not specific to the extended Wasserstein metric.

Notation for statewise and arm-level evaluation. Use μ_s^* and $\hat{\mu}_s$ for the true and fitted interventional laws associated with an evaluation index s. For IHDP, let $\mathcal{I} = \{1, \ldots, n_{\mathrm{test}}\} \times \{0, 1\}$, where $s = (i, t)$ refers to the state (x_i, t). For TCGA, let $\mathcal{I} = \{1, \ldots, n_{\mathrm{test}}\} \times \{1, 2, 3\} \times \mathcal{D}$, where $s = (i, a, d)$ refers to the state (x_i, a, d). For each index s, we draw $Y_{s,1}^*, \ldots, Y_{s,B}^* \sim \mu_s^*$ and $\hat{Y}_{s,1}, \ldots, \hat{Y}_{s,B} \sim \hat{\mu}_s$, with B as in Appendix H.3. The CDFs of μ_s^* and $\hat{\mu}_s$ are denoted by F_s^* and $\hat{F}_s$, and their quantile functions by $Q_s^*(\alpha)$ and $\hat{Q}_s(\alpha)$; numerical evaluation uses the Monte Carlo samples above.

For Jobs, individual counterfactual laws are unavailable. Distributional metrics are therefore evaluated at the randomized arm level, using the model-implied arm CDFs $\hat{F}_a$ and the held-out randomized NSW arm CDFs $\hat{F}_a^{\mathrm{RCT}}$ for $a \in \{0, 1\}$, as defined in Appendix H.3. Metrics reported on the earnings scale use the inverse transformation $re78 = 1000\sinh(Y)$.

Conventional point and policy metrics. We compute conventional causal metrics based on conditional means to check whether distributional gains come at the expense of standard scalar summaries. For IHDP, the precision in estimation of heterogeneous effect (PEHE) is

$$\mathrm{PEHE} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \big\{\hat{\tau}(x_i) - \tau(x_i)\big\}^2,$$

where $\hat{m}_t(x)$ is the estimated conditional mean under treatment t and $\hat{\tau}(x) = \hat{m}_1(x) - \hat{m}_0(x)$. We also compute the absolute average treatment-effect (ATE) error

$$\mathrm{ATEErr} = \left| \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \hat{\tau}(x_i) - \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \tau(x_i) \right|.$$

For TCGA, let $\hat{m}_a(x, d)$ be the estimated conditional mean response. The mean integrated squared error (MISE) is

$$\mathrm{MISE} = \frac{1}{3 n_{\mathrm{test}} |\mathcal{D}|} \sum_{i=1}^{n_{\mathrm{test}}} \sum_{a=1}^{3} \sum_{d \in \mathcal{D}} \big\{\hat{m}_a(x_i, d) - m_a(x_i, d)\big\}^2.$$

For each unit and treatment arm, define the model-selected dosage $\hat{d}_{i,a} \in \arg\max_{d \in \mathcal{D}} \hat{m}_a(x_i, d)$. The dosage policy error (DPE) is

$$\mathrm{DPE} = \frac{1}{3 n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \sum_{a=1}^{3} \big\{\max_{d \in \mathcal{D}} m_a(x_i, d) - m_a(x_i, \hat{d}_{i,a})\big\}^2.$$

For the full treatment–dosage decision, let

$$(\hat{a}_i, \hat{d}_i) \in \operatorname*{arg\,max}_{a \in \{1,2,3\},\ d \in \mathcal{D}} \hat{m}_a(x_i, d).$$

The policy error (PE) is

$$\mathrm{PE} = \frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \big[\max_{a \in \{1,2,3\},\ d \in \mathcal{D}} m_a(x_i, d) - m_{\hat{a}_i}(x_i, \hat{d}_i) \big].$$

For Jobs, $\widehat{\mathrm{ATT}}$ denotes the method-specific estimate of the average treatment effect on the treated, and $\mathrm{ATT}_{\mathrm{RCT}}$ denotes the reference estimate computed from the held-out randomized NSW arms; their absolute discrepancy is $\mathrm{ATTErr} = |\widehat{\mathrm{ATT}} - \mathrm{ATT}_{\mathrm{RCT}}|$. We also calculate an RCT-evaluated policy value. Let $\pi : \mathcal{X} \to \{0, 1\}$ be the treatment rule induced by a fitted method, write $\pi_i = \pi(x_i)$ on held-out randomized covariates, and let $\hat{p} = n_{\mathrm{RCT}}^{-1} \sum_{i \in \mathrm{RCT}} \pi_i$. The policy value on the original earnings scale is estimated by

$$\hat{V}(\pi) = \hat{p} \frac{\sum_{i \in \mathrm{RCT}} \mathbf{1}\{\pi_i = 1, T_i = 1\} re78_i}{\sum_{i \in \mathrm{RCT}} \mathbf{1}\{\pi_i = 1, T_i = 1\}} + (1 - \hat{p}) \frac{\sum_{i \in \mathrm{RCT}} \mathbf{1}\{\pi_i = 0, T_i = 0\} re78_i}{\sum_{i \in \mathrm{RCT}} \mathbf{1}\{\pi_i = 0, T_i = 0\}},$$

with the convention that empty matched arms are replaced by the corresponding randomized arm mean on the original earnings scale. For generative methods, conditional means are estimated from generated samples; for deterministic baselines, we use their native mean estimates.

Proper scoring rule. We compute the continuous ranked probability score (CRPS), a strictly proper scoring rule for univariate predictive distributions [GR07, DZ24]. For a predictive CDF F and observation y,

$$\mathrm{CRPS}(F, y) = \int_{\mathbb{R}} \{F(z) - \mathbf{1}\{y \le z\}\}^2 \mathrm{d}z = \mathbb{E}[|Y - y|] - \frac{1}{2}\mathbb{E}[|Y - Y'|], \quad Y, Y' \overset{\text{i.i.d.}}{\sim} F.$$

For semi-synthetic datasets, we compute the expected CRPS under the true interventional law:

$$\widehat{\mathrm{CRPS}} = \frac{1}{|\mathcal{I}|} \sum_{s \in \mathcal{I}} \frac{1}{B} \sum_{b=1}^{B} \mathrm{CRPS}(\hat{F}_s, Y^*_{s,b}).$$

For Jobs, we report factual CRPS on the held-out randomized NSW sample:

$$\widehat{\mathrm{CRPS}}_{\mathrm{fact}} = \frac{1}{n_{\mathrm{RCT}}} \sum_{i \in \mathrm{RCT}} \mathrm{CRPS}(\hat{F}_{x_i,T_i}, Y_i^{\mathrm{obs}}).$$

We additionally compute the same factual score on the original earnings scale:

$$\widehat{\mathrm{CRPS}}_{\mathrm{earn}} = \frac{1}{n_{\mathrm{RCT}}} \sum_{i \in \mathrm{RCT}} \mathrm{CRPS}(\hat{F}^{\$}_{x_i,T_i}, re78_i),$$

where $\hat{F}^{\$}_{x_i,T_i}$ is the generated predictive distribution after applying $re78 = 1000 \sinh(Y)$. CRPS is useful because it is not a Wasserstein distance, although factual CRPS is included as an auxiliary training loss for GANICE.

