NEIGHBORHOOD-GUIDED FLOW MATCHING FOR POTENTIAL OUTCOME DISTRIBUTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent work in causal inference has increasingly focused on modeling conditional distributions of potential outcomes, which capture outcome variability beyond conditional means. Balanced representation learning aims to mitigate selection bias in observational data, but representation alignment alone does not ensure that fine outcome variations are preserved. To address this challenge, we propose a unified framework that couples representation balancing with neighborhood-guided flow matching. Our method encourages alignment between treatment groups in latent space and uses observed outcomes from covariate neighbors in the target treatment arm as flow-matching endpoints, while conditioning the velocity field on the individual's learned representation and the target treatment. By weighting neighborhood supervision according to local treatment support, the model borrows outcome information where suitable neighboring observations are available. Under standard identification and smoothness assumptions, we show that neighborhood supervision targets a local mixture of conditional potential outcome distributions and bound its deviation from the true distribution. Experiments demonstrate the effectiveness of our method in conditional distribution recovery on synthetic data and treatment-effect estimation on standard causal benchmarks.

1 INTRODUCTION

Learning how an intervention changes an outcome distribution is central to decisions that depend on more than an average response. Two treatments can have similar expected outcomes while differing in variability, tail risk, or the probability of an adverse event. Such differences matter when using observational evidence to inform treatment choices or targeted interventions (Pearl, 2009; Hernán & Robins, 2020; Rzepakowski & Jaroszewicz, 2012). We study the conditional potential-outcome laws $P_a(\cdot \mid x) = \mathcal{L}(Y(a) \mid X = x)$, which retain this distributional information and also determine conditional mean effects (Gische & Voelkle, 2022; Kennedy et al., 2023).

Estimating these laws presents a causal learning problem: each unit supplies an outcome only under the treatment it received (Rubin, 1974). Moreover, observational treatment assignment depends on covariates, so the two treatment groups need not cover the covariate space equally (Rosenbaum & Rubin, 1983). Under conditional exchangeability and positivity, each potential-outcome law is identified from its corresponding factual arm. In finite samples, however, a treatment can be rare near a particular query even when positivity holds. The generator must then recover a conditional distribution from limited local evidence. This difficulty concerns not only the location of the outcome but also its spread and shape.

Existing work provides two useful approaches to this problem. Representation learning makes outcome prediction more tractable and can regularize discrepancies between treatment groups (Johansson et al., 2016; Shalit et al., 2017; Assaad et al., 2021). Generative outcome models represent distributions beyond conditional means, using latent variables, conditional CDFs, normalizing flows, diffusion, or flow matching (Louizos et al., 2017; Zhou et al., 2022; Vanderschueren et al., 2023; Ma et al., 2024; Wu et al., 2026). Recent orthogonal generative learners additionally address sensitivity to estimated nuisance functions (Melnychuk & Feuerriegel, 2026). These developments provide flexible models and principled training objectives for distributional causal estimation.

Yet reducing a global treatment-group discrepancy does not supply the local outcome information needed to recover a missing-arm distribution. An alignment penalty compares representation dis-

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

tributions; it does not specify whether the target outcome law at a query is narrow, heavy-tailed, or bimodal. Meanwhile, a factual generative loss receives endpoints only from the observed arm. When target-arm observations are locally scarce, globally similar representations can therefore coexist with a poorly determined conditional outcome shape. Figure 1(a) illustrates the issue: a single-peaked extrapolation can reproduce a mean while missing two separated modes. Work on overlap and representation sufficiency shows why preserving outcome-relevant information must accompany distributional adjustment (Zhang et al., 2020; Shi et al., 2024; Melnychuk et al., 2026b; Clivio et al., 2026).

This motivates our central question: *how can a conditional generator use nearby target-arm observations to learn a locally faithful outcome distribution, and when should it trust that information?* The key observation is that neighboring factual outcomes carry distributional information when their conditional laws vary smoothly in the chosen metric. Their usefulness depends on two distinct properties: how close the opposite treatment arm is relative to the surrounding data, and how probability is distributed among the available outcomes. Borrowing an average outcome discards the latter information. Retaining a weighted collection of observed endpoints preserves the empirical variation that the generator needs to learn.

We introduce *neighborhood-guided flow matching* (NGFM). A factual warm-up learns a representation whose frozen copy defines opposite-arm neighborhoods. NGFM then supplements factual flow matching and representation alignment with paths ending at every selected donor's observed outcome, conditioned on the query covariates. A query-level reliability weight reduces borrowing when opposite-arm neighbors are distant relative to the pooled local neighborhood; distance weights allocate mass among donors. The learned metric also reduces the cost of comparing high-dimensional covariates: once encoded into d_z dimensions, each distance costs $O(d_z)$ instead of $O(d_x)$ in the original input space. This is particularly useful when $d_z \ll d_x$. Our matched original-X ablation gives comparable effect accuracy, supporting the representation metric as an efficient alternative (Section 5.3). The trainable representation and flow continue to learn against this fixed graph. At prediction time, the fitted flow generates samples without retrieving neighbors.

Our contributions connect this construction to its statistical target and empirical behavior. First, we formulate local distributional supervision through a weighted empirical endpoint law, separating the strength of borrowing from the allocation of donor mass. Second, we characterize the induced endpoint mixture and bound its deviation from the target law through weighted donor distance, effective donor count, and flow approximation error. Third, we evaluate distribution recovery in three simulation families and HC-MNIST, followed by effect estimation on IHDP and ACIC 2018. Density comparisons, matched component ablations, and parameter sweeps examine the roles of local endpoints, representation alignment, and donor reliability.

2 Related work

Treatment effects and representation learning. Propensity weighting and doubly robust estimation adjust for observed confounding (Lunceford & Davidian, 2004; Chernozhukov et al., 2018). Meta-learners, causal forests, and neural estimators extend this task to heterogeneous mean effects (Athey & Imbens, 2015; Künzel et al., 2019; Wager & Athey, 2018; Kennedy, 2023; Shi et al., 2019). Representation-based approaches learn predictive features while controlling treatment-group discrepancy or separating covariate information (Johansson et al., 2016; Shalit et al., 2017; Hassanpour & Greiner, 2019; Assaad et al., 2021). Optimal transport and integral probability metrics provide related adjustment criteria (Wang et al., 2023; Kong et al., 2023). Our objective retains representation alignment, but adds observed local outcomes as distributional training targets.

Generative potential-outcome models. CEVAE and GANITE use latent-variable and adversarial models for causal prediction (Louizos et al., 2017; Yoon et al., 2018). FCCN estimates conditional CDFs, and NOFLITE uses normalizing flows (Zhou et al., 2022; Vanderschueren et al., 2023). DiffPO introduces propensity-adjusted diffusion training, whereas PO-Flow uses flow matching for potential-outcome generation (Ma et al., 2024; Wu et al., 2026). In DiffPO, the observed-arm propensity appears in the denominator of the loss weight. Small propensities can therefore amplify gradients; clipping stabilizes training but changes the weighting and can introduce bias, potentially reducing accuracy. More generally, a flexible generator does not itself resolve a shortage of local target-arm

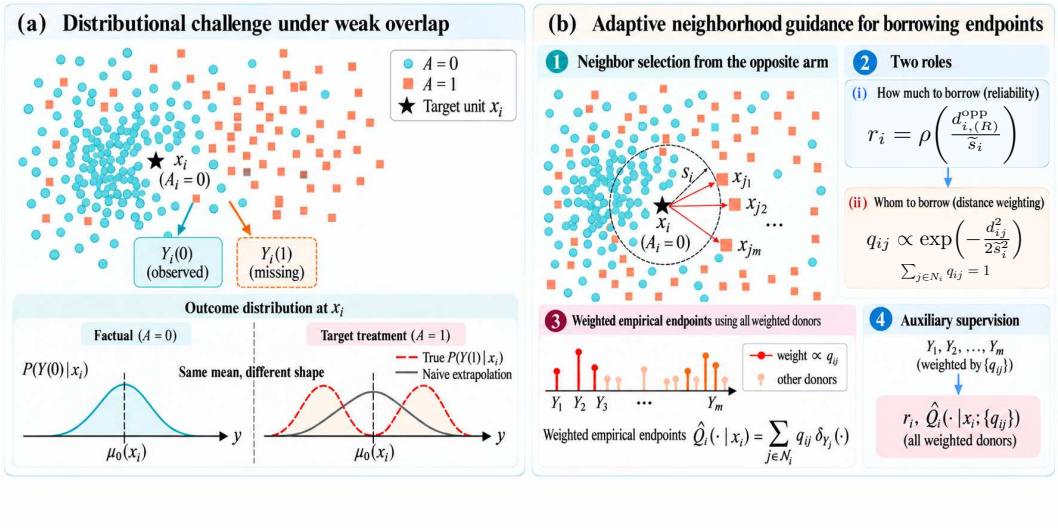

Figure 1: **Local distribution recovery under weak treatment support.** (a) The observed arm can be well represented near a query while the other arm is sparse; matching a conditional mean alone does not recover distributional shape. (b) NGFM retains nearby opposite-arm outcomes as weighted flow endpoints. Query reliability controls their total influence, and distance weights distribute mass among donors. Here $\rho(u) = \exp\{-\frac{1}{2}(u-1)_+^2\}$ and the opposite-arm distance is evaluated at rank R.

observations. GDR-learners construct orthogonal generative risks with double-robustness properties (Melnychuk & Feuerriegel, 2026). Marginal interventional densities, conditional distributional functionals, and bounds on treatment-effect distributions answer complementary causal questions (Kennedy et al., 2023; Melnychuk et al., 2023; Kallus & Oprescu, 2023; Melnychuk et al., 2024a). NGFM changes the endpoint supervision of a conditional flow: its auxiliary target is a weighted local outcome law, with an explicit approximation error.

Overlap, neighborhoods, and information preservation. DKLITE regularizes counterfactual posterior variance, CrossNet learns sufficient representations through cross-group constraints, and OR-learners study representation learning with orthogonal estimation (Zhang et al., 2020; Shi et al., 2024; Melnychuk et al., 2026b). Representation-induced confounding and deconfounding scores clarify which information a reduction must preserve (Melnychuk et al., 2024b; Clivio et al., 2026). OAR adapts mean-effect regularization to overlap (Melnychuk et al., 2026a). PairNet uses observed outcomes of nearby cross-treatment pairs to train effect differences (Nagalapatti et al., 2024). NGFM instead retains those outcomes as atoms of an endpoint distribution and adapts their total training mass to relative cross-arm distance. Flow matching supplies the velocity-regression machinery (Lipman et al., 2023; Tong et al., 2024); the present analysis concerns what that machinery learns from locally borrowed endpoints.

3 Neighborhood-Guided Flow Matching

3.1 Causal Target and Factual Supervision

We observe independent units $\{(X_i, A_i, Y_i)\}_{i=1}^n$, where $A_i \in \{0,1\}$ and $Y_i = Y_i(A_i)$. Our targets are $P_a(\cdot \mid x) = \mathcal{L}(Y(a) \mid X = x)$ for $a \in \{0,1\}$. We assume consistency, conditional exchangeability $Y(a) \perp A \mid X$, and positivity $0 < P(A = 1 \mid X = x) < 1$ on the target population. These conditions identify P_a with $\mathcal{L}(Y \mid X = x, A = a)$. Local treatment scarcity is a finite-sample difficulty within this identified setting.

NGFM combines a shared encoder Φ_ϕ with a treatment-conditioned velocity field $v_\theta(u, t, \Phi_\phi(x), a)$. For an endpoint y, Gaussian source ϵ, and uniform path time t, conditional flow matching uses the path $u_t = (1-t)\epsilon + ty$ and target velocity $y - \epsilon$ (Lipman et al., 2023; Tong et al., 2024). Let $\ell_a(x, y)$

denote its expected squared velocity error. Factual training minimizes $\mathcal{L}_{\text{fact}} = n^{-1} \sum_i \ell_{A_i}(X_i, Y_i)$. The flow is a continuous generative model (Chen et al., 2018): integrating its learned velocity from independent Gaussian draws produces samples from each estimated potential-outcome law.

Factual training determines where endpoint supervision is available. A query observed under $A_i = 0$ contributes a control path, but no treated path at X_i. Representation alignment encourages shared structure across arms; it does not add a treated outcome at that query. NGFM adds such supervision by borrowing observed treated endpoints from a local neighborhood, with a weight that reflects how well the neighborhood supports the query.

3.2 Local support and donor probabilities

Learning a stable metric. We first train the encoder and velocity on factual paths. We then freeze a copy of the encoder and standardize its outputs using training-set moments. Euclidean distance in this space, denoted d_{ij}, defines the K nearest opposite-arm donors N_i. The live encoder remains trainable. Fixing the metric after warm-up gives neighborhoods an outcome-predictive geometry and keeps the borrowed targets stable during subsequent optimization.

How much to borrow. Let s_i be the distance to the $\min(K, n-1)$th nearest training unit pooled across both arms, excluding the query itself, and let $d^{\text{opp}}_{i,(R)}$ be the distance to the Rth opposite-arm neighbor, with $1 \le R \le K$. Set $\widetilde{s}_i = \max(s_i, \varepsilon_s)$. Query reliability is

$$r_i = \exp\left[-\frac{1}{2}\left(\frac{d^{\text{opp}}_{i,(R)}}{\widetilde{s}_i} - 1\right)_+^2\right]. \tag{1}$$

This compares cross-arm availability with ordinary local data density. If R opposite-arm units fit inside the pooled radius, $r_i = 1$. As their distance exceeds that radius, the additional supervision decreases. The reference radius is pooled because normalizing by another opposite-arm distance would obscure the separation between treatments. If fewer than R opposite-arm donors exist, $r_i = 0$.

