LEARNING OUTCOMES OF JOINT INTERVENTIONS VIA SIM2REAL CAUSAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world decisions rarely involve a single lever. Drug combinations, multiplexed gene perturbations, and multi-component policies require predicting outcomes when several variables are manipulated at once, often with limited or no interventional data. Motivated by these demands, we propose a new paradigm: learning outcomes of joint interventions (OJIs) with a single pre-trained model that infers them from observational data alone and flexibly exploits interventional data when available. In their original implementations, causal foundation models (CFMs) such as CausalPFN and Do-PFN address single-variable interventions using observational-only pre-training contexts. We realize this paradigm with **JOINT (Joint Outcome Inference via Neural Transfer)**, a CFM pre-trained entirely on diverse causal systems simulated via next-token generation. JOINT is pre-trained on contexts that randomly include interventional samples to align pre-training with deployment, reducing mean average treatment effect (ATE) error by approximately 39% relative to observational-only pre-training when interventional examples are available. For interventions on up to ten variables, JOINT estimates individual treatment effects more accurately and achieves lower ATE error than competing causal and tabular foundation models. It remains robust under partial observability and generalizes to graphs twice the size of those seen during pre-training. Without fine-tuning, JOINT attains the lowest ATE error on real-world genetic and chemical perturbation datasets, improving on the strongest baseline by approximately 13% and demonstrating effective Sim2Real transfer.

1 INTRODUCTION

Many consequential decisions require manipulating several variables at once, from combination therapies (Srivatsan et al., 2020) and multiplexed gene perturbations (Norman et al., 2019) to economic policies (Athey & Wager, 2021) and public health programs (Null et al., 2018). In each case, practitioners need to anticipate the outcomes of joint interventions (OJIs) to decide which combination of variables to act on. However, causal relationships differ across systems and are typically unknown (Schölkopf et al., 2021). It is often unclear which variables are drivers and which are responses, and the question of interest is simply: if we change some variables, what else changes? Task-specific causal estimators require substantial time and domain knowledge to develop and may not readily transfer across settings (Pearl & Bareinboim, 2014). The available evidence also varies widely: experiments can be costly or infeasible, so practitioners may have only observational records in one setting and a small number of experiments in another (Braga et al., 2025).

Motivated by these demands, we propose a new paradigm for learning OJIs via causal foundation models (CFMs) (Figure 1). A single pre-trained model should support interventions on different subsets of variables and predict the response of any other variable, without requiring prior knowledge of whether it lies downstream of the intervention. It should infer OJIs from observational context alone and exploit interventional evidence when available, applying to previously unseen causal systems without parameter updates. Because real-world systems with known OJIs are rare (Roohani et al., 2024; Stechemesser et al., 2024), we pursue this paradigm through Sim2Real transfer: learning from simulated causal systems and applying the resulting model directly to real data.

CFMs (Stith et al., 2026) such as CausalPFN (Balazadeh et al., 2025) and Do-PFN (Robertson et al., 2025) amortize causal effect estimation through in-context learning, building on tabular foundation

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.

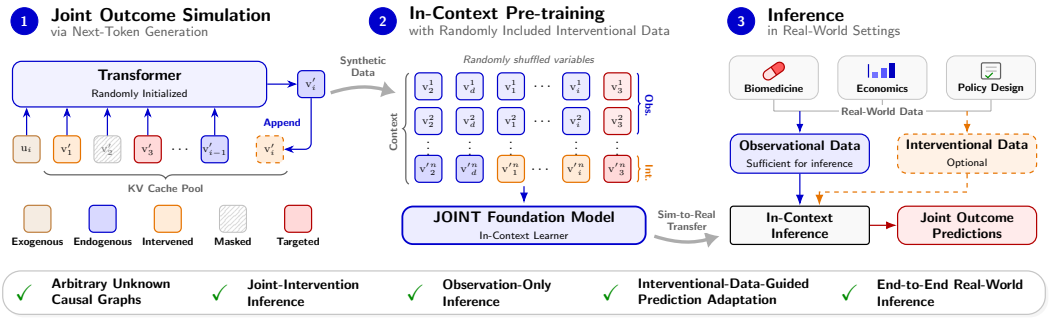

Figure 1: **Overview of the proposed paradigm.** **(1)** Next-token generation simulates outcomes of joint interventions (OJIs). **(2)** Pre-training JOINT by randomly including interventional examples in the context supports learning across heterogeneous causal data and aligns pre-training with real-world inference. **(3)** JOINT enables Sim2Real transfer to OJI prediction with or without interventional data. Obs. and Int. denote observational and interventional data, respectively.