Energy distance. We compute energy distance (ED) as a sample-based distribution discrepancy with geometry different from extended Wasserstein risk [SR13]. For a semi-synthetic state s,

$$\mathrm{ED}(\hat{\mu}_s, \mu_s^*) = 2\mathbb{E}[\|\hat{Y} - Y^*\|] - \mathbb{E}[\|\hat{Y} - \hat{Y}'\|] - \mathbb{E}[\|Y^* - Y^{*\prime}\|],$$

where $\hat{Y}, \hat{Y}' \sim \hat{\mu}_s$ and $Y^*, Y^{*\prime} \sim \mu_s^*$ independently. A hat on ED denotes its sample-based approximation using the evaluation samples. The calculated semi-synthetic value is

$$\widehat{\mathrm{ED}} = \frac{1}{|\mathcal{I}|} \sum_{s \in \mathcal{I}} \widehat{\mathrm{ED}}(\hat{\mu}_s, \mu_s^*).$$

For Jobs, energy distance is computed at the randomized arm level on the original earnings scale and averaged over arms:

$$\widehat{\mathrm{ED}} = \frac{1}{2} \sum_{a=0}^{1} \widehat{\mathrm{ED}}\big(\hat{F}_a, \hat{F}_a^{\mathrm{RCT}}\big).$$

CDF discrepancy. For scalar outcomes, we compute the Kolmogorov–Smirnov discrepancy (KS)

$$\mathrm{KS} = \frac{1}{|\mathcal{I}|} \sum_{s \in \mathcal{I}} \sup_z |\hat{F}_s(z) - F_s^*(z)|.$$

For Jobs, the same discrepancy is computed at the randomized arm level:

$$\mathrm{KS} = \frac{1}{2} \sum_{a=0}^{1} \sup_z \big|\hat{F}_a(z) - \hat{F}_a^{\mathrm{RCT}}(z)\big|.$$

Quantile and quantile-effect errors. Let $\mathcal{A}_Q = \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$ denote the quantile levels. For semi-synthetic datasets, we compute the integrated quantile error (IQE)

$$\mathrm{IQE} = \left[\frac{1}{|\mathcal{I}||\mathcal{A}_Q|} \sum_{s \in \mathcal{I}} \sum_{\alpha \in \mathcal{A}_Q} \{\hat{Q}_s(\alpha) - Q_s^*(\alpha)\}^2 \right]^{1/2}.$$

For IHDP, we additionally compute quantile treatment-effect error (QTEErr):

$$\mathrm{QTEErr} = \left[\frac{1}{n_{\mathrm{test}}|\mathcal{A}_Q|} \sum_{i=1}^{n_{\mathrm{test}}} \sum_{\alpha \in \mathcal{A}_Q} \big\{ [\hat{Q}_{x_i,1}(\alpha) - \hat{Q}_{x_i,0}(\alpha)] - [Q^*_{x_i,1}(\alpha) - Q^*_{x_i,0}(\alpha)] \big\}^2 \right]^{1/2}.$$

For TCGA, the analogous dose-quantile error (DQErr) is

$$\mathrm{DQErr} = \left[\frac{1}{3n_{\mathrm{test}}|\mathcal{D}||\mathcal{A}_Q|} \sum_{i,a,d,\alpha} \{\hat{Q}_{x_i,a,d}(\alpha) - Q^*_{x_i,a,d}(\alpha)\}^2 \right]^{1/2}.$$

For Jobs, we compute both arm-level integrated quantile error and arm-level quantile treatment-effect error. The arm-level IQE is

$$\mathrm{IQE} = \left[\frac{1}{2|\mathcal{A}_Q|} \sum_{a=0}^{1} \sum_{\alpha \in \mathcal{A}_Q} \big\{\hat{Q}_a(\alpha) - \hat{Q}_a^{\mathrm{RCT}}(\alpha)\big\}^2 \right]^{1/2},$$

and the arm-level quantile treatment-effect error is

$$\mathrm{QTEErr} = \left[\frac{1}{|\mathcal{A}_Q|} \sum_{\alpha \in \mathcal{A}_Q} \big\{ [\hat{Q}_1(\alpha) - \hat{Q}_0(\alpha)] - [\hat{Q}_1^{\mathrm{RCT}}(\alpha) - \hat{Q}_0^{\mathrm{RCT}}(\alpha)] \big\}^2 \right]^{1/2}.$$

Tail functionals. Distributional causal inference is often motivated by tail risk rather than central tendencies [Fir07, KO23]. Let $\mathcal{A}_T = \{0.05, 0.10, 0.90, 0.95\}$ be the tail probability levels. For $\alpha \in \mathcal{A}_T$, we define the lower- and upper-tail means (LTM and UTM) as

$$\mathrm{LTM}_\alpha(\mu) = \mathbb{E}_\mu[Y \mid Y \le Q_\mu(\alpha)], \quad \mathrm{UTM}_\alpha(\mu) = \mathbb{E}_\mu[Y \mid Y \ge Q_\mu(\alpha)].$$

Here Y has law μ. A hat on LTM or UTM denotes a sample-based approximation of the corresponding functional using generated outcomes. For semi-synthetic datasets, we compute the tail error (TailErr) as

$$\mathrm{TailErr} = \frac{1}{|\mathcal{I}||\mathcal{A}_T|} \sum_{s \in \mathcal{I}} \sum_{\alpha \in \mathcal{A}_T} \Big(|\widehat{\mathrm{LTM}}_\alpha(\hat{\mu}_s) - \mathrm{LTM}_\alpha(\mu_s^*)| + |\widehat{\mathrm{UTM}}_\alpha(\hat{\mu}_s) - \mathrm{UTM}_\alpha(\mu_s^*)| \Big).$$

For Jobs, the same quantity is computed at the randomized arm level on the earnings scale and averaged across arms.