Which outcomes to emphasize. Within N_i, weights $q_{ij} \propto \exp\{-d_{ij}^2/(2\widetilde{s}_i^2)\}$ sum to one. They define the empirical endpoint law

$$\widehat{Q}_i = \sum_{j \in N_i} q_{ij} \delta_{Y_j}. \tag{2}$$

The distinction between r_i and q_{ij} is essential. Reliability controls the total influence of a neighborhood; donor probabilities preserve and weight the outcomes within it. A distant donor pool can thus receive little influence even though its probabilities sum to one. A well-supported pool with outcomes near two modes retains both sets of endpoints.

3.3 Distributional guidance and the full objective

Each donor supplies its observed Y_j to a path conditioned on the *query* X_i and treatment $1 - A_i$. NGFM evaluates every selected donor loss:

$$\mathcal{L}_{\text{ad}} = \frac{1}{n} \sum_i r_i \sum_{j \in N_i} q_{ij} \ell_{1-A_i}(X_i, Y_j), \qquad \mathcal{L} = \mathcal{L}_{\text{fact}} + \beta \mathcal{L}_{\text{bal}} + \mathcal{L}_{\text{ad}}. \tag{3}$$

The balance term $\mathcal{L}_{\text{bal}}$ is the debiased Sinkhorn divergence between the two minibatch representation measures (Feydy et al., 2019). It regularizes the live representation used by the flow, while $\mathcal{L}_{\text{ad}}$ adds local outcome information. The graph and all donor weights stay fixed during this stage; consequently, changing β in joint training does not recompute the donor metric.

Keeping individual endpoints is consequential for distribution learning. Replacing $\widehat{Q}_i$ by a point mass at $\sum_j q_{ij} Y_j$ preserves its mean but removes its empirical spread and modes. Equation (3) retains the whole weighted law. Evaluating the donor sum also integrates out donor-identity sampling; path times and Gaussian sources remain Monte Carlo draws.

We select the checkpoint by factual validation CRPS, a proper distributional score (Gneiting & Raftery, 2007). At test time, each treatment head transports Gaussian source particles to outcome

samples. Sample means give conditional mean outcomes and their treatment contrast; empirical laws give distributional summaries. Prediction requires only the encoder and flow. Appendix D provides the algorithm, detailed training diagram, path equations, and implementation settings.

4 WHAT DISTRIBUTION DOES NEIGHBORHOOD GUIDANCE LEARN?

The analysis follows the training construction: identify the endpoint law, quantify its local approximation error, and transfer that error to the generated distribution. We consider scalar outcomes with finite second moments. Detailed statements and proofs are in Appendices A and B.

The empirical objective has a distributional target. For a fixed training sample and graph, $\mathcal{L}_{\text{fact}} + \mathcal{L}_{\text{ad}}$ is exactly flow matching against the measure containing factual atoms (X_i, A_i, Y_i) and borrowed atoms $(X_i, 1 - A_i, Y_j)$ with masses proportional to $r_i q_{ij}$. Squared-loss projection gives the oracle velocity as the conditional expectation of $Y - \epsilon$ along the corresponding interpolating paths. Thus all-donor supervision learns from $\widehat{Q}_i$, whereas supervision at the donor mean learns from a point mass. This identity holds for the implemented outcome-trained graph.

A population characterization additionally separates the donor bank from the queries defining the risk. Write $e_a(x) = P(A = a \mid X = x)$, and let $r_a(x)$ and $\widehat{Q}_a(\cdot \mid x)$ denote reliability and the empirical donor law for target arm a. Set $r_a(x) = 0$ when the neighborhood is unavailable.

Proposition 1 (Endpoint mixture)**.** *For a donor bank independent of a fresh query, the normalized endpoint law of $\mathcal{L}_{\text{fact}} + \mathcal{L}_{\text{ad}}$ at (x, a), before representation pooling, is*

$$T_a = (1 - \alpha_a)P_a + \alpha_a \widehat{Q}_a, \qquad \alpha_a(x) = \frac{e_{1-a}(x) r_a(x)}{e_a(x) + e_{1-a}(x) r_a(x)}. \tag{4}$$

For scalar outcomes, $W_1(P_a, T_a) = \alpha_a W_1(P_a, \widehat{Q}_a)$.

The propensity in this expression records the frequency of factual and borrowed paths; NGFM does not estimate it. The formula identifies the role of reliability precisely: for fixed treatment probabilities and donor law, reducing r_a attenuates donor mismatch. In a very scarce arm, even a small reliability can still correspond to appreciable borrowed mass because factual mass is also small.

Locality and effective sample size govern donor error. Let $\mathcal{G}$ contain donor covariates, arm labels, and a metric fitted independently of donor outcomes. Suppose that, conditional on $\mathcal{G}$, the donor outcomes are independent with laws $P_a(\cdot \mid X_j)$. For a nonempty donor set, define the weighted radius $\bar{d}_a(x) = \sum_j q_j d(x, X_j)$ and effective donor count $K_{\text{eff}} = (\sum_j q_j^2)^{-1}$.

Theorem 1 (Conditional donor approximation)**.** *Assume local regularity $W_1(P_a(\cdot \mid X_j), P_a(\cdot \mid x)) \le L_a d(x, X_j)$. If outcomes lie in a common interval of length D, then*

$$\mathbb{E}[W_1(\widehat{Q}_a, P_a) \mid \mathcal{G}] \le L_a \bar{d}_a(x) + \frac{D}{2\sqrt{K_{\text{eff}}}}. \tag{5}$$

For unbounded outcomes with a uniform conditional qth-moment bound M_q about a common center, $q > 2$, replace $D/2$ by $H_q = \frac{q}{q-2}(4M_q)^{1/q}$.

The first term measures how much the conditional law changes across the neighborhood. The second measures the noise in its weighted empirical approximation. Increasing K can reduce sampling noise while admitting more distant donors; concentrating weights can reduce the radius while lowering K_{eff}. This tradeoff explains why a larger donor pool need not improve distribution recovery. The moment version covers Gaussian and Laplace outcomes under the stated uniform moment condition.

From donor endpoints to generated outcomes. Let $\mathcal{E}_{a,x}$ be the population excess velocity risk for fitting paths from T_a, and let $\nu_{a,x}$ be the numerical integration error in W_1. Under the state-Lipschitz and integrability conditions in Appendix A, flow stability and Theorem 1 give

$$\mathbb{E}[W_1(P_a, \widehat{P}_{\theta,a}) \mid \mathcal{G}] \le \underbrace{\alpha_a \left(L_a \bar{d}_a + \frac{H}{\sqrt{K_{\text{eff}}}} \right)}_{\text{borrowed-endpoint error}} + \underbrace{\mathbb{E}[\sqrt{C_L \mathcal{E}_{a,x}} + \nu_{a,x} \mid \mathcal{G}]}_{\text{flow approximation and integration}}, \tag{6}$$

where $H = D/2$ or H_q and C_L depends on the velocity's Lipschitz envelope. The residual includes representation approximation; Appendix G discusses the scope of these guarantees.

Corollary 1 (Recovery and mean-effect consequences)**.** *Under the preceding assumptions, vanishing $\bar{d}_a$, diverging K_{eff}, and vanishing expected flow and integration residual imply vanishing conditional mean W_1 error, for bounded L_a and H. For integrable estimated and true laws,*

$$|\widehat{\tau}(x) - \tau(x)| \leq W_1(\widehat{P}_1, P_1) + W_1(\widehat{P}_0, P_0). \tag{7}$$

Averaging the right-hand side bounds absolute ATE error; its $L^2(P_X)$ norm bounds rPEHE.

These conclusions connect distribution learning to causal summaries while retaining the distinction between them. Conditional means can be accurate even when outcome modes are poorly recovered. We therefore evaluate both full distributions and mean effects.

5 Experiments

The experiments examine three consequences of neighborhood guidance: recovery of conditional distributional shape, accuracy of the resulting mean effects, and the contribution of each source of supervision. Simulations provide known conditional laws, allowing distribution recovery to be evaluated directly. Semi-synthetic benchmarks extend the analysis to observed tabular and image covariates.

Settings, metrics, and comparison protocol. Setting I combines local treatment imbalance with bimodal outcomes. Setting II has global treatment scarcity and arm-specific noise. Setting III follows Melnychuk et al. (2026b). Setting IV uses HC-MNIST with $\Gamma = 1$, giving conditional exchangeability; Appendix F.3 adds a $\Gamma = e$ confounding test. IHDP and ACIC 2018 provide effect-estimation benchmarks (Hill, 2011; Shimoni et al., 2018). Simulation baselines are CEVAE, DKLITE, FCCN, NOFLITE, GDR-CNF, PO-Flow, and DiffPO.

Conditional Wasserstein-1 distance (cW1) averages distributional error over test queries and both treatment arms. Root PEHE (rPEHE) measures effect error against the supplied benchmark truth, and absolute ATE error measures the error in the average effect. Lower values are better. Methods receive the same train, validation, and test observations within each repetition; NGFM checkpoints in Settings I–IV are selected by factual validation CRPS. Saved train/validation/test allocations are approximately 70%/15%/15% or 70%/20%/10% for tabular data; exact counts and truth definitions are in Appendix C. Tables report means and sample SDs across repeated experiments. Simulation results use five independently generated datasets, while benchmark results use repeated fits with shared partitions across methods.

5.1 Recovery of potential-outcome distributions

Table 1 shows that NGFM has the lowest cW1 in Settings I, III, and IV. Setting I contains two outcome modes in scarce treatment cells, so its conditional mean alone does not determine the outcome law. On HC-MNIST with $\Gamma = 1$, NGFM has the lowest cW1 (0.748 ± 0.016) and rPEHE (1.224 ± 0.044); GDR-CNF has the lowest ATE error.

Setting II isolates global treatment scarcity. FCCN obtains the lowest cW1, 0.100 ± 0.013, while NGFM and GDR-CNF obtain 0.118 ± 0.017 and 0.121 ± 0.011. The contrast across settings is informative: a smaller arm does not necessarily create the same local distributional challenge as treatment-dependent support coupled with multimodality. Mean-effect rankings also differ. DKLITE has the smallest rPEHE in Settings I and II, GDR-CNF in Setting III, and NGFM in Setting IV. These results motivate inspecting the recovered densities alongside effect errors.

Figure 2 examines two low-support conditional laws. NGFM recovers their separated peaks and low-density valley. This is the feature that a mean-only assessment cannot reveal: placing mass between two modes can leave the mean nearly unchanged. The donor objective supplies outcomes on both sides of the valley as separate endpoints, consistent with the empirical-law characterization in Section 4.

Figure 3 presents a related sample-size study; Setting I uses $p = 0.2$, compared with $p = 0.15$ in Table 1. Setting I error decreases as more observations become available, whereas Setting III

Table 1: **Distribution and effect recovery.** Settings I–III use 2,000 observations; Setting IV is HC-MNIST with $\Gamma = 1$. Entries are mean $\pm$ SD over five runs. rPEHE is the square root of PEHE in every setting. Bold marks the smallest mean in each setting and metric. Dashes indicate unavailable entries.

	Setting I			Setting II		
Method	cW1	rPEHE	ATE error	cW1	rPEHE	ATE error
CEVAE	0.343±0.039	0.141±0.029	**0.042±0.032**	0.183±0.012	0.114±0.027	0.075±0.021
DKLITE	0.425±0.002	**0.106±0.037**	0.043±0.042	0.226±0.010	**0.081±0.030**	0.047±0.030
FCCN	0.270±0.016	0.176±0.078	0.051±0.020	**0.100±0.013**	0.170±0.027	0.079±0.035
NOFLITE	0.399±0.042	0.178±0.079	0.150±0.103	0.194±0.027	0.217±0.079	0.075±0.063
GDR-CNF	0.243±0.062	0.228±0.028	0.113±0.105	0.121±0.011	0.083±0.026	**0.032±0.022**
PO-Flow	0.399±0.081	0.155±0.056	0.096±0.069	0.125±0.016	0.142±0.019	0.051±0.031
DiffPO	2.114±1.268	1.810±1.348	1.590±1.451	2.697±0.873	2.075±0.806	2.003±0.863
NGFM (ours)	**0.227±0.020**	0.267±0.037	0.071±0.025	0.118±0.017	0.151±0.042	0.084±0.051
	Setting III			**Setting IV**		
Method	cW1	rPEHE	ATE error	cW1	rPEHE	ATE error
CEVAE	1.154±0.050	1.890±0.145	0.175±0.154	1.355±0.041	2.674±0.130	0.120±0.064
DKLITE	1.818±0.035	1.402±0.155	0.149±0.100	1.354±0.003	1.812±0.032	0.062±0.017
FCCN	1.026±0.123	1.853±0.131	0.327±0.192	1.323±0.013	1.962±0.079	0.342±0.178
NOFLITE	1.134±0.048	1.572±0.140	0.225±0.126	1.687±0.006	3.002±0.016	0.176±0.101
GDR-CNF	0.406±0.064	**0.868±0.119**	0.108±0.099	1.094±0.024	2.110±0.042	**0.034±0.015**
PO-Flow	0.658±0.165	1.068±0.201	0.235±0.157	0.838±0.038	1.628±0.163	0.187±0.051
DiffPO	10.200±3.940	10.320±5.070	9.447±5.178	0.973±0.181	1.742±0.303	0.222±0.127
NGFM (ours)	**0.389±0.047**	0.905±0.160	**0.094±0.104**	**0.748±0.016**	**1.224±0.044**	0.127±0.118

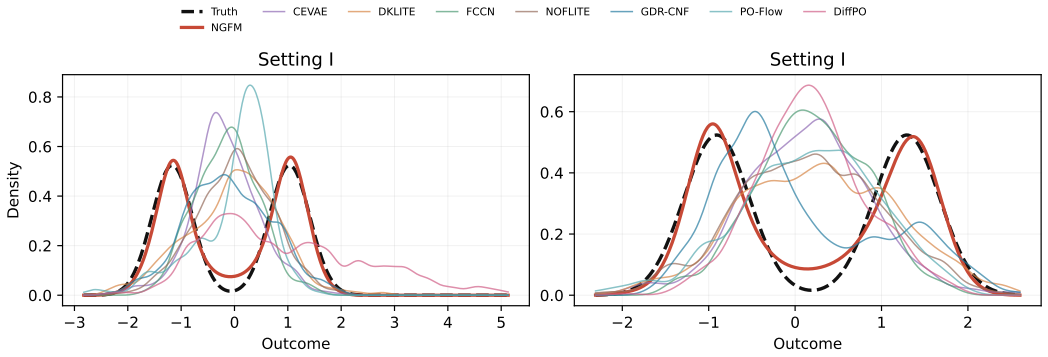

Figure 2: **Recovering outcome shape in low-support regions.** Two Setting I query–arm pairs. The dashed black curve smooths quantiles of the true conditional law; colored curves smooth generated samples using the same bandwidth. NGFM retains the separation between modes.

remains more difficult at smaller samples. These curves complement the fixed-size table by showing how each generator uses additional data. Appendix F adds rPEHE curves, the complete Setting III tables, and further conditional-density comparisons. The interval widths also reveal variation across independently generated datasets, which is substantial in some small-sample comparisons.