models like TabPFN (Hollmann et al., 2025). In their original implementations, CausalPFN assumes strong ignorability for a single binary treatment, while Do-PFN supports varying causal graphs but uses a designated scalar treatment and an additive-noise synthetic prior. Both use observational-only pre-training contexts and are not explicitly trained to incorporate interventional examples into the context. Realizing the proposed paradigm therefore requires both an efficient way to simulate diverse causal systems under joint interventions and a pre-training strategy that prepares one model for varying variable roles and available evidence.

We address both requirements. For simulation, we cast endogenous variable generation as next-token generation. In our Transformer (Vaswani et al., 2017) implementation, each variable is generated by attending only to its parents, with non-parent variables masked out, and interventions overwrite selected values before generation proceeds to downstream variables. This separates causal structure from functional mechanisms and enables the synthesis of diverse OJIs data with non-additive dependencies.

For learning, we introduce JOINT (Joint Outcome Inference via Neural Transfer), a CFM pre-trained on these simulated OJIs using prompts that always contain observational context and are randomly augmented with interventional examples. This dual-regime design provides contextual guidance for learning across heterogeneous causal systems. Intervention and target-role conditioning specify the intervened variables and indicate whether the target is known to be a downstream outcome. Together, these components allow JOINT to infer OJIs from observational context alone and improve its predictions when interventional examples are available.

Experiments show that JOINT estimates individual treatment effects more accurately and achieves lower average treatment effect (ATE) error than competing causal and tabular foundation models in both zero-shot settings (observational data only) and few-shot settings (interventional examples included), maintains strong performance under partial observability in query variables, and generalizes to graphs larger than those seen during pre-training. On two real-world perturbation datasets, Norman (Norman et al., 2019) and ComboSciPlex (Srivatsan et al., 2020), JOINT achieves the lowest mean ATE error among the evaluated methods, supporting Sim2Real transfer across genetic and chemical perturbations. Dual-regime pre-training reduces mean ATE error by $\sim 39.2\%$ across the evaluated few-shot settings compared with observational-only pre-training. Our contributions:

1. We propose a new paradigm for learning OJIs via CFMs through Sim2Real transfer, enabling a single pre-trained model to accommodate different intervention subsets and prediction targets across causal systems without parameter updates.
2. We instantiate this paradigm with JOINT and a dual-regime pre-training strategy that randomly includes interventional examples in contexts. This design guides learning across heterogeneous causal systems and aligns training with the interventional evidence available at inference time.
3. Extensive experiments demonstrate JOINT's accuracy, robustness, and generalization in OJI estimation, validate the benefit of dual-regime pre-training, and establish effective transfer to real-world genetic and chemical perturbation datasets.

2 Background and Related Work

Tabular Foundation Models. Tabular foundation models such as TabPFN (Hollmann et al., 2025) learn from synthetic datasets during pre-training and perform in-context prediction on new datasets without dataset-specific parameter updates. Under the Prior-Data Fitted Network (PFN) framework, a model learns to predict held-out query targets from their covariates and labeled context samples. This procedure approximates Bayesian posterior predictive inference under the pre-training prior, transferring the computational cost of learning to pre-training (Müller et al., 2021). Subsequent models such as TabICL (QU et al., 2025) extend tabular in-context learning to larger datasets. Both TabPFN and TabICL use multilayer perceptrons (MLPs) as a core component of data synthesis, with neurons representing causal variables and network connectivity determining the causal structure. This couples graph structure with the architecture used to parameterize the generation mechanisms, making it difficult to explicitly control the causal structure among selected variables and flexibly specify diverse interventions. Details about TabPFN and TabICL are provided in Appendix B.