Calibration, interval width, and PIT diagnostics. For the nominal coverage levels $\mathcal{C} = \{0.50, 0.80, 0.90, 0.95\}$, we define the central predictive interval $I_s(c) = [\hat{Q}_s\{(1-c)/2\}, \hat{Q}_s\{(1+c)/2\}]$, where $\hat{Q}_s(\cdot)$ denotes the estimated quantile function. For semi-synthetic datasets, the empirical coverage is defined as

$$\widehat{\mathrm{Cov}}(c) = \frac{1}{|\mathcal{I}|B} \sum_{s \in \mathcal{I}} \sum_{b=1}^{B} \mathbf{1}\{Y_{s,b}^* \in I_s(c)\}.$$

The reported calibration error (CalErr) is

$$\mathrm{CalErr} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} |\widehat{\mathrm{Cov}}(c) - c|.$$

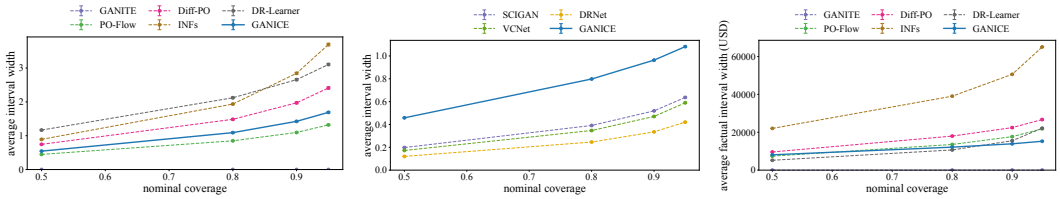
We also report the average interval width at each nominal level, which serves as a diagnostic of sharpness. For Jobs, calibration and interval width are evaluated on factual outcomes from the held-out randomized NSW sample. Finally, we report probability integral transform (PIT) histograms of $\hat{F}_s(Y_{s,b}^*)$ for semi-synthetic benchmarks and $\hat{F}_{x_i,T_i}(Y_i^{\mathrm{obs}})$ for Jobs.

H.5 Additional Results

This subsection reports additional empirical results on the three benchmarks used in Section 7. Each benchmark is evaluated over 100 repetitions. We focus on diagnostics complementary to those in Table 2, including non-Wasserstein distributional metrics, quantile and tail errors, calibration, randomized-arm CDF fits, and an objective ablation. Unless stated otherwise, reported values are means with standard errors computed over the same repetitions as in Table 2. Figures display a subset of methods for readability.

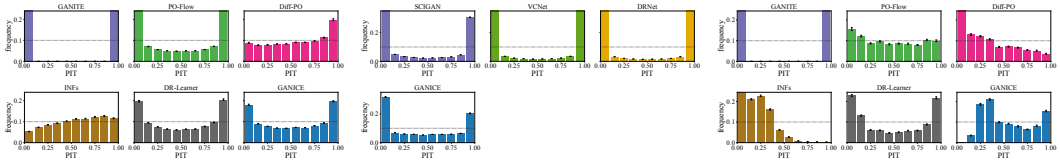
Robustness across distributional metrics. Tables 10–12 report the additional evaluation metrics introduced in Appendix H.4. The results show that the performance gains of GANICE are not merely a consequence of optimizing the extended Wasserstein objective. On TCGA, GANICE achieves the best performance across all reported distributional metrics, including CRPS, energy distance, KS distance, dose-quantile error, tail error, and calibration error. On Jobs, GANICE also performs best on all reported distributional and factual predictive metrics except calibration, while attaining the lowest ATT error against the held-out randomized NSW sample. On IHDP, Causal-DRF achieves the best distributional scores, followed by CoDiTE on several metrics, while GANICE attains the lowest PEHE and ATE error. Compared with INFs, GANICE performs substantially better on integrated quantile error, quantile treatment-effect error, and tail error, whereas INFs has lower CRPS, energy distance, KS distance, and calibration error. This distinction is important: a method may accurately capture marginal distributional structure while still failing to recover treatment-induced distributional contrasts.

Relation to scalar causal metrics. The additional metrics distinguish distributional causal learning from conventional mean-response or policy-oriented evaluation. On IHDP, GANICE attains the lowest PEHE and ATE error, while Causal-DRF performs best on the distributional functionals.



(a) IHDP (b) TCGA (c) Jobs

Figure 2: Predictive interval widths across nominal coverage levels. Widths should be interpreted together with calibration error: narrower intervals are preferable only when empirical coverage is maintained.



(a) IHDP (b) TCGA (c) Jobs

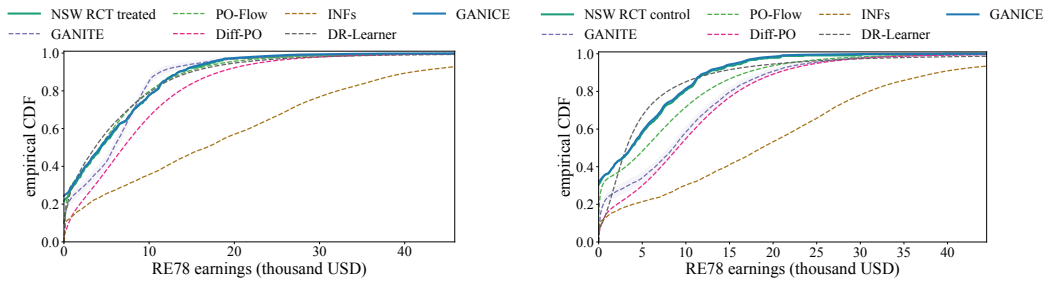
Figure 3: Probability integral transform diagnostics. For calibrated predictive distributions with continuous CDFs, PIT histograms should be close to uniform. For Jobs, PIT is evaluated factually on the held-out randomized NSW sample.

On TCGA, DRNet performs best on mean-response and policy metrics, which is expected because it is specifically optimized for scalar dose-response estimation; however, its distributional metrics are considerably worse. This shows the limits of point-estimation benchmarks for evaluating interventional distribution learning. On Jobs, DR-Learner achieves the highest policy value, whereas GANICE attains the best arm-level distributional fit, factual CRPS, QTE error, tail error, and ATT error. GANICE thus improves estimation of the full interventional distribution while remaining competitive on scalar causal summaries.