5.2 Treatment-effect estimation

Table 2 compares effect estimates from generated conditional means and direct outcome regressors. NGFM has the lowest reported rPEHE on IHDP, 0.330 ± 0.066, while CFR has the smallest ATE error. On ACIC 2018, NGFM has the lowest values in both columns: rPEHE 0.409 ± 0.172 and ATE error 0.399 ± 0.169. Thus local distributional supervision supports accurate mean contrasts. The different IHDP rankings reflect the distinction in Corollary 1: ATE error permits cancellation across units, whereas rPEHE summarizes root mean squared unit-level effect error.

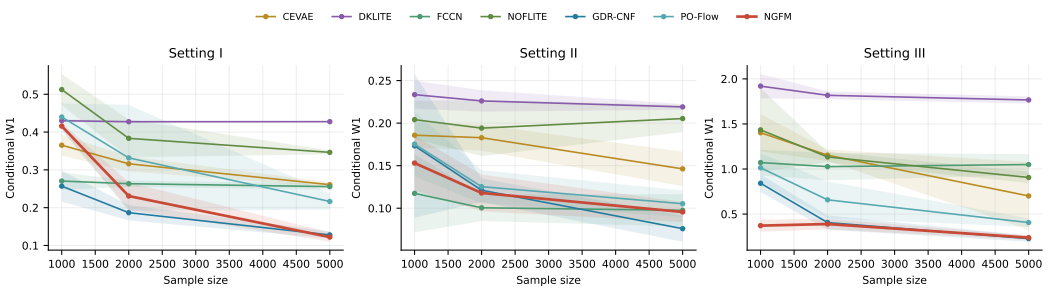

Figure 3: **Conditional W1 across sample sizes.** Lines are means over five matched repetitions; shading gives 95% confidence intervals. Each panel holds the mechanism fixed while varying sample size. DiffPO is omitted here because its much larger errors would compress the other curves; its numerical results remain in the tables and supplementary code.

Table 2: **IHDP and ACIC 2018.** Mean ± SD of rPEHE and absolute ATE error across repeated experiments. Bold marks the smallest mean per dataset and metric. Dashes indicate unavailable validated results.

	IHDP		ACIC 2018	
Method	rPEHE	ATE error	rPEHE	ATE error
TARNet	0.768±0.092	0.135±0.096	1.189±0.103	0.652±0.113
CFR	0.699±0.090	**0.092±0.074**	1.151±0.123	0.451±0.215
BWCFR	0.736±0.093	0.114±0.066	1.104±0.156	0.474±0.189
CrossNet-CMMD	0.703±0.107	0.111±0.098	1.188±0.123	0.495±0.272
PairNet	0.854±0.116	0.137±0.139	1.961±0.300	0.780±0.340
Orthogonal DR	2.147±0.478	0.497±0.196	1.621±0.185	0.562±0.133
GANITE	1.634±0.562	0.534±0.429	3.236±3.428	0.558±0.460
CEVAE	0.628±0.140	0.114±0.103	0.924±0.176	0.691±0.104
DKLITE	0.377±0.067	0.093±0.074	0.929±0.039	0.597±0.017
FCCN	0.793±0.140	0.159±0.106	0.929±0.180	0.620±0.271
NOFLITE	0.423±0.059	0.096±0.040	0.752±0.014	0.518±0.037
GDR-CNF	0.568±0.135	0.109±0.075	0.873±0.055	0.662±0.065
PO-Flow	0.665±0.084	0.191±0.109	0.636±0.193	0.452±0.315
DiffPO	1.183±0.343	0.645±0.259	0.955±0.254	0.473±0.250
NGFM (ours)	**0.330±0.066**	0.133±0.078	**0.409±0.172**	**0.399±0.169**

5.3 Which components contribute to the improvement?

Matched ablations compare the complete objective with a direct-X factual flow, removal of $\mathcal{L}_{\text{bal}}$, removal of $\mathcal{L}_{\text{ad}}$, and donor selection by Euclidean distance in the original covariate space. Each comparison fixes data partitions, optimization seeds, and shared hyperparameters.

In Setting III at $n = 1{,}000$, adding $\mathcal{L}_{\text{bal}}$ reduces cW1 from 0.383 to 0.372 and rPEHE from 0.680 to 0.650 across five repetitions. Removing donor guidance or using original-X donor distances also increases cW1.

Table 3 shows that adding $\mathcal{L}_{\text{bal}}$ lowers rPEHE from 0.380 to 0.330 on IHDP and from 0.455 to 0.409 on ACIC 2018; ATE error also decreases. Direct-X factual flow has larger rPEHE on both benchmarks. Original-X donor distances give comparable effect accuracy, especially on IHDP. The compact learned metric preserves useful neighborhoods while lowering per-distance computation; its donor graph stays fixed after warm-up.

5.4 Do endpoint shape and reliability matter?

Replacing each donor outcome by the weighted mean $\bar{y}_i = \sum_j q_{ij} y_j$ removes the empirical endpoint law's spread while retaining its mean. Setting $r_i = 1$ removes query-level attenuation. Crossing these two changes gives the four variants in Table 4; within each simulation repetition they share the frozen donor graph, distance probabilities, initialization, and training budget.

Table 3: **Matched component ablations.** Mean ± SD across repeated experiments. Variants share partitions, seeds, and common hyperparameters. Bold marks the smallest mean per dataset and metric.

	IHDP		ACIC 2018	
Variant	rPEHE	ATE error	rPEHE	ATE error
Direct-X factual flow	0.473±0.059	**0.116±0.092**	1.578±0.239	0.709±0.168
No adaptive term ($\mathcal{L}_{\mathrm{fact}} + \beta\mathcal{L}_{\mathrm{bal}}$)	0.393±0.123	0.221±0.146	0.493±0.285	0.485±0.289
No balance term ($\mathcal{L}_{\mathrm{fact}} + \mathcal{L}_{\mathrm{ad}}$)	0.380±0.089	0.144±0.111	0.455±0.139	0.441±0.134
Original-X donor distance	0.343±0.080	0.126±0.054	0.477±0.281	0.468±0.283
NGFM (full)	**0.330±0.066**	0.133±0.078	**0.409±0.172**	**0.399±0.169**

Table 4: **Endpoint shape and reliability in Settings I–II.** cW1, mean ± SD over five repetitions. These matched endpoint runs use a separate configuration from Table 1, with $p = 0.2$ in both settings. Donor mean replaces endpoints by their weighted mean. Bold marks the smallest mean in each row.

Setting	n	NGFM	$r_i = 1$	Donor mean	$r_i = 1$ + donor mean
I	500	0.420±0.003	**0.419±0.002**	0.435±0.011	0.442±0.014
I	1,000	0.416±0.007	**0.415±0.005**	0.433±0.009	0.436±0.011
I	2,000	0.328±0.076	**0.311±0.093**	0.424±0.006	0.429±0.004
I	5,000	**0.122±0.010**	0.132±0.013	0.374±0.016	0.382±0.017
II	500	**0.156±0.024**	0.170±0.031	0.256±0.031	0.266±0.033
II	1,000	**0.153±0.026**	0.160±0.023	0.216±0.036	0.227±0.034
II	2,000	**0.137±0.019**	0.145±0.023	0.187±0.018	0.198±0.022
II	5,000	**0.096±0.009**	0.096±0.013	0.194±0.010	0.206±0.014

Retaining the full endpoint distribution gives lower mean cW1 than donor-mean supervision in all eight setting–sample-size combinations, both with reliability weighting and with $r_i = 1$. This identifies information beyond the neighbor mean as useful for distribution recovery. Reliability has a smaller, setting-dependent effect: with full endpoints it improves all four Setting II means and Setting I at $n = 5{,}000$; with donor-mean supervision it improves all eight means.

Table 5: **IHDP endpoint comparison.** Mean ± SD of rPEHE and absolute ATE error across the fixed benchmark partitions. NGFM reuses the reference fits in Tables 2–3; the three endpoint controls use standardized training data (Appendix C). Bold marks the smallest mean per metric.

Variant	rPEHE	ATE error
NGFM	**0.330±0.066**	**0.133±0.078**
$r_i = 1$	0.451±0.108	0.195±0.083
Donor mean	0.456±0.089	0.165±0.124
$r_i = 1$ + donor mean	0.469±0.079	0.184±0.099

Table 5 reports the corresponding IHDP variants. The NGFM reference has the lowest rPEHE and ATE error among these rows. Appendix F.1 reports the sweeps of donor count, reliability rank, and balance strength.

6 Conclusion

NGFM combines representation alignment with local distributional supervision. Its weighted donor law preserves observed outcome variation, while query reliability controls borrowing under scarce treatment support. The theory connects locality and sampling error to distribution recovery; the experiments identify where conditional laws and mean effects improve.

AI USE STATEMENT

Generative AI tools assisted manuscript drafting, mathematical checking, experimental exposition, and plotting code. Scientific claims and numerical records remain the authors' responsibility.

REPRODUCIBILITY STATEMENT

We will submit supplementary code with the manuscript. The supplementary code contains the NGFM implementation, simulation generators, benchmark configurations, training and evaluation routines, distribution-comparison figures, and complete available NGFM result tables, including ablations and parameter sweeps. Appendices A–B provide the analysis and proofs; Appendices C–E describe data generation, metrics, training, and network architectures. Appendix D contains the training and inference framework in Figure 4 and the complete procedure in Algorithm 1; Section D.1 gives the flow-matching derivation and sampling details. Appendix E details the network architectures in Table 6. Appendix F gives extended numerical and graphical results. Machine-readable records and plotting inputs accompany the tables and figures.

REFERENCES

Serge Assaad, Shuxi Zeng, Chenyang Tao, Shounak Datta, Nikhil Mehta, Ricardo Henao, Fan Li, and Lawrence Carin. Counterfactual representation learning with balancing weights. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1972–1980, 2021. URL https://proceedings.mlr.press/v130/assaad21a.html.

Susan Athey and Guido W. Imbens. Machine learning methods for estimating heterogeneous causal effects. *arXiv preprint arXiv:1504.01132v1*, 2015. URL https://arxiv.org/abs/1504.01132v1.

John Charles Butcher. A history of Runge–Kutta methods. *Applied Numerical Mathematics*, 20(3): 247–260, 1996.

Patrick Chao, Patrick Blöbaum, and Shiva Prasad Kasiviswanathan. Interventional and counterfactual inference with diffusion models. *arXiv preprint arXiv:2302.00860*, 2023. URL https://arxiv.org/abs/2302.00860v1.

Li Chen, Xiaotong Shen, and Wei Pan. Enhancing causal effect estimation with diffusion-generated data. *arXiv preprint arXiv:2504.03630*, 2025.

Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6571–6583, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018. ISSN 1368-4221, 1368-423X. doi: 10.1111/ectj.12097.

Oscar Clivio, Alexander Nicholas D'Amour, Alexander Franks, David Bruns-Smith, Christopher C. Holmes, and Avi Feller. Deconfounding scores and representation learning for causal effect estimation with weak overlap. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 1288–1296, 2026. URL https://proceedings.mlr.press/v300/clivio26a.html.

Giovanni Corrao, Federica Nicotra, Andrea Parodi, Antonella Zambon, Davide Soranna, Franca Heiman, Luca Merlino, and Giuseppe Mancia. External adjustment for unmeasured confounders improved drug–outcome association estimates based on health care utilization data. *Journal of Clinical Epidemiology*, 65(11):1190–1199, 2012.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Chaorui Deng, Deyao Zhu, Kunchang Li, Shi Guang, and Haoqi Fan. Causal diffusion transformers for generative modeling. *arXiv preprint arXiv:2412.12095*, 2024.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Daniel Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical science*, 34(1):43–68, 2019.

Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using sinkhorn divergences. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 2681–2690, 2019. URL https://proceedings.mlr.press/v89/feydy19a.html.

Christian Gische and Manuel C Voelkle. Beyond the mean: A flexible framework for studying causal effects using linear models. *Psychometrika*, 87(3):868–901, 2022.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Negar Hassanpour and Russell Greiner. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019.

Yinan He, Gengshi Huang, Siyu Chen, Jianing Teng, Kun Wang, Zhenfei Yin, Lu Sheng, Ziwei Liu, Yu Qiao, and Jing Shao. X-learner: Learning cross sources and tasks for universal visual representation. In *European Conference on Computer Vision*, pp. 509–528. Springer, 2022.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, Boca Raton, 2020.

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

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 4829–4838. PMLR, 2021. URL https://proceedings.mlr.press/v139/jesson21a.html.

Fredrik Johansson, Uri Shalit, and David Sontag. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029. PMLR, 2016.

Nathan Kallus and Miruna Oprescu. Robust and agnostic learning of conditional distributional treatment effects. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6037–6060, 2023. URL https://proceedings.mlr.press/v206/kallus23a.html.