Interventions and Outcomes of Joint Interventions (OJIs). Consider a causally ordered set $\mathcal{V}$ of d endogenous variables, with corresponding random vector $\mathbf{V} \in \mathbb{R}^d$ and realization $\mathbf{v} \in \mathbb{R}^d$. An intervention $\mathrm{do}(\mathbf{V}_\mathcal{I} = \mathbf{v}_\mathcal{I})$ on a nonempty index set $\mathcal{I} \subseteq \{1, \ldots, d\}$ replaces the generation mechanisms of the selected variables with fixed assignments $\mathbf{v}_\mathcal{I}$, while leaving all other mechanisms and the distribution of exogenous noise unchanged (Pearl, 2009). The intervention-set size $|\mathcal{I}|$ counts all variables whose values are fixed by the intervention. The values of downstream variables may nevertheless change in response to the intervention. For a target variable V_y with $y \notin \mathcal{I}$, its outcome distribution under this intervention is

$$p(\mathrm{V}_y \mid \mathrm{do}(\mathbf{V}_\mathcal{I} = \mathbf{v}_\mathcal{I})) . \tag{1}$$

Recent work, Joint Intervention Learning (JIL) by Kekić et al. (2025) estimates effects of joint interventions using single-variable interventional data under nonlinear additive outcome assumptions. Their formulation avoids requiring data of OJIs. Details about JIL are provided in Appendix B.

Causal In-Context Learning. Building on the distinction between observational and interventional prediction, recent works such as CausalPFN (Balazadeh et al., 2025) and Do-PFN (Robertson et al., 2025) amortize causal effect estimation through pre-training on simulated causal tasks. These approaches pre-train tabular Transformer models using observational contexts containing treatment, covariate, and outcome values, with simulated causal quantities serving as query targets. This enables inference on unseen datasets without dataset-specific parameter updates. However, their original implementations remain constrained by specific causal assumptions and intervention settings.

For CausalPFN, the original implementation considers a single binary treatment T, covariates $\mathbf{X}$, and an outcome Y under strong ignorability, requiring unconfoundedness given $\mathbf{X}$ and treatment overlap. Its prior enforces these conditions, and its treatment interface does not directly represent joint interventions on arbitrary variable subsets.

For Do-PFN, the synthetic prior uses additive noise models, where each endogenous variable is generated as $\mathrm{V}_i = f_i(\mathbf{V}_{\mathrm{PA}_i}) + \mathrm{U}_i$, with PA_i denoting its parents and U_i an independent exogenous noise term. While f_i can be nonlinear, the noise enters additively rather than interacting with parent values. Do-PFN samples diverse causal graphs and accommodates some settings with unobserved confounding, but its original implementation places a designated scalar treatment T in the first input column and is designed for interventions $\mathrm{do}(\mathrm{T} = \mathrm{t})$ rather than joint interventions on arbitrary subsets of variables. Further training details of CausalPFN and DoPFN are provided in Appendix B.

Counterfactuals and Twin Networks. Constructing training data for OJI prediction involves comparing outcomes across intervention scenarios. Independently sampled interventional outcomes generally differ in both intervention assignments and exogenous noise, whereas counterfactual comparisons require evaluating the same underlying unit under alternative interventions. Twin networks represent this relationship using two copies of a causal system that share exogenous variables (Balke & Pearl, 2022). By holding the realized exogenous noise fixed across the two copies, paired interventional outcomes can be generated whose differences reflect only changes in intervention assignments. We adapt this construction to generate paired training examples for pretraining, as described in the next section.

3 METHODS

We implement the proposed paradigm with Transformer-based causal data synthesis and dual-regime in-context pre-training of JOINT. The synthesis stage generates observational and paired interventional samples across heterogeneous causal systems, providing known intervention outcomes for supervision. JOINT is then pre-trained to predict these outcomes from contexts that optionally include interventional examples, enabling inference with or without interventional evidence. Crucially, the causal graphs guide simulation but are never provided to JOINT.

3.1 CAUSAL DATA SYNTHESIS VIA NEXT-TOKEN GENERATION

Attention Mechanisms and Causal Masking. To implement causal structure-constrained next-token generation, we generate variables in topological order and restrict each variable to attend only to its causal parents. An additive mask is applied to the attention logits before softmax, with 0 permitting attention and $-\infty$ blocking it. The temperature $\tau > 0$ controls attention sharpness. Unlike standard next-token generation, which only blocks future tokens, our mask additionally excludes non-parent variables, enforcing the causal structure specified by $\mathcal{G}$. The key–value representations of preceding variables, denoted by $\mathbf{h}_j^K$ and $\mathbf{h}_j^V$, are cached for sequential generation.