Calibration and sharpness. Tables 10–12 report calibration errors, while Figures 2 and 3 present interval sharpness and PIT diagnostics. GANICE achieves the lowest calibration error on TCGA and the second-lowest on Jobs, where SKCD (adapted) performs best. The interval-width analysis on TCGA shows that this improvement arises from capturing the broader dose-dependent uncertainty, rather than from producing overly narrow prediction intervals. On Jobs, GANICE maintains substantially narrower factual prediction intervals than INFs and DiffPO at high coverage levels. On IHDP, Causal-DRF has the lowest calibration error. Among the methods displayed in Figure 3, INFs exhibits the best PIT calibration on IHDP, but has larger quantile treatment-effect and tail errors than GANICE. These results support the main conclusion: calibration alone is insufficient for causal distribution learning, because the estimand is a target-design family of interventional conditional distributions and their distributional contrasts.

Randomized arm-level distributional fits on Jobs. Figure 4 compares the generated treated- and control-arm CDFs with the empirical CDFs from the held-out randomized NSW sample. The treated-arm panel complements Figure 1, while the control-arm panel verifies that the improvement is not specific to a single treatment arm. Across both arms, GANICE closely follows the randomized empirical distribution throughout the main body of the earnings distribution and avoids the large tail discrepancies observed for INFs. PO-Flow and DR-Learner are competitive in parts of the treated-arm curve, but Table 12 shows that GANICE is better under integrated CDF, quantile, tail, and Wasserstein-type discrepancies.

Objective ablation. Figure 5 evaluates the role of the proposed finite-resolution causal objective using IHDP. Replacing the statewise objective with a pooled WGAN-style objective substantially increases the empirical extended Wasserstein error, indicating that ordinary joint distribution matching can obscure statewise causal discrepancies. The full GANICE objective and the no-cell-



(a) Treated arm

(b) Control arm

Figure 4: Jobs randomized-arm CDF diagnostics. The panels compare model-implied interventional arm CDFs with held-out randomized NSW empirical CDFs on the earnings scale.

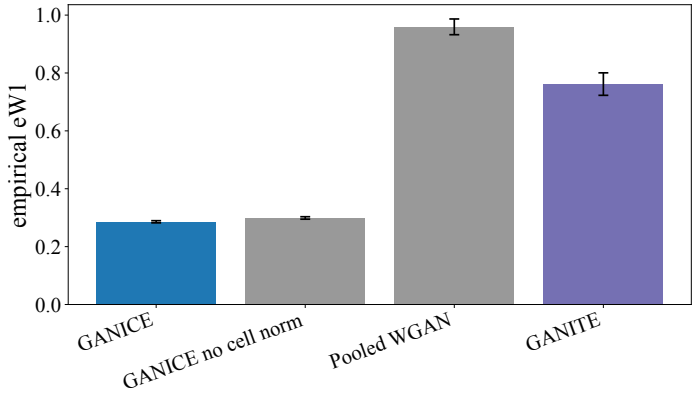


Figure 5: Objective ablation on IHDP. The full method is compared with variants that remove cell normalization or use a pooled WGAN-style objective. Lower empirical extended Wasserstein error is better.

normalization variant perform similarly on this binary-treatment benchmark, although the full objective performs slightly better and is theoretically aligned with the target-design aggregation analyzed above. The comparison with GANITE further suggests that adversarial imputation on factual coordinates alone is insufficient for accurate interventional distribution learning. The ablation thus supports the two central design choices of GANICE: statewise extended-Wasserstein comparison and stratified target-design aggregation.

H.6 Implementation Diagnostics

The main benchmark experiments use one shared conditional generator. We compare it with a routed implementation, which fits one outcome generator per cell, on paired repetitions with the same split and evaluation target (10 for IHDP, 10 for Jobs, and 5 for TCGA).

Estimating the transfer discrepancy. Let $\hat{g}$ be the shared generator and $\hat{\nu}_C$ the fitted law in cell C of the routed reference. Draw $B = 2{,}048$ independent samples $Y_{C,1}, \dots, Y_{C,B}$ from $\nu^{\rho}_{\hat{g},C}$ and an independent sample $Z_{C,1}, \dots, Z_{C,B}$ from $\hat{\nu}_C$. Sampling the cell mixture uses $W \sim Q_\rho(\cdot \mid C)$ and independent latent noise. For scalar outcomes, sorting these samples gives the empirical Wasserstein distance:

$$\hat{\varepsilon} = \sum_C q_C \frac{1}{B} \sum_{b=1}^{B} \left| Y_{C,(b)} - Z_{C,(b)} \right|.$$

Here parenthesized subscripts indicate order statistics. The reported $\hat{\varepsilon}$ averages five independent replications of this computation. Two independent samples from the same routed law provide a Monte Carlo self-distance. The measured discrepancies exceeded this reference on all benchmarks.

As a sanity check, substituting the routed experts themselves for the shared generator in this computation gives zero discrepancy on every cell retained by the minimum-count rule. For gated models using the same experts, the gate discrepancy η in Proposition 7 gives another computable bound.

Routing yields lower mean error on IHDP, similar error on Jobs, and higher error on TCGA. For Jobs, $\hat{\varepsilon}$ compares cell distributions before arm-level calibration, whereas the primary metric compares calibrated arm distributions. These diagnostics measure finite-sample discrepancies and do not establish the asymptotic transfer condition.

H.7 Sensitivity Analysis

We use a fully synthetic finite-state experiment to examine overlap, sample size, target-design shift, cell resolution, covariate dimension, and treatment–covariate heterogeneity. The covariates are $Z \sim \mathrm{Unif}(\{0,1\}^d)$, with $d \geq 2$, and the conditioning state is $W = (Z, T)$. For binary treatment, $\mathbb{P}(T = 1 \mid Z = z)$ takes the values $\kappa/2$ and $1 - \kappa/2$ on the two parity classes of $\sum_j z_j$, with $\kappa \in (0, 1]$. Define $c_1(z) = (-1)^{\sum_{j=1}^d z_j}$ and $c_2(z) = (-1)^{\sum_{j=1}^d j z_j}$. The functions m_t and θ_t determine the conditional mean and distributional shape, with heterogeneity parameter $h \in [0, 1]$:

$$m_t(z) = 0.4c_1(z) + 0.2c_2(z) + t\big\{0.8 + 0.35h\, c_1(z)\big\},$$
$$\theta_t(z) = \frac{\pi}{12} + \frac{\pi}{4} z_1 z_2 + \frac{\pi}{12} ht\big\{1 + c_2(z)\big\}.$$