Leonid V Kantorovich. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4): 1381–1382, 2006. doi: 10.1007/s10958-006-0049-2.

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

Edward H Kennedy, Sivaraman Balakrishnan, and LA Wasserman. Semiparametric counterfactual density estimation. *Biometrika*, 110(4):875–896, 2023.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations (ICLR 2015), Conference Track Proceedings*, 2015. URL https://arxiv.org/abs/1412.6980.

James K Kirklin, David C Naftel, Robert L Kormos, Lynne W Stevenson, Francis D Pagani, Marissa A Miller, Karen L Ulisney, J Timothy Baldwin, and James B Young. Second INTERMACS annual report: more than 1,000 primary left ventricular assist device implants. *The Journal of Heart and Lung Transplantation*, 29(1):1–10, 2010.

Insung Kong, Yuha Park, Joonhyuk Jung, Kwonsang Lee, and Yongdai Kim. Covariate balancing using the integral probability metric for causal inference. In *International Conference on Machine Learning*, pp. 17430–17461. PMLR, 2023.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Changyu Liu, Yuling Jiao, Junhui Wang, and Jian Huang. Wasserstein distributionally robust nonparametric regression. *arXiv preprint arXiv:2505.07967*, 2025.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/94b5bde6de888ddf9cde6748ad2523d1-Abstract.html.

Jared K Lunceford and Marie Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine*, 23(19): 2937–2960, 2004.

Yuchen Ma, Valentyn Melnychuk, Jonas Schweisthal, and Stefan Feuerriegel. DiffPO: A causal diffusion model for learning distributions of potential outcomes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 43663–43692, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/4d254c96c5ff500dda2ac0a58987aeba-Abstract-Conference.html.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Valentyn Melnychuk and Stefan Feuerriegel. GDR-learners: Orthogonal learning of generative models for potential outcomes. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ac12d0596191a29629f151d97a93653e-Abstract-Conference.html.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Normalizing flows for interventional density estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24361–24397, 2023. URL https://proceedings.mlr.press/v202/melnychuk23a.html.

Valentyn Melnychuk, Stefan Feuerriegel, and Mihaela van der Schaar. Quantifying aleatoric uncertainty of the treatment effect: A novel orthogonal learner. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bdabb5d4262bcfb6a1d529d690a6c82b-Abstract-Conference.html.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *International Conference on Learning Representations*, 2024b. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/abe31a12e83111fdf2cfd54deed5a2ce-Abstract-Conference.html.

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Overlap-adaptive regularization for conditional average treatment effect estimation. In *International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=HMMSnGgYOy.

Valentyn Melnychuk, Dennis Frauen, Jonas Schweisthal, and Stefan Feuerriegel. Orthogonal representation learning for estimating causal quantities. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5086–5094, 2026b. URL https://proceedings.mlr.press/v300/melnychuk26a.html.

Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pp. 666–704, 1781.

Lokesh Nagalapatti, Pranava Singhal, Avishek Ghosh, and Sunita Sarawagi. PairNet: Training with observed pairs to estimate individual treatment effect. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 37237–37259, 2024. URL https://proceedings.mlr.press/v235/nagalapatti24a.html.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Paul R. Rosenbaum and Donald B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983. doi: 10.1093/biomet/70.1.41. URL https://doi.org/10.1093/biomet/70.1.41.

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

Piotr Rzepakowski and Szymon Jaroszewicz. Decision trees for uplift modeling with single and multiple treatments. *Knowledge and Information Systems*, 32(2):303–327, 2012.

Pedro Sanchez, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. Diffusion models for causal discovery via topological ordering. In *International Conference on Learning Representations*, 2023.

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017. URL https://proceedings.mlr.press/v70/shalit17a.html.

Claudia Shi, David Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. *Advances in neural information processing systems*, 32, 2019.

Pengfei Shi, Wei Zhong, Xinyu Zhang, Ningtao Wang, Xing Fu, Weiqiang Wang, and Yin Jin. Estimating conditional average treatment effects via sufficient representation learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pp. 4894–4901, 2024. doi: 10.24963/ijcai.2024/541. URL https://www.ijcai.org/proceedings/2024/541.

Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv preprint arXiv:1802.05046*, 2018. URL https://arxiv.org/abs/1802.05046.

Pablo Sánchez-Martin, Miriam Rateike, and Isabel Valera. Vaca: Designing variational graph autoencoders for causal queries. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (7):8159–8168, Jun. 2022. doi: 10.1609/aaai.v36i7.20789. URL https://ojs.aaai.org/index.php/AAAI/article/view/20789.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=CD9Snc73AW.

Toon Vanderschueren, Jeroen Berrevoets, and Wouter Verbeke. NOFLITE: Learning to predict individual treatment effect distributions. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=EjqopDxLbG.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Hao Wang, Jiajun Fan, Zhichao Chen, Haoxuan Li, Weiming Liu, Tianqiao Liu, Quanyu Dai, Yichao Wang, Zhenhua Dong, and Ruiming Tang. Optimal transport for treatment effect estimation. *Advances in Neural Information Processing Systems*, 36:5404–5418, 2023.

Dongze Wu, David I. Inouye, and Yao Xie. Flow-based generative modeling of potential outcomes and counterfactuals. In *IEEE International Symposium on Information Theory*, 2026. URL https://arxiv.org/abs/2505.16051.

Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ByKWUeWA-.

Yao Zhang, Alexis Bellot, and Mihaela van der Schaar. Learning overlapping representations for the estimation of individualized treatment effects. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1005–1014, 2020. URL https://proceedings.mlr.press/v108/zhang20c.html.

Tianhui Zhou, William E. Carson, IV, and David Carlson. Estimating potential outcome distributions with collaborating causal networks. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=q1Fey9feu7.

A Detailed analysis of the NGFM objective

The outcome of a donor is a *path endpoint*, not a substitute for the query's missing potential outcome. We analyze the endpoint measure first, then the error from replacing the target law by weighted observed donors, and finally the error of fitting a flow to that measure. These statements concern scalar outcomes.

The empirical endpoint measure. Condition on the training sample and the frozen graph. Write $c_i = \mathbf{1}\{|N_i| > 0\}$ and choose any law for $\widehat{Q}_i$ when $c_i = 0$. The unnormalized measure

$$M_n(dx, da, dy) = \frac{1}{n}\sum_{i=1}^{n}\Big[\delta_{(X_i,A_i,Y_i)}(dx, da, dy) + c_i r_i\, \delta_{(X_i,1-A_i)}(dx, da)\widehat{Q}_i(dy)\Big] \tag{8}$$

has total mass $m_n = 1 + n^{-1}\sum_i c_i r_i$. Sampling (x, a, y) from M_n/m_n and then t, ϵ gives exactly $(\mathcal{L}_{\text{fact}} + \mathcal{L}_{\text{ad}})/m_n$. In particular, $\sum_j q_{ij}\ell_{1-A_i}(X_i, Y_j)$ is a mixture of losses at the observed Y_j, whereas $\ell_{1-A_i}(X_i, \sum_j q_{ij}Y_j)$ would define a point-mass endpoint. The alignment penalty $\mathcal{L}_{\text{bal}}$ affects the learned representation but contributes no endpoints to M_n.

For any endpoint law T with a finite second moment, let $U_t = (1-t)\epsilon + tY$, $Y \sim T$, and $V = Y - \epsilon$. Squared-loss projection identifies the unrestricted optimal velocity and its excess risk:

$$v_T^*(u,t) = \mathbb{E}[V \mid U_t = u, t], \qquad \mathbb{E}|v(U_t,t) - V|^2 - \mathbb{E}|v_T^*(U_t,t) - V|^2 = \mathbb{E}|v(U_t,t) - v_T^*(U_t,t)|^2. \tag{9}$$

With conditioning restricted to $Z = \Phi_\phi(X)$, the optimum instead pools endpoints from covariates sharing the same representation. Thus the target at a particular x below is the covariate-conditional target before representation pooling.

Borrowing mass and donor error. For a population interpretation, take a donor bank independent of a new query, fix its graph, and let $e_a(x) = P(A = a \mid X = x)$. Let $c_a(x)$ indicate an available arm-a neighborhood for a query in arm $1-a$, with reliability $r_a(x)$ and donor law $\widehat{Q}_a(\cdot \mid x)$. The endpoint law assigned to head a at x is

$$T_a = (1-\alpha_a)P_a + \alpha_a\widehat{Q}_a, \qquad \alpha_a(x) = \frac{e_{1-a}(x)c_a(x)r_a(x)}{e_a(x) + e_{1-a}(x)c_a(x)r_a(x)}. \tag{10}$$

The formula follows by normalizing factual mass $e_a(x)$ and borrowed mass $e_{1-a}(x)c_a(x)r_a(x)$. It shows why reliability should attenuate a distant opposite arm. For real-valued outcomes, the CDF representation of W_1 gives the exact identity $W_1(P_a, T_a) = \alpha_a W_1(P_a, \widehat{Q}_a)$.

For a nonempty donor set, let $\mathcal{G}$ contain an outcome-independent graph, its donor covariates and arm labels, and the fixed query. Suppose donor outcomes are conditionally independent with laws $P_a(\cdot \mid X_j)$, all supported on a common interval of length D, and $W_1\{P_a(\cdot \mid X_j), P_a(\cdot \mid x)\} \le L_a d(x, X_j)$. Then, with $K_{\text{eff}} = (\sum_j q_j^2)^{-1}$,

$$\mathbb{E}[W_1(\widehat{Q}_a, P_a) \mid \mathcal{G}] \le L_a \sum_{j \in N_a(x)} q_j d(x, X_j) + \frac{D}{2\sqrt{K_{\text{eff}}}}. \tag{11}$$

The first term is weighted locality bias. The second is finite-donor noise: placing more weight on the closest donors may reduce the first term while reducing K_{eff}. Appendix B proves the bound and an unbounded-outcome version appropriate to Gaussian and Laplace noise.

From endpoints to a generated law. When $\alpha_a = 0$, the donor contribution is defined as zero without assigning an effective donor count to an empty set. For a fitted velocity $v_{\theta,a}$, let $\mathcal{E}_{a,x}$ be the population excess risk in Eq. (9) under T_a, and let $\nu_{a,x}$ be the W_1 error from numerical ODE integration. If $v_{\theta,a}$ is $L(t)$-Lipschitz in its state and its ODE is well posed, set $C_L = \int_0^1 \exp(2\int_t^1 L(s)\,ds)\,dt$. Coupling its trajectory to an oracle characteristic process yields

$$\mathbb{E}[W_1(P_a, \widehat{P}_{\theta,a}) \mid \mathcal{G}] \le \alpha_a \left(L_a \sum_j q_j d(x, X_j) + \frac{D}{2\sqrt{K_{\text{eff}}}} \right) + \mathbb{E}[\sqrt{C_L \mathcal{E}_{a,x}} + \nu_{a,x} \mid \mathcal{G}]. \tag{12}$$

The fitted velocity and representation affect the residual $\mathcal{E}_{a,x}$; raw training loss is not a bound on it. The empirical graph uses an outcome-trained warm-up encoder, so the conditional donor guarantee requires an independent or sample-split metric. Equation (8) remains exact for the implemented objective. Appendix B gives the coupling argument and the precise assumptions.

Consequences for causal summaries. For integrable scalar laws, the Kantorovich–Rubinstein duality gives $|\widehat{\mu}_a(x) - \mu_a(x)| \le W_1(\widehat{P}_a(\cdot \mid x), P_a(\cdot \mid x))$. Hence the conditional treatment-effect error obeys

$$|\widehat{\tau}(x) - \tau(x)| \le W_1(\widehat{P}_1, P_1) + W_1(\widehat{P}_0, P_0).$$

The same inequality bounds absolute ATE error after averaging over the target covariate distribution. For the target covariate law, Minkowski's inequality further gives

$$\text{rPEHE} \le \{\mathbb{E}_X W_1^2(\widehat{P}_1, P_1)\}^{1/2} + \{\mathbb{E}_X W_1^2(\widehat{P}_0, P_0)\}^{1/2}.$$

Thus controlling mean cW1 alone does not control rPEHE without second-moment information. If all true and generated outcomes lie in a common interval of length D, then $W_1 \le D$ and the displayed inequality yields $\text{rPEHE} \le 2\sqrt{D\,\text{cW1}}$, where cW1 averages the two arms over X. These relationships motivate reporting both distributional and mean-effect metrics. A small weighted donor radius is useful when the outcome law is regular in the chosen frozen metric; L_a states that requirement explicitly.

B Theoretical details and proofs

Identification and conditioning. Let $(X_i, A_i, Y_i(0), Y_i(1))$ be independent observations, with $Y_i = Y_i(A_i)$. Under consistency, $Y(a) \perp A \mid X$, and $e_a(x) > 0$ on the target population, $P_a(\cdot \mid x) = \mathcal{L}(Y \mid X = x, A = a)$ almost everywhere. This identifies each marginal potential-outcome law. It does not identify the joint law of $(Y(0), Y(1)) \mid X$. A graph based on X and A alone leaves donor outcomes conditionally independent, even though the same donor can be reused in losses for many queries.

Exact empirical objective. Let M_n be Eq. (8) and $m_n = M_n(\mathcal{X} \times \{0,1\} \times \mathbb{R})$. Its auxiliary part is $n^{-1}\sum_i r_i \delta_{(X_i, 1-A_i)} \sum_j q_{ij}\delta_{Y_j}$. Therefore integration of the path loss against M_n equals $\mathcal{L}_{\text{fact}} + \mathcal{L}_{\text{ad}}$. This equality requires no independence between graph edges. For fixed encoder Φ_ϕ, the optimal measurable velocity at (u, t, z, a) is

$$v^*(u, t, z, a) = \mathbb{E}[Y - \epsilon \mid U_t = u, t, \Phi_\phi(X) = z, A^\star = a],$$

where $A^\star$ denotes the endpoint arm in M_n. The conditional expectation is an orthogonal projection among square-integrable functions: expanding a candidate's squared residual around v^* makes the cross term vanish. This proves Eq. (9) and shows precisely where representation pooling enters. Replacing the weighted donor sum by a loss at $\sum_j q_{ij} Y_j$ would replace $\widehat{Q}_i$ by a point mass and change this projection.