Endogenous Variable Generation. Unlike MLP-based synthetic priors such as those used in TabPFN (Hollmann et al., 2025) and TabICL (QU et al., 2025), where the causal structure is tied to the feed-forward network topology, we explicitly separate causal structure from functional mechanisms. A randomly initialized Transformer $g_{\boldsymbol{\Phi}}$ models nonlinear interactions among parent values and exogenous noise, while attention masks $\mathbf{m}_i \in \{0, -\infty\}^{i-1}$ define each variable's parents during topological generation. Each mask entry is 0 with probability p_{edge} and $-\infty$ otherwise, and both weights $\boldsymbol{\Phi}$ and masks remain fixed within each causal system. Given noise $\mathrm{u}_i \sim p(\mathrm{U}_i)$ and cached representations $\mathcal{KV}_{<i}$ of preceding values, we generate

$$\mathrm{v}_i = g_{\boldsymbol{\Phi}}(\mathrm{u}_i \mid \mathcal{KV}_{<i}, \mathbf{m}_i). \tag{2}$$

For root variables without parents, $\mathbf{m}_i$ contains only $-\infty$ entries, ensuring that their generated values depend solely on exogenous noise. Under interventions, selected values are overwritten by prescribed assignments before being written to the KV, allowing intervention effects to propagate downstream. Algorithm 1 summarizes the procedure, with further details in Appendix C.

Algorithm 1 Endogenous Variable Generation with a Transformer and KV Cache

Require: Dimension d, edge probability p_{edge}, intervention set $\mathcal{I}$ ($\mathcal{I} = \emptyset$ for purely observational samples) and assignments $\mathbf{v}_{\mathcal{I}}$.
Ensure: Realized values $\mathbf{v} \in \mathbb{R}^d$.
1: Randomly initialize Transformer weights $\boldsymbol{\Phi}$ and attention masks $\{\mathbf{m}_i\}_{i=1}^d$, where $\mathbf{m}_i \in \{0, -\infty\}^{i-1}$, with entries 0 with probability p_{edge} and $-\infty$ otherwise.
2: Initialize $\mathcal{KV}_{<1} \leftarrow \emptyset$.
3: **for** $i = 1$ to d **do**
4: Sample exogenous noise $\mathrm{u}_i \sim p(\mathrm{U}_i)$.
5: $\mathrm{v}_i \leftarrow g_{\boldsymbol{\Phi}}(\mathrm{u}_i \mid \mathcal{KV}_{<i}, \mathbf{m}_i)$. ▷ Noise only if no parents
6: **if** $i \in \mathcal{I}$ **then**
7: Overwrite v_i with its prescribed value in $\mathbf{v}_{\mathcal{I}}$.
8: **end if**
9: $(\mathbf{h}_i^K, \mathbf{h}_i^V) \leftarrow$ Extract-KV$(g_{\boldsymbol{\Phi}}(\mathrm{v}_i))$.
10: $\mathcal{KV}_{<i+1} \leftarrow \mathcal{KV}_{<i} \cup \{(\mathbf{h}_i^K, \mathbf{h}_i^V)\}$.
11: **end for**
12: **return** $\mathbf{v}$.

Observational and Counterfactual Simulation. Within each causal system, Transformer parameters and masks are fixed. We generate observational samples $\mathcal{D}^{\text{obs}} = \{\mathbf{v}^{(o)}\}_{o=1}^{N^{\text{obs}}}$ without interventions using independent noise draws $\mathbf{u}^{(o)} \sim p(\mathbf{U})$. For paired samples $\mathcal{D}^{\text{twin}} = \{(\mathbf{v}'^{(p)}, \mathbf{v}''^{(p)})\}_{p=1}^{N^{\text{twin}}}$, each pair shares $\mathbf{u}^{(p)} \sim p(\mathbf{U})$ under complementary assignments to a sampled intervention set $\mathcal{I}$. We independently draw $\mathrm{v}_j'^{(p)} \sim \text{Bernoulli}(0.5)$ and set $\mathrm{v}_j''^{(p)} = 1 - \mathrm{v}_j'^{(p)}$