Since $z \in \{0,1\}^d$, the product $z_1 z_2$ equals $\mathbf{1}\{z_1 = z_2 = 1\}$. With independent $U_1, U_2 \sim \mathrm{Unif}([0, 1])$, independent of (Z, T), the potential outcomes are

$$Y(t) = m_t(Z) + 0.45\Big\{\sqrt{2}\cos(\theta_t(Z))\sin(2\pi U_1) + \frac{4}{\sqrt{5}}\sin(\theta_t(Z))\sin^3(2\pi U_2)\Big\}.$$

This generator is bounded and smooth in the latent variables. For $H = \{z \in \{0,1\}^d : z_1 = z_2 = 1\}$ and $\delta \in [0, 1]$, the target design is

$$Q_\rho = \big\{(1-\delta)\mathrm{Unif}(\{0,1\}^d) + \delta \mathrm{Unif}(H)\big\} \otimes \mathrm{Unif}(\{0,1\}).$$

The target cell masses are known exactly, and a valid overlap constant for this design is $\kappa/(1 + 3\delta)$. The nominal configuration is $(n, d, \kappa, \delta, h) = (2048, 4, 1, 0, 0.5)$ with 32 cells. We examine each factor separately, refit GANICE, and average the statewise W_1 errors over all (z, t) using their target masses. Each statewise W_1 error is computed from $1{,}024$ samples of the true law and $1{,}024$ generated samples. Sample-size comparisons use nested prefixes of one observational sample; target-design comparisons use the same observations.

The errors change little over the tested overlap, sample-size, and cell-resolution ranges, suggesting an approximation or optimization floor at these sample sizes. Errors increase with covariate dimension and heterogeneity. At $\delta = 0.9$, the error is 0.426 of the nominal value under the corresponding target weighting.

Table 4: Notation for distances, partitions, sieves, and rates

Symbol	Description
eW_1	Extended Wasserstein-1 distance on $\mathcal{W} \times \mathcal{Y}$ for laws sharing the same $\mathcal{W}$-marginal; equivalently, the average of statewise W_1 distances.
$\mathrm{eW}_1^{(\boldsymbol{m})}$	Finite-resolution discrepancy induced by $\mathcal{F}^{(\boldsymbol{m})}$; it equals the target-mass-weighted sum of cellwise W_1 distances.
$\Gamma(\eta, \zeta)$, $\Gamma_\Delta(P, R)$	Couplings with marginals η, ζ, and the subset of couplings between P, R that preserve the conditioning state, respectively.
$\mathcal{L}_1$	Anchored outcome critic class $\{f : f(y_0) = 0,\ \mathrm{Lip}(f) \le 1\}$.
$\mathcal{F}$	Exact extended-Wasserstein dual class; each section $h(w, \cdot)$ belongs to $\mathcal{L}_1$.
$\boldsymbol{m}, \Pi_{\boldsymbol{m}}$	Resolution vector $\boldsymbol{m} = (m_1, \dots, m_{d_W}) \in \mathbb{N}_0^{d_W}$ and its anisotropic dyadic partition, with side length 2^{-m_j} in coordinate j.
$\boldsymbol{m}_n$	Optimized resolution vector with $m_{n,j} = \lfloor r \log_2(\kappa n)/\alpha_j \rfloor \vee 0$; the generic IPM result uses $r_{\mathcal{V}}$ in place of r.
$\lvert\boldsymbol{m}\rvert_1$	Total resolution $\sum_{j=1}^{d_W} m_j$.
C, q_C, π_C	A cell C, its target mass $q_C = Q_\rho(C)$, and observational mass $\pi_C = Q_{\mathrm{obs}}(C)$.
$C_{\boldsymbol{m}}(w)$	Cell of $\Pi_{\boldsymbol{m}}$ containing w.
N_C^{obs}, M_C^ρ	Realized observational count and target-design Monte Carlo count in cell C.
$\mu_C^{\mathrm{obs}}, \mu_C^\rho, \nu_{g,C}^\rho$	Outcome cell mixtures under Q_{obs} and Q_ρ, and generated cell mixture under Q_ρ, respectively, normalized by cell masses; μ_C^ρ uses the kernel of P, with $\mu_w = \mu_w^*$ when $P = P^*$.
$P_g^{(\boldsymbol{m})}$	Cell-resolution generated law $Q_\rho(\mathrm{d}w)\nu^\rho_{g,C_{\boldsymbol{m}}(w)}(\mathrm{d}y)$.
$\mathcal{F}^{(\boldsymbol{m})}$	Finite-resolution critic class, piecewise constant in w and outcome-Lipschitz within each cell.
k, ℓ	Numbers of observed outcomes and independent latent draws in a one-state estimation problem.
$N, \mathcal{G}_N, \mathcal{D}_N$	Deterministic sieve index $N \ge 2$ and finite one-state generator and critic classes; the oracle bound applies whenever $k, \ell \ge c_* N$ for a fixed $c_* > 0$.
$N_{\boldsymbol{m}}$	Deterministic effective cell sample size $\lceil \kappa n 2^{-\lvert\boldsymbol{m}\rvert_1} \rceil \vee 2$, used as the one-state sieve index in every cell at resolution $\boldsymbol{m}$, giving critic class $\mathcal{D}_{N_{\boldsymbol{m}}}$.
$\mathcal{G}_n^{\mathrm{cont}}, \mathcal{H}_{n,\boldsymbol{m}}^{\mathrm{cont}}$	Implemented continuous-conditioning generator class and finite-resolution critic class.
$\hat{P}_{n,\boldsymbol{m}}^{\mathrm{cont}}$	Cell-resolution estimate $P^{(\boldsymbol{m})}_{\hat{g}_{n,\boldsymbol{m}}^{\mathrm{cont}}}$.
$\mathcal{H}_{K_0}^s$	Periodic Hölder ball of maps $\mathbb{T}^{d_U} \to \mathbb{R}^p$ with image in $B^p(0, K_0)$ and smoothness s.
β	Smoothness parameter of the one-state latent generator class $\mathcal{H}_{K_0}^{\beta+1}$.
a	Local one-state WGAN rate exponent $a = \min\{(\beta+1)/(2\beta + d_U), 1/2\}$.
$\boldsymbol{\alpha}$	Anisotropic conditional regularity vector $(\alpha_1, \dots, \alpha_{d_W})$.
$b_{\boldsymbol{m}}$	Localization bias scale $\sum_j 2^{-\alpha_j m_j}$.
$\bar{d}$	Effective anisotropic conditioning dimension $\sum_j \alpha_j^{-1}$.
r	Continuous-conditioning rate exponent $(a^{-1} + \bar{d})^{-1}$.
$p_B, \boldsymbol{s}$	Integrability and smoothness indices in the anisotropic Besov–Nikolskii critic embedding.
$\mathcal{M}_{\mathrm{disc}}, \mathcal{M}_{\mathrm{cont}}$	Finite-state and continuous-conditioning causal model classes.
$\mathcal{R}_n^{\mathrm{disc}}, \mathcal{R}_n^{\mathrm{cont}}$	Minimax risks in finite-state and continuous-conditioning settings.