Population endpoint mixture. For Eq. (10), take a donor bank independent of a fresh query (X, A, Y). Conditional on that bank and $X = x$, factual paths ending in arm a have measure $e_a(x)P_a(dy \mid x)$. Queries

whose factual arm is $1-a$ provide borrowed paths of measure $e_{1-a}(x)c_a(x)r_a(x)\widehat{Q}_a(dy \mid x)$. Their sum has mass $w_a(x) = e_a(x) + e_{1-a}(x)c_a(x)r_a(x) > 0$. Division by $w_a(x)$ proves the mixture. In one dimension,

$$W_1(P,Q) = \int_{\mathbb{R}} |F_P(y) - F_Q(y)|\, dy.$$

Since $F_{T_a} - F_{P_a} = \alpha_a(F_{\widehat{Q}_a} - F_{P_a})$, the equality $W_1(P_a, T_a) = \alpha_a W_1(P_a, \widehat{Q}_a)$ follows. The propensity in this derivation represents the frequency of factual and borrowed paths; the algorithm does not estimate it.

Weighted-donor bound. Condition on $\mathcal{G}$, the donor covariates, arm labels, and outcome-independent weights. Put $\overline{Q} = \sum_j q_j P_a(\cdot \mid X_j)$. Convexity and local W_1 continuity give

$$W_1(\overline{Q}, P_a(\cdot \mid x)) \le \sum_j q_j W_1(P_a(\cdot \mid X_j), P_a(\cdot \mid x)) \le L_a \sum_j q_j d(x, X_j).$$

Writing $\widehat{F}(y) = \sum_j q_j \mathbf{1}\{Y_j \le y\}$ and $\overline{F}(y) = \sum_j q_j F_j(y)$, conditional independence gives

$$\mathrm{Var}(\widehat{F}(y) \mid \mathcal{G}) = \sum_j q_j^2 F_j(y)\{1 - F_j(y)\} \le \tfrac{1}{4}\sum_j q_j^2.$$

Jensen's inequality bounds $\mathbb{E}|\widehat{F}(y) - \overline{F}(y)|$ by $\frac{1}{2}\sqrt{\sum_j q_j^2}$. Integrating across the common support interval of length D and applying the triangle inequality proves Eq. (11). Reuse of donor j for another query does not enter this single-query variance calculation.

For an unbounded law, suppose $\max_j \mathbb{E}[|Y_j - m|^q \mid \mathcal{G}] \le M_q$ for some $q > 2$. Split the CDF integral at $m \pm h$. The central interval contributes at most $h\sqrt{\sum_j q_j^2}$. Cauchy–Schwarz and Markov's inequality bound each tail by

$$\sqrt{\sum_j q_j^2}\int_h^\infty \sqrt{M_q/u^q}\, du = \sqrt{\sum_j q_j^2}\,\frac{2\sqrt{M_q}}{q-2}h^{1-q/2}.$$

For $M_q = 0$, all donors equal the common center and the sampling term vanishes. Otherwise, choosing $h = (4M_q)^{1/q}$ yields the bound in Eq. (11) with $D/2$ replaced by $H_q = \frac{q}{q-2}(4M_q)^{1/q}$. Thus Gaussian and Laplace outcomes can be covered under a uniform moment condition. For bounded outcomes, changing Y_j changes $W_1(\widehat{Q}_a, P_a)$ by at most Dq_j; McDiarmid's inequality additionally gives, with probability at least $1-\delta$ conditional on $\mathcal{G}$,

$$W_1(\widehat{Q}_a, P_a) \le L_a \sum_j q_j d(x, X_j) + D\sqrt{\sum_j q_j^2}\left(\frac{1}{2} + \sqrt{\frac{\log(1/\delta)}{2}}\right).$$

Flow stability with empirical endpoints. The donor law may be discrete, so regularity of the oracle velocity at $t = 1$ cannot simply be assumed. For any target T with finite second moment, $U_t = (1-t)\epsilon + tY$ has a positive density p_t for $t < 1$, with CDF F_t. Differentiation of the Gaussian convolution gives $\partial_t F_t(u) = -p_t(u)v_T^*(u,t)$. If $S \sim \mathrm{Unif}(0,1)$ and $B_t = F_t^{-1}(S)$, then $\dot{B}_t = v_T^*(B_t, t)$ for $t < 1$, $B_0 \sim \mathcal{N}(0,1)$, and $\mathcal{L}(B_t) = \mathcal{L}(U_t)$. Conditional Jensen yields $\int_0^1 \mathbb{E}|v_T^*(B_t,t)|^2 dt \le \mathbb{E}|Y - \epsilon|^2$, so the characteristic has an absolutely continuous extension with $B_1 \sim T$, even when T contains atoms.

For a learned velocity v, suppose $|v(u,t) - v(u',t)| \le L(t)|u - u'|$ and $\int_0^1 [L(t) + |v(0,t)|]dt < \infty$. Couple its ODE trajectory $\widehat{B}_t$ to B_t using the same B_0. Gronwall's inequality and Cauchy–Schwarz give

$$\mathbb{E}|\widehat{B}_1 - B_1|^2 \le \left\{\int_0^1 \exp\left(2\int_t^1 L(s)\, ds\right) dt\right\}\int_0^1 \mathbb{E}|v(B_t,t) - v_T^*(B_t,t)|^2 dt = C_L\mathcal{E}_T(v).$$

Their joint law is a coupling, hence $W_1(\mathcal{L}(\widehat{B}_1), T) \le \sqrt{C_L\mathcal{E}_T(v)}$. Add numerical integration error ν and $W_1(P_a, T_a) = \alpha_a W_1(P_a, \widehat{Q}_a)$, condition on the graph, and apply the donor bound to obtain Eq. (12). Here α_a, the weights, and the donor radius are $\mathcal{G}$-measurable, while the fitted velocity, its Lipschitz envelope, and the excess risk may depend on donor outcomes and remain inside the conditional expectation. The excess velocity risk is a population quantity. The result does not replace it by an empirical training loss.

Scope of the graph assumption. The implemented warm-up encoder uses factual outcomes before it is frozen. Conditioning on its graph therefore does not automatically preserve donor-outcome independence. An encoder fitted on a separate sample independent of the entire donor bank satisfies the stated premise. A sample-split implementation must fit the metric outside the donor bank being analyzed; excluding only the query or one donor at a time is insufficient if the graph still depends on other donor outcomes. The exact empirical-measure identity remains valid for the implemented graph without this premise; the donor and flow bounds describe the independent-metric surrogate.

Recovery and effect bounds (Corollary 1). Because $0 \le \alpha_a \le 1$, the first conclusion follows directly from Eq. (12) when its radius, effective-sample, and expected flow residual terms vanish. This is convergence of the conditional expected W_1 error; conditional convergence in probability follows by Markov's inequality. It does not by itself assert almost-sure recovery. For mean effects, the function $y \mapsto y$ is 1-Lipschitz, so Kantorovich–Rubinstein duality (Villani et al., 2008) gives $|\widehat{\mu}_a(x) - \mu_a(x)| \le W_1(\widehat{P}_a, P_a)$. Applying the triangle inequality to the two arm means proves Eq. (7). Jensen's inequality yields the ATE bound, and squaring and integrating the pointwise inequality yields the stated $L^2(P_X)$ bound for rPEHE. Minkowski's inequality further separates it into the sum of the two arm-specific L^2 norms recorded in Appendix A. These consequences require finite corresponding moments; a fixed finite donor count alone does not make the sampling term vanish.

C Data and experimental protocol

Simulation families. Setting I is a local-support design: treatment probability changes across a covariate boundary, and outcomes in the locally rare treatment cells have two separated Gaussian modes. Setting II makes one arm globally scarce and gives treated outcomes Laplace noise. Setting III uses the nonlinear outcome mechanism of Melnychuk et al. (2026b). Table 1 uses 2,000 observations in each family; the related sample-size study uses the mechanisms specified below. Each simulation result summarizes five independently generated data repetitions. Within each repetition, all methods use identical train, validation, and test indices. The saved partitions are stratified by treatment, with approximately 70%/15%/15% for training, validation, and testing. Integer rounding within arms gives 1,399–1,400 training, 299–300 validation, and 300–302 test observations at this sample size.

Semi-synthetic benchmarks. IHDP uses fixed splits of one supplied 747-unit realization, each with 522 training, 149 validation, and 76 test observations (approximately 70%/20%/10%). ACIC 2018 uses a 25,000-unit task with 17,500 training, 3,750 validation, and 3,750 test observations (70%/15%/15%); five optimization runs reuse this partition. Setting IV (HC-MNIST, $\Gamma = 1$) uses five generated realizations, each with 35,000 training, 15,000 validation, and 10,000 test images; its construction and evaluation are detailed below. IHDP and ACIC 2018 use their supplied effect truth. Neither record supplies a verified full conditional potential-outcome law for cW1.

C.1 Data-generating mechanisms

All simulated outcomes below are given on their original scale. For the simulations, model-input covariates and factual outcomes are standardized using training-set moments; generated outcomes are transformed back before evaluation. Treatment is sampled as $A \mid X = x \sim \mathrm{Bernoulli}(e(x))$, and only $Y = Y(A)$ is available for fitting. The remaining coordinates in higher-dimensional variants are independent standard Gaussian noise covariates.

Setting I: local scarcity and bimodal outcomes. We construct a local-support stress test with independent $X_j \sim \mathcal{N}(0,1)$. Let $\mu_a(x) = 0.2(x_1 + x_2) + a$ and

$$e(x) = \begin{cases} p, & x_1 < 0, \\ 1-p, & x_1 \ge 0, \end{cases} \qquad b_a(x) = \mathbf{1}\{x_1 < 0, a = 1\} + \mathbf{1}\{x_1 \ge 0, a = 0\}.$$

With independent $S_a \in \{-1, 1\}$ of equal probability and $\epsilon_a \sim \mathcal{N}(0,1)$, the potential outcomes are

$$Y(a) = \mu_a(X) + 1.1 b_a(X) S_a + 0.35\epsilon_a.$$

Thus a locally common arm has a unimodal Gaussian law, while a locally rare arm has an equally weighted two-Gaussian mixture with modes separated by 2.2. The mean effect is one throughout. The main table uses $d = 2$, $p = 0.15$, and $n = 2{,}000$; the sample-size curves use $d = 2$ and $p = 0.2$. This is our constructed design, rather than a copied benchmark mechanism.

Setting II: global scarcity and different tail shapes. Independently draw $X_j \sim \mathcal{N}(0,1)$ and set $e(x) = p$. With the same μ_a as Setting I,

$$Y(0) = \mu_0(X) + 0.35\epsilon_0, \qquad Y(1) = \mu_1(X) + 0.65 L_1,$$

where $\epsilon_0 \sim \mathcal{N}(0,1)$ and $L_1 \sim \mathrm{Laplace}(0,1)$ are independent. This constructed design separates global arm scarcity from covariate-dependent support. The main table and sample-size curves use $d = 2$ and $p = 0.2$.

Setting III: nonlinear outcome mechanism. Following the nonlinear simulation in Melnychuk et al. (2026b), write $Z = X_1 \sim \mathrm{Unif}[-2, 2]$ and $U = X_2 \sim \mathcal{N}(0,1)$, independently. The implemented propensity

and potential outcomes are

$$\begin{aligned} e(z,u) &= \text{sigmoid}(0.75z - u + 0.5), \\ \mu_a(z,u) &= (2a-1)z + a - 2\sin\{2(2a-1)z + u\} - 2u(1+0.5z), \\ Y(a) &= \mu_a(Z,U) + \epsilon_a, \qquad \epsilon_a \sim \mathcal{N}(0,1). \end{aligned}$$

Both Z and U are observed. The treatment-dependent sine argument is retained for both arms. The main table uses $d = 2$ and $n = 2{,}000$; varying sample size leaves this mechanism unchanged.

Setting IV: HC-MNIST. HC-MNIST combines MNIST images with synthetic treatment and outcome mechanisms following Jesson et al. (2021). The input is a 28×28 grayscale image I, flattened to 784 covariates after pixel normalization $(I/255 - 0.1307)/0.3081$. For digit class $c \in \{0, \ldots, 9\}$, let $m(I)$ be the average pixel intensity on the $[0,1]$ scale and let m_c, s_c be its class-specific mean and sample SD, computed from the 60,000 official MNIST training images. Define

$$v(I) = \text{clip}\left(\frac{m(I) - m_c}{s_c}, -1.4, 1.4\right), \qquad z = \phi(I) = -2 + 0.4c + \frac{0.4}{2.8}\{v(I) + 1.4\}.$$

Thus each digit occupies one of ten equal intervals in $[-2, 2]$. Independently draw $U \sim \text{Bernoulli}(1/2)$ and set $\Gamma = 1$ for the main Setting IV results. The appendix confounding test uses $\Gamma = e$. With

$$\pi(z) = \text{sigmoid}(0.75z + 0.5), \quad a_\Gamma(z) = \{\Gamma\pi(z)\}^{-1} + 1 - \Gamma^{-1}, \quad b_\Gamma(z) = \Gamma\pi(z)^{-1} + 1 - \Gamma,$$

the implemented mechanism is

$$\begin{aligned} A \mid z,u &\sim \text{Bernoulli}\{u/a_\Gamma(z) + (1-u)/b_\Gamma(z)\}, \\ \mu_a(z,u) &= (2a-1)z + (2a-1) - 2\sin\{(4a-2)z\} - (4u-2)(1+0.5z), \\ Y(a) &= \mu_a(z,u) + \epsilon_a, \qquad \epsilon_a \sim \mathcal{N}(0,1). \end{aligned}$$

The digit label, z, U, propensity, and potential outcomes are retained for data generation or evaluation, but only image pixels, treatment, and the factual outcome enter fitting. At $\Gamma = 1$, $a_\Gamma(z) = b_\Gamma(z) = 1/\pi(z)$, so treatment depends on the image through z but not on hidden U after conditioning on the image. The main Setting IV experiment therefore satisfies conditional exchangeability. At $\Gamma = e$, treatment also depends on U, yielding the appendix confounding test. The contribution of U to μ_a is the same in both arms, so it cancels from the conditional mean contrast.