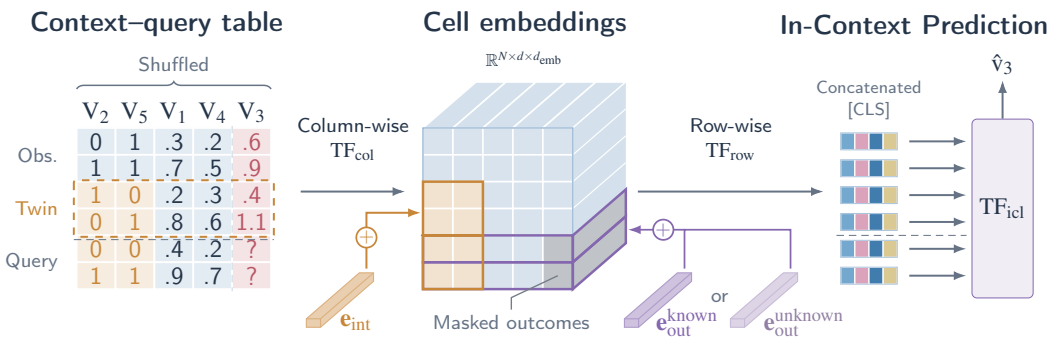

Figure 2: **JOINT framework.** During pre-training, the context contains observational samples (Obs.) and optional interventional samples from a twin network, while query outcomes are masked. Columns are shuffled. V_2 and V_5 are intervention variables, and V_3 is the prediction target. Column-wise encoding ($\mathrm{TF}_{\mathrm{col}}$) produces cell representations, augmented with $\mathbf{e}_{\mathrm{int}}$ at directly intervened cells and either $\mathbf{e}_{\mathrm{out}}^{\mathrm{known}}$ or $\mathbf{e}_{\mathrm{out}}^{\mathrm{unknown}}$ across each query row to indicate the available target-role information. Row-wise aggregation ($\mathrm{TF}_{\mathrm{row}}$) concatenates CLS representations into a sample representation, and the in-context Transformer ($\mathrm{TF}_{\mathrm{icl}}$) predicts the masked query outcomes $\widehat{\mathrm{v}}_3$ using the context.

for $j \in \mathcal{I}$, yielding

$$\mathbf{v}'^{(p)} = g_{\boldsymbol{\Phi}}\left(\mathbf{u}^{(p)} \mid \mathrm{do}(\mathbf{V}_{\mathcal{I}} = \mathbf{v}_{\mathcal{I}}'^{(p)})\right), \qquad \mathbf{v}''^{(p)} = g_{\boldsymbol{\Phi}}\left(\mathbf{u}^{(p)} \mid \mathrm{do}(\mathbf{V}_{\mathcal{I}} = \mathbf{v}_{\mathcal{I}}''^{(p)})\right). \tag{3}$$

Shared noise ensures that paired differences reflect only the intervention change and its downstream effects.

3.2 Model Construction

Figure 2 illustrates the JOINT framework, including context–query organization, intervention and target-role conditioning, and in-context prediction of OJIs.

Base Architecture. JOINT adopts the TabICL backbone (QU et al., 2025), comprising column-wise encoding, row-wise aggregation, and in-context prediction (Figure 2). It accommodates varying numbers of variables without fixed column identities across datasets. Combined with the conditioning described below, it supports different intervention subsets and prediction targets.

Context-Query Split. We partition samples from $\mathcal{D}^{\mathrm{obs}}$ and $\mathcal{D}^{\mathrm{twin}}$ into a context $\mathcal{C}$ and held-out queries $\mathcal{Q}$. Pre-training randomly uses either zero-shot contexts (observation data only), $\mathcal{C} = \mathcal{D}_{\mathrm{ctx}}^{\mathrm{obs}} \subseteq \mathcal{D}^{\mathrm{obs}}$, or few-shot contexts (interventional examples included), $\mathcal{C} = \mathcal{D}_{\mathrm{ctx}}^{\mathrm{obs}} \cup \mathcal{D}_{\mathrm{ctx}}^{\mathrm{int}}$, where $\mathcal{D}_{\mathrm{ctx}}^{\mathrm{int}}$ contains samples from twin-network pairs. For each interventional query $\mathbf{v}^{(q)} \in \mathcal{Q}$, the target $\mathrm{v}_y^{(q)}$ is omitted from the model input but retained as the ground-truth label for computing the pre-training loss; the available non-target values $\mathbf{v}_{\setminus y}^{(q)}$ realized under the same intervention serve as input in both regimes. Context samples retain their target values. At inference time, interventional context samples need not be paired, and no model parameters are updated.