Table 5: Datasets used in the experiments.

Dataset	Size	Treatment	Evaluation target
IHDP	$n = 747$, 25 covariates	binary	known $\{\mu^*_{x,t}\}_{t=0,1}$
TCGA	$n = 9{,}659$, 4,000 features	3 treatments + dosage	known $\{\mu^*_{x,a,d}\}$ on a dosage grid
Jobs	NSW $297/425$, PSID $2{,}490$	binary	randomized arm-level CDFs

Table 6: Data sources and licenses.

Dataset	Source	License / terms
IHDP	Supplementary materials of [Hil11]	Not specified in the supplied supplementary files.
TCGA	`https://paperdatasets.s3.amazonaws.com/tcga.db`	Not explicitly specified for the processed benchmark database.
	TCGA Research Network / Genomic Data Commons: `https://portal.gdc.cancer.gov`	Governed by NIH/GDC open-data access policies.
Jobs	`https://users.nber.org/~rdehejia/data/nsw_treated.txt` `https://users.nber.org/~rdehejia/data/nsw_control.txt` `https://users.nber.org/~rdehejia/data/psid_controls.txt`	CC BY-NC 2.0.

Notes. For IHDP, we found no explicit license statement in the supplied supplementary files of [Hil11]. For TCGA, the first row reports the processed benchmark database used in continuous-dose causal learning; the second row records the underlying TCGA/GDC data-access regime [WCM+13]. The MIT license of the Perfect Match code repository covers the code and does not extend to `tcga.db`.

Table 7: GANICE hyperparameters used in the experiments.

Dataset	Architecture and optimization	Finite-resolution and auxiliary settings
IHDP	latent dim. 4; G/C hidden $(128, 128)$; 520 adversarial steps; 1 critic step per generator step; shared conditional generator; 800 pretraining steps.	Cell map: first two training-set PCA coordinates plus treatment; resolution $(1, 1, 1)$; minimum cell size 6; target-mass Monte Carlo size $50{,}000$. Validation eW_1 selects among four random restarts and the grid of generator transport weights $\{0.75, 1.25\}$, factual CRPS weights $\{8, 12\}$, and factual MSE weights $\{1, 2\}$. Residual quantile calibration uses 12 samples per observation, grid size 256, and blend 0.75.
TCGA	latent dim. 4; G/C hidden $(96, 96)$; 620 adversarial steps; 2 critic steps per generator step; shared conditional generator; $1{,}600$ pretraining steps.	Covariate resolution 0; treatment resolution 2; dosage resolution 3; minimum cell size 4; target-mass Monte Carlo size $45{,}000$. Generator transport weight 0.6; factual CRPS weight 10 with 6 samples; factual MSE weight 8 with 6 samples.
Jobs	latent dim. 4; G/C hidden $(96, 96)$; 3 critic steps per generator step; shared conditional generator; 420 pretraining steps.	Cell map: standardized $re75$, Black indicator, and treatment; resolution $(1, 1, 1)$; minimum cell size 4; target-mass Monte Carlo size $40{,}000$. Validation selects among $(180, 220, 260)$ adversarial steps, generator transport weights in $\{4, 6\}$, and factual CRPS weights in $\{4, 5\}$. The Jobs implementation additionally matches the zero-earnings mass using the NSW training split and fits arm-level quantile calibration on the validation RCT split.

Table 8: Baseline hyperparameters for IHDP and Jobs.

Dataset	Method	Hyperparameters
IHDP	GANITE	hidden dim. 96; counterfactual generator iterations 500; ITE inference iterations 500; counterfactual discriminator steps 2; ITE discriminator steps 1; reconstruction weight 2; ITE weight 5; learning rate 10^{-3}.
IHDP	PO-Flow	hidden dim. 64; training steps 750; learning rate 10^{-3}; weight decay 10^{-5}; RK4 integration steps 20.
IHDP	DiffPO	hidden dim. 64; time-embedding dim. 128; residual blocks 4; propensity steps 220; diffusion steps 350; diffusion time steps 45; denoiser learning rate 5×10^{-4}; propensity learning rate 10^{-3}; weight decay 10^{-5}.
IHDP	INFs	hidden dim. 64; outcome bins 64; nuisance steps 700; target-flow steps 700; nuisance learning rate 10^{-3}; target learning rate 4×10^{-3}; weight decay 10^{-5}; propensity-loss weight 1; propensity clipping 0.05; outcome noise std. 0.01.
IHDP	DR-Learner	hidden dim. 64; 2-fold cross-fitting; nuisance steps 600; final CATE steps 750; nuisance and final learning rates 10^{-3}; weight decay 10^{-5}; propensity clipping 0.05.
Jobs	GANITE	hidden dim. 128; counterfactual generator iterations 650; ITE inference iterations 650; counterfactual discriminator steps 2; ITE discriminator steps 1; reconstruction weight 1; ITE weight 5; learning rate 10^{-3}.
Jobs	PO-Flow	hidden dim. 64; training steps 850; learning rate 10^{-3}; weight decay 10^{-5}; RK4 integration steps 18.
Jobs	DiffPO	hidden dim. 64; time-embedding dim. 128; residual blocks 3; propensity steps 260; diffusion steps 520; diffusion time steps 55; denoiser learning rate 5×10^{-4}; propensity learning rate 10^{-3}; weight decay 10^{-5}.
Jobs	INFs	hidden dim. 64; outcome bins 72; nuisance steps 700; target-flow steps 700; nuisance learning rate 10^{-3}; target learning rate 4×10^{-3}; weight decay 10^{-5}; propensity-loss weight 1; propensity clipping 0.05; outcome noise std. 0.01.
Jobs	DR-Learner	hidden dim. 64; 2-fold cross-fitting; nuisance steps 650; final CATE steps 800; nuisance and final learning rates 10^{-3}; weight decay 10^{-5}; propensity clipping 0.05.
IHDP, Jobs	CoDiTE	Gaussian covariate and outcome kernels; arm-specific covariate bandwidths and a pooled outcome bandwidth selected by the median-distance heuristic; bandwidth computation uses at most 20,000 pair distances; common arm regularization $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}\}$ selected by validation; arm-specific kernel systems $K_a + n_a \lambda I$; all training observations retained; predictive sampling uses outcome-RKHS projection onto the empirical probability simplex, with at most 256 iterations and tolerance 5×10^{-4}.
IHDP, Jobs	DNet	shared hidden layers $(200, 100)$; treatment-specific outcome hidden layer (100); ELU activations; 50 noncrossing quantiles at levels $k/51$, $k = 1, \ldots, 50$; Adam learning rate 10^{-3}; training steps selected from $\{750, 1000, 1250\}$; quantile-Huber threshold 1; propensity-loss weight 1; treatment-tower kernel L2 coefficient 10^{-5}; initial quantile increment 0.02; unbounded quantile heads; predictive sampling assigns equal mass to the estimated quantiles.
IHDP, Jobs	Causal-DRF	2,500 trees in 50 groups on IHDP and 4,000 trees in 40 groups on Jobs; independent Bernoulli$(1/2)$ group half-samples; tree subsampling fraction 0.8 within each group, without replacement; random $50/50$ honest splits; minimum leaf size per treatment arm 4; maximum depth 12; 5 candidate covariates per split on IHDP and 2 on Jobs (the integer part of the square root of the number of covariates); split-balance parameter 0.2; all admissible thresholds evaluated; Gaussian outcome-kernel bandwidth from the full-sample all-pairs median heuristic; splitting uses 32 random Fourier frequencies with sine and cosine features.
IHDP, Jobs	SKCD (adapted)	2-fold stratified cross-fitting; Gaussian covariate and outcome kernels with full-sample all-pairs median bandwidths; outcome-embedding systems $K + 10^{-3} I$; LightGBM propensity estimation with 30 Optuna trials per outer training fold; internal stratified $80/20$ split; binary log-loss selection; maximum 2,000 boosting rounds; 10-round early stopping; propensity clipping 10^{-6}; predictive sampling equally mixes the two fold-specific outcome nuisance distributions after Euclidean simplex projection of their signed embedding coefficients.