For each of five seeds, the 60,000 official training images are permuted; the first 50,000 supply 35,000 training and 15,000 validation images. The 10,000 official test images form the test set. The retained sample therefore has 60,000 units with proportions $7/12$, $1/4$, and $1/6$ (58.33%/25%/16.67%). The training and validation image partitions and synthetic variables vary across seeds; test images remain fixed while their synthetic variables are regenerated. Factual outcomes are standardized using training-set moments and predictions are restored to the original scale.

Effect metrics use all 10,000 test images and the supplied mean contrast. The stored conditional W1 evaluation uses the first 128 test images and 100 predictive particles per arm. Writing $b_a(z) = (2a-1)z + (2a-1) - 2\sin\{(4a-2)z\}$ and $h(z) = 2(1+0.5z)$, the reference interventional law integrates out hidden U:

$$P_a(\cdot \mid I) = \tfrac{1}{2}\mathcal{N}(b_a(z) - h(z), 1) + \tfrac{1}{2}\mathcal{N}(b_a(z) + h(z), 1).$$

Its quantiles are computed by numerical CDF inversion. The stored score named PEHE is already root PEHE, as confirmed by the evaluation code.

IHDP. The semi-synthetic Infant Health and Development Program benchmark combines observed infant and maternal covariates with simulated potential outcomes (Hill, 2011; Louizos et al., 2017). The supplied realization has 747 units and 25 pretreatment covariates (six continuous and 19 binary), with 139 treated and 608 control units. We use its supplied conditional means as effect truth. Each partition contains 522 training, 149 validation, and 76 test observations. Treatment and outcomes are excluded from the covariate matrix.

ACIC 2018. The ACIC 2018 benchmark of Shimoni et al. (2018) uses real covariates with simulated assignments and potential outcomes. Our task contains 25,000 observations and 177 input covariates after preprocessing. The fixed split contains 17,500 training, 3,750 validation, and 3,750 test units. We compute rPEHE and ATE error from the supplied potential-outcome truth. This is the ACIC 2018 benchmark, distinct from the ACIC 2016 competition analyzed by Dorie et al. (2019).

Training and selection. NGFM first minimizes factual flow loss. Its encoder is copied and frozen for donor distances; representation coordinates are standardized with training-set moments. The main encoder continues to train. The reference Sinkhorn penalty uses unit-normalized representation vectors and `geomloss.SamplesLoss(loss="sinkhorn", p=2, blur=0.05)`. Each post-warm-up update evaluates the distance-weighted loss for every selected opposite-arm outcome. Validation factual CRPS selects a post-warm-up checkpoint, including in Setting IV. Test cW1 uses 500 generated particles per query and treatment arm in Settings I–III; Setting IV uses the 100-particle protocol below. The same data split and seed are used for each matched component ablation.

For Setting IV, the main-table $\Gamma = 1$ NGFM results use factual validation CRPS for checkpoint selection: cW1 0.748 ± 0.016, rPEHE 1.224 ± 0.044, and ATE error 0.127 ± 0.118 across five repetitions. The same selection rule is used for the $\Gamma = e$ confounding test in Appendix F.3. Both use the reliability rule in Eq. (1).

The simulation DiffPO adapter uses its propensity-adjusted diffusion objective with the corrected per-unit weight shape. For the reported rerun, estimated propensities are clipped to $[0.10, 0.90]$, inverse-propensity weights to $[0.5, 2.0]$, and the gradient norm to 0.25; the final forward noise-schedule coefficient is 0.1. The simulation grid contains 26 combinations of sample size, dimension, and treatment probability: 11 in Setting I, 12 in Setting II, and three in Setting III. Each uses five complete repetitions and the same saved splits as the other generators; the supplementary tables enumerate the grid.

Parameter sweeps. The sensitivity grid uses one five-repetition case from each simulation family. Setting I uses 5,000 observations; Settings II and III use 1,000. The K sweep uses $8, 16, 32$ in Settings I and II and additionally 4 in Setting III. The R sweep uses $4, 8, 16, 32$ with $K = 32$ fixed. The β sweep uses $0, 0.1, 1, 3$. Within each family and repetition, all points reuse the same saved data partition and optimization seed; only the named parameter changes. Means and 95% confidence intervals are computed across five repetitions.

Metrics and repetition units. For each test set $\mathcal{T}$ of size m, write $P_a(\cdot \mid X_i)$ for the true conditional outcome law, $\widehat{P}_a(\cdot \mid X_i)$ for its generated estimate, and $\widehat{\mu}_a(X_i) = \mathbb{E}_{\widehat{P}_a}[Y]$. Let $\widehat{\tau}_i = \widehat{\mu}_1(X_i) - \widehat{\mu}_0(X_i)$ and let τ_i denote the supplied benchmark effect truth. Effect errors for one repetition are

$$\mathrm{PEHE} = \frac{1}{m}\sum_{i\in\mathcal{T}}(\widehat{\tau}_i - \tau_i)^2, \qquad \mathrm{rPEHE} = \sqrt{\mathrm{PEHE}}, \qquad \text{ATE error} = \left|\frac{1}{m}\sum_{i\in\mathcal{T}}\widehat{\tau}_i - \frac{1}{m}\sum_{i\in\mathcal{T}}\tau_i\right|. \quad (13)$$

Every displayed rPEHE value is the mean of repetition-level square roots, with its sample SD across the same repetitions. In particular, the square root is taken *before* aggregation, never on a reported mean PEHE or its SD. The Setting IV NGFM scores are calculated from the saved test predictions; the baseline workbook labels its already-rooted scores "PEHE", which we relabel rPEHE without a second square root.

Where conditional laws are available, the per-repetition distribution error is

$$\mathrm{cW1} = \frac{1}{2m}\sum_{i\in\mathcal{T}}\sum_{a\in\{0,1\}} W_1\Big(\widehat{P}_a(\cdot \mid X_i), P_a(\cdot \mid X_i)\Big). \quad (14)$$

The simulation evaluation approximates $\widehat{P}_a$ with 500 generated particles per query and arm; Setting IV uses 100 particles on 128 test images and the interventional mixture law described above. We do not report cW1 for IHDP or ACIC 2018 because the available records do not specify full conditional laws. Arm-specific cW1 uses the same test queries with only one summand, $m^{-1}\sum_i W_1(\widehat{P}_a, P_a)$, so overall cW1 is the average of the two arm-specific errors.

NGFM checkpoint selection uses factual validation CRPS: for a validation outcome y and two independent draws S, S' from the model's distribution under the observed arm, $\mathrm{CRPS}(\widehat{P}, y) = \mathbb{E}|S - y| - \frac{1}{2}\mathbb{E}|S - S'|$. Tables show means $\pm$ sample SD of repetition-level metrics. Figure 3 and Figure 5 instead show pointwise 95% confidence bands for each plotted mean from five matched repetitions. The machine-readable result inventory records the source and repetition count of each aggregated metric.

Endpoint-comparison protocol. Table 4 uses five repetitions at each sample size, with $d = 2$ and $p = 0.2$ in both settings. These are separate matched endpoint-ablation runs, with their own reference configuration rather than the fits in Table 1; within each repetition all four variants share the factual warm-up and its frozen donor graph. The donor-mean variant retains the original donor probabilities and replaces every borrowed endpoint by their weighted mean; the $r_i = 1$ variant replaces the query weight by one. The combined variant applies both changes. The new IHDP variants in Table 5 use training-fitted standardization of covariates and factual outcomes, restore predictions to outcome units before evaluation, and select checkpoints by factual validation CRPS. The NGFM row reuses the unstandardized reference fits from Tables 2–3 (rPEHE 0.330 ± 0.066, ATE error 0.133 ± 0.078). The three controls share standardized preprocessing, whereas this reference retains its recorded unstandardized preprocessing; this table therefore also differs in preprocessing between the reference and controls. The result records retain each partition, optimization seed, configuration, and preprocessing transform.

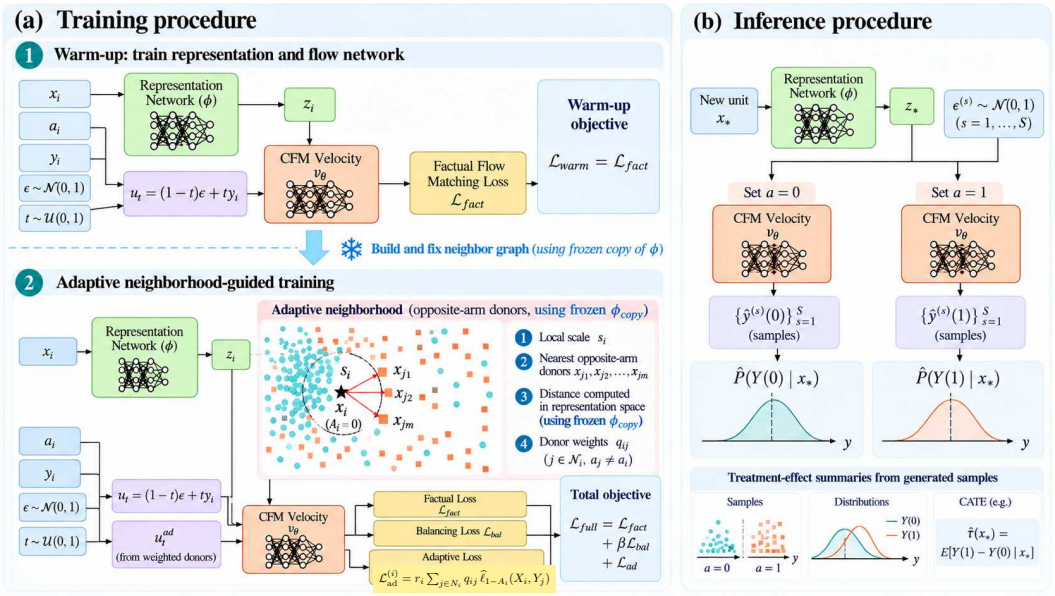

Figure 4: **Training and inference for NGFM.** Factual warm-up supplies the frozen metric. The live encoder and flow then optimize factual, alignment, and weighted all-donor losses. Gaussian source particles generate each marginal outcome law at inference. The auxiliary-loss box shows one query's sampled contribution $\mathcal{L}_{\mathrm{ad}}^{(i)}$, conditioned on X_i.

D Implementation and Training Flow

The adaptive encoder is a residual representation network. The velocity model has a FiLM-conditioned shared trunk with treatment-selected outputs. Setting I, Setting II, Setting III, and IHDP use representation width 64, latent dimension 10, and velocity width 64; ACIC uses 128, 32, and 128. Only the copy g is frozen after factual warm-up; the main encoder remains trainable. Donor indices, distances, reliability, and probabilities are stored for each training query. In the current weighted-all implementation, each selected donor produces its own path draw, the path losses are multiplied by the fixed probabilities, and the sum is multiplied by the fixed query reliability.

Weighted path evaluation. For $\epsilon \sim \mathcal{N}(0,1)$ and $t \sim \mathrm{Unif}[0,1]$, the path and expected squared loss are

$$u_t = (1-t)\epsilon + ty, \qquad \ell_a(x,y) = \mathbb{E}_{t,\epsilon}\big[v_\theta(u_t,t,\Phi_\phi(x),a) - (y-\epsilon)\big]^2, \qquad \mathcal{L}_{\mathrm{fact}} = \frac{1}{n}\sum_i \ell_{A_i}(X_i,Y_i).$$

For a minibatch query i, each valid donor $j \in N_i$ supplies an observed endpoint Y_j. The implementation draws a path time and Gaussian source for that endpoint, computes its squared velocity residual at the query representation and opposite treatment head, and multiplies it by the stored q_{ij}. The sum over donors is then multiplied by r_i. Conditional on the minibatch, the auxiliary term has Monte Carlo noise only from the path draws; donor identity is fully averaged. If fewer than K donors exist, padded positions are masked before summation. A query with no eligible opposite-arm donor has zero adaptive loss. The graph is fixed for all post-warm-up updates, so these weights do not receive gradients.

D.1 Flow Matching: Paths, Regression, and Generation

We give the conditional flow-matching construction used by NGFM (Lipman et al., 2023; Tong et al., 2024). Throughout this subsection, $c = (x,a)$ is a fixed covariate–treatment condition and T_c is its endpoint law. The network receives x through the trainable representation $\Phi_\phi(x)$. For factual supervision the endpoint is an observed outcome; for the combined objective it follows the factual–donor measure described in Appendix A.

A tractable path from noise to an outcome. Independently draw $Y \sim T_c$, $\epsilon \sim \mathcal{N}(0,1)$, and $t \sim \mathrm{Unif}[0,1]$. Define

$$U_t = (1-t)\epsilon + tY, \qquad \dot{U}_t = Y - \epsilon. \tag{15}$$

The path starts at Gaussian noise and ends at Y. Conditional on an endpoint $Y = y$, its density and velocity for $t < 1$ are

$$p_t(u \mid y) = \mathcal{N}\big(u; ty, (1-t)^2\big), \qquad b_t(u \mid y) = \frac{y-u}{1-t}. \tag{16}$$

On a sampled path, $b_t(U_t \mid Y) = Y - \epsilon$, so training uses the finite target in Eq. (15) without dividing by $1-t$. No trajectory integration or likelihood evaluation is required for this regression update.