Table 9: Baseline hyperparameters for TCGA.

Method	Hyperparameters
SCIGAN	noise dim. 8; hidden dim. 64; set dim. 16; batch size 128; GAN iterations 700; inference-network iterations 1,000; dosage samples per treatment 5; factual reconstruction weight 1; learning rate 10^{-3}.
VCNet	hidden dim. 64; spline degree 2; spline knots $(1/3, 2/3)$; treatment-density grid size 10; batch size 128; training steps 1,100; learning rate 10^{-3}; weight decay 10^{-4}; density-loss weight 0.2; targeted regularization disabled; outcome standardization enabled.
DRNet	hidden dim. 64; dosage strata 5; base layers 2; treatment-specific layers 1; dosage-head layers 2; repeated dosage input enabled; batch size 128; training steps 1,100; learning rate 10^{-3}; weight decay 10^{-4}; outcome standardization enabled.

Table 10: Additional evaluation metrics on IHDP. Lower is better for all columns. Bold and underline for PEHE follow the unrounded means (0.049561 for GANICE, 0.049718 for Causal-DRF).

Method	CRPS	ED	KS	IQE	QTEErr	TailErr	CalErr	PEHE	ATEErr
GANITE	0.769 (0.039)	0.899 (0.075)	0.790 (0.010)	0.986 (0.038)	0.675 (0.021)	1.840 (0.069)	0.787 (0.000)	0.403 (0.031)	0.309 (0.025)
PO-Flow	0.469 (0.007)	0.298 (0.008)	0.396 (0.003)	0.595 (0.010)	0.729 (0.015)	0.988 (0.016)	0.312 (0.004)	0.457 (0.024)	0.132 (0.011)
DiffPO	0.391 (0.004)	0.142 (0.005)	0.261 (0.003)	0.451 (0.007)	0.688 (0.011)	0.748 (0.011)	0.077 (0.004)	0.345 (0.013)	0.536 (0.012)
INFs	0.381 (0.004)	0.124 (0.003)	0.237 (0.003)	0.569 (0.007)	0.713 (0.009)	0.963 (0.013)	0.028 (0.001)	0.279 (0.010)	0.466 (0.010)
DR-Learner	0.864 (0.008)	1.089 (0.013)	0.534 (0.003)	1.300 (0.012)	2.116 (0.021)	1.907 (0.019)	0.174 (0.003)	4.504 (0.095)	0.380 (0.023)
CoDiTE	0.347 (0.004)	0.055 (0.001)	0.207 (0.001)	0.389 (0.008)	0.501 (0.011)	0.561 (0.009)	0.024 (0.001)	0.060 (0.002)	0.085 (0.006)
DNet	0.642 (0.009)	0.644 (0.013)	0.616 (0.003)	0.873 (0.014)	1.075 (0.020)	1.491 (0.019)	0.571 (0.002)	1.070 (0.043)	0.268 (0.020)
Causal-DRF	**0.346** (0.003)	**0.053** (0.001)	**0.194** (0.002)	**0.287** (0.004)	**0.375** (0.005)	**0.486** (0.006)	**0.020** (0.001)	0.050 (0.002)	0.078 (0.006)
SKCD (adapted)	0.372 (0.004)	0.105 (0.001)	0.309 (0.001)	0.433 (0.007)	0.604 (0.010)	0.669 (0.008)	0.088 (0.002)	0.118 (0.004)	0.084 (0.006)
GANICE	0.382 (0.004)	0.125 (0.003)	0.288 (0.003)	0.389 (0.005)	0.399 (0.006)	0.693 (0.009)	0.164 (0.003)	**0.050** (0.002)	**0.074** (0.006)

Table 11: Additional evaluation metrics on TCGA. Lower is better for all columns.