Why endpoint regression learns the marginal velocity. Mixing over the endpoint law gives

$$p_t(u \mid c) = \int p_t(u \mid y) T_c(dy), \qquad v^*(u,t,c) = \frac{\int b_t(u \mid y) p_t(u \mid y) T_c(dy)}{p_t(u \mid c)} = \mathbb{E}[Y - \epsilon \mid U_t = u, t, c]. \tag{17}$$

The weights in this average depend on the current state through the conditional density. Integrating the conditional continuity equations yields $\partial_t p_t(u \mid c) + \partial_u\{p_t(u \mid c) v^*(u,t,c)\} = 0$ for $t < 1$. Thus the conditional expectation is the velocity associated with the mixture path, rather than a velocity evaluated at the mean endpoint.

For $v_{\theta,\phi}(u,t,c) = v_\theta(u,t,\Phi_\phi(x),a)$, define

$$\mathcal{J}_{\mathrm{CFM}}(\theta,\phi \mid c) = \mathbb{E}\big[\{v_{\theta,\phi}(U_t,t,c) - (Y-\epsilon)\}^2 \mid c\big], \tag{18}$$

$$\mathcal{J}_{\mathrm{FM}}(\theta,\phi \mid c) = \mathbb{E}\big[\{v_{\theta,\phi}(U_t,t,c) - v^*(U_t,t,c)\}^2 \mid c\big]. \tag{19}$$

Orthogonal projection gives

$$\mathcal{J}_{\mathrm{CFM}} = \mathcal{J}_{\mathrm{FM}} + \mathbb{E}[\mathrm{Var}(Y - \epsilon \mid U_t, t, c) \mid c]\,. \tag{20}$$

For a fixed endpoint law at the original condition $c = (x,a)$, the last term is independent of both network parameter sets. Under differentiation under the expectation, the two risks therefore have identical gradients. If the oracle is instead conditioned only on the learned representation, its target pools covariates and can change with ϕ; it must not be treated as a fixed oracle during representation updates. The covariate-conditioned excess risk above includes this representation approximation.

The weighted all-donor loss. For a minibatch B, draw an independent path time and source for each factual endpoint and each selected query–donor pair. If $\widehat{\ell}_{i,a}(y)$ denotes the resulting sampled squared residual conditioned on X_i, the implemented path-loss estimate is

$$\widehat{\mathcal{L}}_{\mathrm{paths}}(B) = \frac{1}{|B|} \sum_{i \in B} \left[\widehat{\ell}_{i,A_i}(Y_i) + r_i \sum_{j \in N_i} q_{ij} \widehat{\ell}_{i,1-A_i}(Y_j) \right]. \tag{21}$$

Its expectation over uniform minibatch and path sampling is $\mathcal{L}_{\mathrm{fact}} + \mathcal{L}_{\mathrm{ad}}$ for the fixed graph. The sum over all selected donors exactly averages donor identity, but path sampling remains stochastic. Substituting $\sum_j q_{ij} Y_j$ for the separate endpoints changes the endpoint law and is not an equivalent loss. The Sinkhorn term is added separately to regularize the live representation; it does not define an optimal-transport coupling between noise and endpoints. Our source and endpoint draws are independent during training.

Generation and numerical integration. For a fitted model and condition c, solve

$$\frac{dU_t}{dt} = v_\theta(U_t, t, \Phi_\phi(x), a), \qquad U_0 \sim \mathcal{N}(0,1). \tag{22}$$

For step size h, Heun's method first predicts $\widetilde{u} = u + h\, v_\theta(u,t,c)$ and then updates

$$u^+ = u + \frac{h}{2}\{v_\theta(u,t,c) + v_\theta(\widetilde{u}, t+h, c)\},$$

where the representation argument is suppressed. Fresh independent Gaussian draws provide ordinary generated samples. For the reported scalar-outcome evaluations, the implementation instead uses the deterministic source grid $u_0^{(m)} = \Phi_{\mathcal{N}}^{-1}((m - \frac{1}{2})/M)$, $m = 1, \ldots, M$, where $\Phi_{\mathcal{N}}$ is the standard-normal CDF. This quadrature reduces sampling variability when estimating means, CRPS, and cW1; the resulting particles are not independent Monte Carlo draws. Reusing a source grid across arms is a numerical convention, not an identified coupling of the two potential outcomes. Integration error and finite-particle error are distinct: the former is the $\nu_{a,x}$ term in the theory, while the latter arises when approximating the fitted law by finitely many particles.

For a discrete donor endpoint law, the conditional density is smooth for $t < 1$ and approaches that law at $t = 1$ in Wasserstein distance. The oracle velocity need not remain regular at the endpoint; Appendix B constructs its limiting characteristic. This does not require a finite, globally Lipschitz neural flow to reproduce atomic endpoints exactly. Its approximation and numerical errors are retained in the theoretical bound.

Algorithm 1 Neighborhood-guided flow matching

Require: Training and validation data; donor count K, rank $R \leq K$, balance coefficient β
1: Fit (ϕ, θ) on $\mathcal{L}_{\text{fact}}$ during factual warm-up.
2: Copy and freeze $g \leftarrow \Phi_\phi$; standardize $g(X)$ using training moments.
3: **for** each training query i **do**
4: Compute the pooled radius s_i excluding i; retrieve opposite-arm Top-K donors N_i.
5: Store r_i from Eq. (1) and normalized distance weights q_{ij}.
6: **end for**
7: **for** each joint-training minibatch B **do**
8: Draw factual paths and compute $\mathcal{L}_{\text{fact}}(B)$.
9: Draw a separate path for every valid (i, j) with $i \in B$, $j \in N_i$.
10: Compute $\mathcal{L}_{\text{ad}}(B) = |B|^{-1} \sum_{i \in B} r_i \sum_{j \in N_i} q_{ij} \ell_{1-A_i}(X_i, Y_j)$ using the sampled paths.
11: Compute $\mathcal{L}_{\text{bal}}(B)$ from the live treatment-arm representations.
12: Update (ϕ, θ) using $\mathcal{L}_{\text{fact}}(B) + \beta \mathcal{L}_{\text{bal}}(B) + \mathcal{L}_{\text{ad}}(B)$; keep g, N_i, r_i, q_{ij} fixed.
13: **end for**
14: Select the post-warm-up checkpoint with lowest factual validation CRPS.
15: For each test query and arm, integrate Gaussian source particles through the fitted flow.
16: **return** Conditional outcome samples and their distributional and mean-effect summaries.

Selection and generation. Optimization uses Adam (Kingma & Ba, 2015). Numerical integration follows the Runge–Kutta family of methods (Butcher, 1996); the saved benchmark protocol uses 32-step Heun integration. After warm-up, checkpoints are compared by factual validation continuous ranked probability score. The criterion uses only observed outcomes under their factual treatment. A fixed grid of Gaussian quantiles is then integrated through each treatment-conditioned velocity field; their empirical laws yield conditional means, effect estimates, and cW1 when conditional truth exists. Neither validation nor test observations enter graph construction. Donor indices and weights require at most $O(nK)$ storage; prediction does not use the graph, so its cost is independent of K.

Matched component protocol. Each ablation starts from the complete model's data partition, optimization seed, architecture, learning rate, batch size, training budget, and checkpoint rule. The no-balance variant sets $\beta = 0$ while retaining weighted donor guidance. The no-adaptive variant removes the donor loss while retaining Sinkhorn alignment. The original-X variant changes only the distance space used to form donors and reliability, retaining the same objective. Direct-X flow conditions the velocity on covariates directly and omits the warm-up graph and auxiliary losses. These definitions apply to Setting III, IHDP, and ACIC 2018; Table 3 reports the two benchmark ablations.

Paired effect of alignment. The improvement from adding $\mathcal{L}_{\text{bal}}$ is computed within each repeated run as the metric without alignment minus the metric of the full model. The average cW1 improvement in Setting III is 0.011, and the corresponding rPEHE and ATE-error improvements are 0.030 and 0.105. For IHDP, the paired rPEHE gain is 0.049; for ACIC 2018 it is 0.047. These calculations use the same test observations in each pair and take the error difference within a run before averaging. The full run-level values are retained in the result records.

Sensitivity run records. The nine parameter sweeps in Figure 5 use five matched repetitions at every level. The archived run-level records retain cW1, rPEHE, and ATE error for each checkpoint and make the plotted confidence intervals reproducible.

E Network architectures and computation

Table 6 summarizes the network skeletons used by the implementation and benchmark adapters. The tabular NGFM encoder has a SiLU input projection, two residual blocks with ReLU activations between their linear layers, and a unit-normalized latent projection. Its velocity network embeds time with 32 sinusoidal coordinates, combines time, representation, and treatment through FiLM and two gated residual blocks with aggregated skip outputs, and produces two treatment-selected scalar outputs. A frozen encoder copy defines neighborhoods. The live encoder and velocity network are jointly optimized after warm-up.

Table 6: **Network skeletons.** Widths describe the tabular simulation adapters unless a benchmark is specified. MLP denotes a multilayer perceptron.

Method	Architecture and implementation
NGFM	Two-block residual representation encoder; FiLM-conditioned velocity MLP; two arm outputs. Encoder/latent/velocity widths are 64/10/64 for Settings I–III and IHDP, and 128/32/128 for ACIC 2018. Setting IV uses the MLP architecture described below.
CEVAE	Variational latent-variable model with treatment and outcome decoders and inference networks; latent dimension 20, hidden width 128, three hidden layers.
DKLITE	Three-layer width-50 encoder and decoder; 50-dimensional features with treatment-specific Bayesian linear outcome posteriors.
FCCN	Separate invariant and treatment-specific feature generators, a propensity head, and two neural conditional CDF heads; sampling uses monotone CDF inversion.
NOFLITE	Width-32 representation network with three encoder layers, two balancing layers, and treatment-specific conditional priors; one SigmoidX flow with width-16 transforms and width-32 conditioning networks.
GDR-CNF	Conditional normalizing-flow nuisance and target models, width-64 conditioning networks and ten-bin transforms; propensity estimation and a doubly robust generative objective.
PO-Flow	FiLM-conditioned residual flow network with hidden width 128 and treatment-specific velocity outputs; Gaussian source and numerical ODE integration.
DiffPO	Conditional denoising network with four residual layers, 64 channels, 128-dimensional diffusion embedding, and two attention heads; 100 diffusion steps and an estimated propensity model.
TARNet / CFR / BWCFR	Shared MLP representation with two treatment heads; the adapters use width-128, two-layer MLPs and 32 representation outputs. CFR adds discrepancy regularization; BWCFR uses overlap weighting.
CrossNet-CMMD / PairNet	Shared representation and two treatment heads; conditional MMD constraints or observed cross-arm pair losses, respectively.
Orthogonal DR	Neural nuisance models for outcomes and propensity, followed by an effect regression on doubly robust pseudo-outcomes.
GANITE	Adversarial counterfactual generator and discriminator, followed by a two-head effect-prediction network.

The tabular NGFM configurations use Adam, factual warm-up, validation CRPS selection, and Heun integration. The supplementary code records the optimizer settings, training budgets, and neighborhood parameters. For n training units, precomputing a d_z-dimensional representation changes the arithmetic cost of an exhaustive distance search from $O(n^2 d_x)$ to $O(n^2 d_z)$, in addition to a one-time encoder pass; it does not remove the quadratic dependence on n. Only the selected donor graph needs to be retained after construction.

HC-MNIST implementation. Setting IV uses a two-hidden-layer SiLU encoder of width 128 with a 32-dimensional representation, followed by a three-hidden-layer width-128 velocity MLP conditioned on the representation, outcome, treatment, and time. After 500 factual-only updates, a frozen, coordinate-standardized copy of the encoder determines the pooled rank-32 radius and the 32 nearest opposite-arm donors for each training query. The rank-8 opposite-arm distance sets query reliability through Eq. (1); the pooled radius also sets the normalized donor probabilities. Each joint update combines factual flow loss, the reliability-weighted losses at all 32 observed donor outcomes, and debiased Sinkhorn divergence between the unit-normalized live minibatch representations of the two arms with coefficient $\beta = 1$. Training uses AdamW. Factual validation CRPS selects the post-warm-up checkpoint reported in Table 1. Predictions use Heun integration. The supplementary code provides the complete configurations.

All experiments were completed on servers equipped with NVIDIA L40 GPUs.

F Additional experimental results

The main table fixes Settings I–III at $n = 2{,}000$. Table 7 reports all available sample sizes for Setting III. It also separates distributional error by treatment arm. The supplementary code retains the complete tables for all 26 combinations of sample size, dimension, and treatment probability in Settings I–III, including machine-readable means, SDs, and individual-run metrics, together with NGFM benchmark, ablation, and sensitivity results.

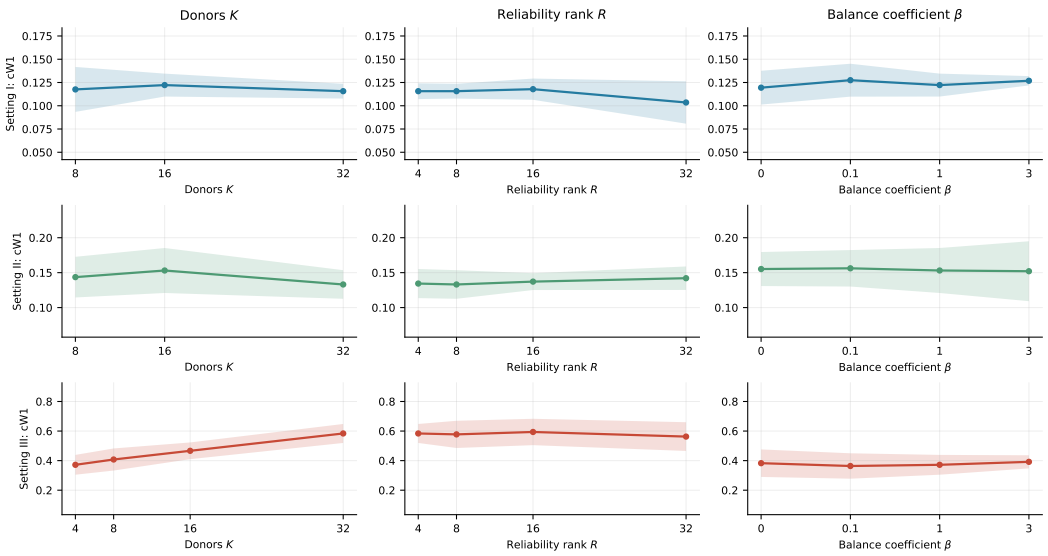

Figure 5: **Sensitivity of cW1.** Rows are Settings I–III; columns vary donor count K, reliability rank R, and balance strength β. Means and 95% confidence intervals use five matched repetitions. The R sweeps fix $K = 32$. Panels within each row share a vertical scale.

F.1 Neighborhood and alignment sensitivity

Figure 5 illustrates the locality–sampling tradeoff in Theorem 1. Larger donor pools need not improve recovery: in Setting III, increasing K from 4 to 32 raises cW1 from 0.372 to 0.584, as the neighborhood reaches less similar conditional laws. Setting II has a milder, nonmonotone response. Effective donor count and donor distance therefore matter together.

Reliability and alignment control different parts of the objective. Setting I benefits from a larger tested reliability rank, while Setting II is comparatively stable across both reliability and balance sweeps. Setting III favors a small positive balance coefficient, although its confidence bands overlap. These patterns support adapting guidance to the data mechanism; stronger borrowing or alignment is not uniformly better. Complete sweep values are included in the supplementary code.

F.2 Further distribution and effect results

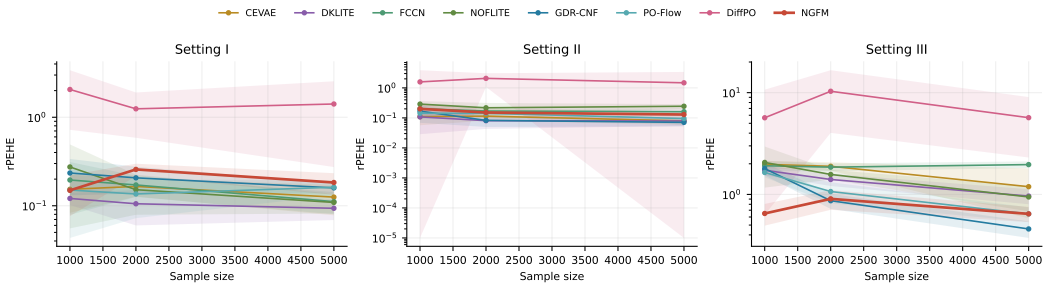

Figure 6: **Effect recovery across sample sizes.** rPEHE on the same mechanisms as Figure 3, with mean curves and 95% confidence bands. A logarithmic vertical scale retains DiffPO while making the other methods visible. Each repetition's PEHE is square-rooted before aggregation.

The rPEHE curves show that improvements in distributional recovery do not translate into identical rankings for mean effects. A model can approximate a mean contrast without reproducing the full outcome shape. The density comparisons below illustrate this distinction directly.

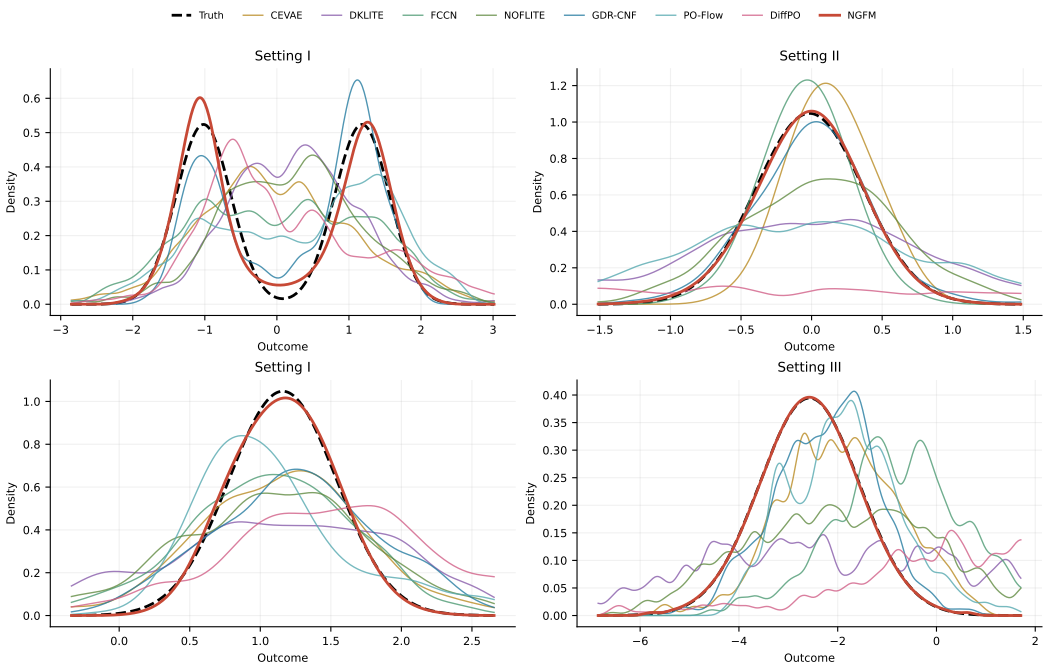

Figure 7: **Further conditional distribution comparisons.** Additional query–arm examples from Settings I–III, including DiffPO. All curves use the same Gaussian smoothing bandwidth. Axes focus on the true distribution's central range; generated samples outside that range are retained in the supplementary plotting inputs. NGFM closely follows the reference shape in these examples, while competing generators can smooth away modes, misplace the center, or spread mass too widely.

Table 7: Complete Setting III comparisons. Entries are mean ± SD; bold marks the lowest mean within each condition and metric.

Method	cW1	cW1 ($A=0$)	cW1 ($A=1$)	rPEHE	ATE error
			Setting III: $n = 1000$, $d = 2$		
CEVAE	1.403±0.167	1.535±0.255	1.270±0.109	1.975±0.134	0.464±0.314
DKLITE	1.920±0.108	2.020±0.112	1.819±0.118	1.728±0.153	**0.124±0.108**
FCCN	1.071±0.115	1.175±0.099	0.968±0.212	1.913±0.179	0.435±0.262
NOFLITE	1.434±0.374	1.464±0.069	1.405±0.772	2.062±0.722	0.383±0.242
GDR-CNF	0.842±0.078	1.043±0.151	0.640±0.143	1.794±0.207	0.207±0.110
PO-Flow	1.014±0.129	1.114±0.184	0.913±0.088	1.644±0.138	0.182±0.151
DiffPO	9.002±4.458	11.095±6.636	6.909±7.486	5.665±4.055	3.591±4.922
NGFM	**0.372±0.054**	**0.432±0.070**	**0.311±0.038**	**0.650±0.126**	0.156±0.156
			Setting III: $n = 2000$, $d = 2$		
CEVAE	1.154±0.050	1.182±0.113	1.125±0.039	1.890±0.145	0.175±0.154
DKLITE	1.818±0.035	1.858±0.036	1.778±0.050	1.402±0.155	0.149±0.100
FCCN	1.026±0.123	1.051±0.092	1.002±0.159	1.853±0.131	0.327±0.192
NOFLITE	1.134±0.048	1.321±0.056	0.946±0.077	1.572±0.140	0.225±0.126
GDR-CNF	0.406±0.064	**0.465±0.094**	0.347±0.068	**0.868±0.119**	0.108±0.099
PO-Flow	0.658±0.165	0.721±0.190	0.595±0.228	1.068±0.201	0.235±0.157
DiffPO	10.200±3.940	11.588±5.678	8.811±7.610	10.320±5.070	9.447±5.178
NGFM	**0.389±0.047**	0.472±0.102	**0.307±0.028**	0.905±0.160	**0.094±0.104**

Method	cW1	cW1 ($A = 0$)	cW1 ($A = 1$)	rPEHE	ATE error
		Setting III: $n = 5000$**,** $d = 2$			
CEVAE	0.701±0.303	0.722±0.316	0.680±0.294	1.192±0.491	0.203±0.145
DKLITE	1.766±0.032	1.781±0.033	1.751±0.032	0.963±0.130	0.072±0.044
FCCN	1.049±0.094	1.177±0.093	0.922±0.175	1.964±0.119	0.263±0.182
NOFLITE	0.906±0.051	0.959±0.106	0.854±0.031	0.944±0.212	0.211±0.118
GDR-CNF	**0.228±0.028**	0.258±0.045	**0.197±0.014**	**0.458±0.068**	0.054±0.054
PO-Flow	0.406±0.039	0.492±0.057	0.320±0.055	0.644±0.084	**0.050±0.035**
DiffPO	6.336±3.718	9.989±6.856	2.683±0.717	5.678±2.717	4.446±3.267
NGFM	0.238±0.021	**0.241±0.019**	0.236±0.038	0.644±0.089	0.121±0.082

F.3 HC-MNIST WITH HIDDEN CONFOUNDING

The same image and outcome mechanism with $\Gamma = e$ makes treatment depend on the unobserved U. Table 8 reports the five-run results under this additional test.

Table 8: **HC-MNIST with hidden confounding ($\Gamma = e$).** Mean ± sample SD over five runs; lower is better. Bold marks the smallest mean in each metric.

Method	cW1	rPEHE	ATE error
CEVAE	1.529±0.065	3.303±0.245	1.642±0.260
DKLITE	2.228±0.007	2.370±0.024	1.650±0.027
FCCN	1.519±0.064	2.536±0.207	**1.414±0.325**
NOFLITE	1.766±0.011	3.376±0.020	1.554±0.040
GDR-CNF	1.385±0.038	2.875±0.099	1.590±0.053
PO-Flow	2.072±0.042	**2.060±0.031**	2.049±0.023
DiffPO	2.150±0.056	2.127±0.049	2.105±0.046
NGFM (ours)	**1.142±0.026**	2.416±0.104	1.870±0.130

G METHOD LIMITATIONS

Causal identification and representation information. NGFM relies on conditional exchangeability and positivity given the observed covariates. Neither a flexible generator nor neighborhood borrowing identifies effects under unmeasured confounding without further assumptions. Main-table Setting IV uses $\Gamma = 1$ and satisfies conditional exchangeability; the $\Gamma = e$ experiment in Appendix F.3 evaluates robustness when this condition fails. A learned metric can also discard outcome-relevant information: global treatment-group alignment alone does not ensure representation sufficiency (Melnychuk et al., 2024b). Donor guidance is most useful when nearby representations correspond to similar conditional outcome laws.

Scope of the theoretical guarantees. The empirical endpoint-law identity holds for the implemented outcome-trained graph. The conditional donor concentration bound additionally assumes a metric fitted independently of donor outcomes, as in a construction that trains the metric on a separate sample independent of the entire donor bank. Factual warm-up on the same outcomes does not establish this independence. The population recovery conclusion therefore applies under the assumptions stated in Section 4; it is not an unconditional consistency guarantee for every fitted graph.

Sparse support and computational cost. Reliability attenuates borrowing from distant donors but cannot create observations in an unsupported treatment region. Larger neighborhoods can introduce approximation bias, and a frozen warm-up metric cannot adapt its donor identities during later training. Compact representations reduce per-distance arithmetic, while exhaustive graph construction still scales quadratically with training-set size. The weighted all-donor objective also increases training work with K, although inference requires no neighbor retrieval.

Distributional target. NGFM estimates the two conditional marginal potential-outcome laws. Their means identify the conditional average effect under the stated assumptions, but the marginals alone do not identify the joint law of $(Y(0), Y(1))$ or the distribution of individual treatment effects. Additional dependence assumptions would be required for those targets.

H FURTHER CONTEXT AND DISTINCTIONS

Observational evidence and applications. Clinical registries and healthcare-utilization records illustrate the motivation for learning from observational outcomes. The INTERMACS report describes outcomes among recipients of ventricular-assist devices (Kirklin et al., 2010), while Corrao et al. (2012) study adjustment for unmeasured confounders using external information. These applications motivate careful causal assumptions; a flexible outcome generator alone cannot replace confounder control.

Related generative tasks. Diffusion models have also been developed for interventional and counterfactual queries (Chao et al., 2023) and for generating data to assist effect estimation (Chen et al., 2025). Structural causal modeling and causal discovery include graph autoencoders and diffusion-based topological ordering (Sánchez-Martin et al., 2022; Sanchez et al., 2023). These tasks use different inputs or target different objects from the two conditional marginal laws studied here. The word *causal* also appears in autoregressive generative architectures, such as causal diffusion transformers (Deng et al., 2024); an autoregressive ordering is distinct from treatment-effect identification.

Representation learning and diagnostics. For image covariates, cross-source visual representation learning (He et al., 2022) is a possible ingredient for a future feature extractor. Its X-Learner is a visual representation method, distinct from the causal X-learner of Künzel et al. (2019). Embedding visualizations such as UMAP (McInnes et al., 2018) can display learned geometry, but distributional metrics and matched ablations are needed to assess the effect on conditional outcome recovery. Our current evaluation uses those quantitative comparisons.

Transport as a distance and as a regularizer. Classical transport formulations originate with Monge (1781) and Kantorovich (2006); Villani et al. (2008) provides the metric and duality background for the Wasserstein arguments used here. Entropic transport enables efficient Sinkhorn computation (Cuturi, 2013), and debiasing removes the entropic self-interaction terms (Feydy et al., 2019). Distributionally robust regression also uses Wasserstein geometry (Liu et al., 2025), but its uncertainty-set objective differs from both our representation penalty and our local endpoint law. NGFM's theory bounds the latter in W_1; it does not equate the squared-cost Sinkhorn penalty with that error.