Method	CRPS	ED	KS	DQErr	TailErr	CalErr	MISE	DPE	PE
SCIGAN	0.420 (0.001)	0.306 (0.003)	0.503 (0.001)	0.558 (0.002)	0.901 (0.004)	0.510 (0.001)	0.155 (0.002)	0.212 (0.004)	0.450 (0.005)
VCNet	0.570 (0.004)	0.606 (0.009)	0.617 (0.002)	0.788 (0.007)	1.226 (0.009)	0.585 (0.002)	0.448 (0.010)	0.134 (0.005)	0.399 (0.011)
DRNet	0.439 (0.001)	0.344 (0.002)	0.559 (0.001)	0.600 (0.002)	0.980 (0.003)	0.605 (0.002)	**0.149** (0.002)	**0.065** (0.003)	**0.176** (0.003)
GANICE	**0.409** (0.003)	**0.284** (0.007)	**0.382** (0.003)	**0.532** (0.007)	**0.820** (0.009)	**0.310** (0.002)	0.225 (0.007)	0.213 (0.005)	0.542 (0.010)

Table 12: Additional evaluation metrics on Jobs. Distributional metrics compare interventional arm distributions with held-out randomized NSW arm distributions; both CRPS columns and calibration use factual outcomes. Factual CRPS uses the transformed scale, and CRPS earn. uses the earnings scale. Lower is better except for policy value.

Method	Factual CRPS	CRPS earn.	ED	KS	IQE	QTEErr	TailErr	CalErr	ATTErr	Policy value
GANITE	1.463 (0.021)	7291 (338)	4093 (562)	0.491 (0.016)	6122 (399)	5128 (310)	10547 (733)	0.787 (0.000)	3979 (283)	5330 (74)
PO-Flow	0.801 (0.004)	3552 (26)	337 (22)	0.176 (0.004)	2729 (93)	2966 (131)	5015 (176)	0.068 (0.004)	2147 (122)	5779 (84)
DiffPO	0.892 (0.007)	3998 (35)	1504 (70)	0.281 (0.005)	4927 (162)	2864 (127)	8833 (333)	0.121 (0.006)	2418 (123)	5409 (73)
INFs	1.195 (0.007)	8198 (30)	9688 (93)	0.525 (0.003)	19171 (94)	2753 (87)	37232 (234)	0.159 (0.003)	1574 (98)	5670 (79)
DR-Learner	0.859 (0.004)	3769 (37)	335 (14)	0.208 (0.003)	2637 (123)	2481 (153)	7187 (268)	0.200 (0.003)	1377 (117)	**5946** (76)
CoDiTE	0.839 (0.004)	3832 (24)	315 (16)	0.160 (0.003)	2492 (75)	3003 (133)	6231 (198)	0.067 (0.002)	2577 (141)	5647 (80)
DNet	0.844 (0.005)	3649 (29)	296 (16)	0.222 (0.004)	1907 (78)	2586 (148)	3490 (175)	0.246 (0.005)	1731 (119)	5735 (84)
Causal-DRF	0.871 (0.004)	4565 (23)	2270 (40)	0.258 (0.003)	8620 (69)	11932 (166)	12873 (162)	0.078 (0.002)	9169 (142)	5196 (61)
SKCD (adapted)	0.854 (0.004)	3979 (25)	640 (28)	0.194 (0.004)	3495 (80)	4674 (168)	6281 (190)	**0.047** (0.002)	3836 (155)	5471 (62)
GANICE	**0.769** (0.003)	**3408** (29)	**167** (10)	**0.127** (0.003)	**1644** (60)	**2107** (96)	**3210** (149)	0.058 (0.002)	**1068** (85)	5753 (81)

Table 13: Paired implementation diagnostics, with standard errors in parentheses and paired repetition counts beside benchmark names. Primary errors use the metrics in Table 2. The ratio is $\hat{\varepsilon}$ divided by the shared primary error, using the displayed means.

Benchmark	Shared error	Routed error	$\hat{\varepsilon}$	Ratio
IHDP (10)	0.292 (0.012)	0.233 (0.009)	0.155 (0.012)	0.53
Jobs (10)	0.181 (0.017)	0.177 (0.020)	0.531 (0.049)	2.93
TCGA (5)	0.413 (0.015)	0.526 (0.032)	0.258 (0.034)	0.62

Table 14: Finite-state sensitivity analysis of GANICE. Entries are relative errors with the nominal configuration normalized to 1. Each setting uses three optimization seeds. Reported standard errors of the ratios are below 0.01, except for $d = 3$, where the standard error is 0.03.

Factor	Settings	Relative error
Overlap	$\kappa = 1, 0.5, 0.25$	$1.000, 1.005, 1.010$
Sample size	$n = 2048, 1024, 768, 512$	$1.000, 0.996, 1.001, 1.002$
Target-design shift	$\delta = 0, 0.45, 0.9$	$1.000, 0.888, 0.426$
Cell resolution	$\lvert\Pi\rvert = 2, 4, 8, 16, 32$	$0.998, 1.000, 0.996, 0.997, 1.000$
Covariate dimension	$d = 2, 3, 4, 5, 6$	$0.267, 0.753, 1.000, 1.026, 1.030$
Heterogeneity	$h = 0, 0.25, 0.5, 0.75, 1$	$0.820, 0.910, 1.000, 1.091, 1.183$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims stated in the abstract and introduction accurately reflect the contributions and scope presented in the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the proposed approach in the second paragraph of Section 8.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We have provided the full set of assumptions, statements, and complete proofs in the main text and in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We briefly describe the experimental setup in Section 7, and provide all necessary information to reproduce the results in Appendix H.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide the code in the supplementary materials, and all datasets used in the paper (IHDP, TCGA, and Jobs) are publicly available and open access.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We have described the experimental setting and training details in Appendix H.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We run each experiment multiple times with different random seeds and report the standard error computed from these independent replications to reflect the variability across runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Although the paper does not provide detailed compute specifications, our experiments use relatively small datasets and can be run on a standard laptop or desktop, requiring only modest computational resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have checked the NeurIPS Code of Ethics and conducted all aspects of this research in accordance with its guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

Justification: We consider a theoretical analysis and methodological development in causal inference using GANs. As our work focuses on theoretical properties and does not introduce new datasets or pre-trained models, it does not have direct societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: We do not release new datasets or pre-trained models, and our work focuses on theoretical properties; therefore, the results do not present any significant risk of misuse that would require additional safeguards.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We properly cite the papers or URLs for the datasets used in our experiments. Table 6 reports the available license or access-term information for each dataset.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not introduce any new assets in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We do not conduct any crowdsourcing experiments or research involving human subjects in this paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

Justification: We do not conduct any research involving human subjects in this work, so there are no associated risks that would require IRB approval.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: We have used an LLM only for drafting, editing, formatting, and refactoring code.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.