Intervention and Target-Role Conditioning. To equip JOINT with intervention-aware inference capabilities, we add learnable conditioning embeddings to the cell representations produced by $\mathrm{TF}_{\mathrm{col}}$ before row-wise aggregation, encoding both intervention locations and the available information about the target's causal role. Let $\mathbf{h}_i^{(p)}$ denote the representation of variable V_i in sample p. For each interventional sample in either the context or query set, we add $\mathbf{e}_{\mathrm{int}}$ only to the directly intervened variables:

$$\mathbf{h}_i^{(p)} \leftarrow \mathbf{h}_i^{(p)} + \mathbf{e}_{\mathrm{int}}, \qquad i \in \mathcal{I}^{(p)}. \tag{4}$$

All other cells receive no intervention embedding. This conditioning distinguishes assigned values from system-generated values and identifies the intervention subset.

Unlike intervention status, which is specific to individual variables, the target's causal role provides prior information about the query as a whole. Therefore, for each query sample, we add a target-role embedding to all variable representations, indicating whether the target is known to be a downstream outcome or its topological role is unknown:

$$\mathbf{h}_i^{(q)} \leftarrow \mathbf{h}_i^{(q)} + \begin{cases} \mathbf{e}_{\text{out}}^{\text{known}}, & \text{if } \mathrm{V}_y \text{ is known to be a downstream outcome,} \\ \mathbf{e}_{\text{out}}^{\text{unknown}}, & \text{if its topological role is unknown,} \end{cases} \qquad i = 1, \dots, d. \tag{5}$$

Optimization Objective. Given a context $\mathcal{C} \in \{\mathcal{D}_{\text{ctx}}^{\text{obs}}, \mathcal{D}_{\text{ctx}}^{\text{obs}} \cup \mathcal{D}_{\text{ctx}}^{\text{int}}\}$ and query values $\mathbf{v}_{\setminus y}^{(q)}$ realized under the corresponding intervention, JOINT predicts the conditional density $p_\theta(\mathrm{v}_y^{(q)} \mid \mathbf{v}_{\setminus y}^{(q)}, \mathcal{I}^{(q)}, \mathbf{e}_{\text{out}}^{(q)}, \mathcal{C})$, where θ denotes its learnable parameters. Here, $\mathcal{I}^{(q)}$ specifies the query intervention set, $\mathbf{e}_{\text{out}}^{(q)} \in \{\mathbf{e}_{\text{out}}^{\text{known}}, \mathbf{e}_{\text{out}}^{\text{unknown}}\}$ denotes the embedding corresponding to the randomly sampled target-role setting, and $\mathcal{C}$ includes the intervention annotations of its samples. The intervention assignments are contained in $\mathbf{v}_{\setminus y}^{(q)}$. The same objective applies under both context regimes. We represent the predictive density using a discretized histogram (Müller et al., 2021) and train the model by minimizing the negative log-likelihood of the held-out query targets:

$$\mathcal{L}(\theta) = \mathbb{E}_{\psi \sim p(\psi),\, \mathcal{C},\, \mathcal{Q}} \left[-\frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \log p_\theta \left(\mathrm{v}_y^{(q)} \mid \mathbf{v}_{\setminus y}^{(q)}, \mathcal{I}^{(q)}, \mathbf{e}_{\text{out}}^{(q)}, \mathcal{C} \right) \right], \tag{6}$$

where ψ specifies the sampled causal data generator configuration, and $\mathcal{C}$ and $\mathcal{Q}$ are constructed from the resulting samples. Here, $\mathbf{v}_{\setminus y}^{(q)}$ denotes the available non-target query values, which may include only a subset of the remaining variables. At inference time, the predicted outcome is the mean of the learned conditional distribution:

$$\hat{\mathrm{v}}_y^{(q)} = \mathbb{E}_{p_\theta} \left[\mathrm{V}_y \mid \mathbf{v}_{\setminus y}^{(q)}, \mathcal{I}^{(q)}, \mathbf{e}_{\text{out}}^{(q)}, \mathcal{C} \right]. \tag{7}$$

3.3 Evaluation

All $2^{|\mathcal{I}|} - 1$ non-control binary assignments are evaluated, with N_q query instances for each joint assignment $\mathbf{v}_{\mathcal{I}}$. Let $\mathrm{v}_y^{(i)}$ and $\hat{\mathrm{v}}_y^{(i)}$ denote the realized and predicted outcomes for instance i; the subscript ctrl denotes outcomes under the all-zero control assignment $\mathbf{0}_{\mathcal{I}}$. The same N_q noise realizations are used across assignments. Omitting assignment dependence from the notation, the true and predicted individual effects are

$$\tau^{(i)} = \mathrm{v}_y^{(i)} - \mathrm{v}_{y,\mathrm{ctrl}}^{(i)}, \qquad \hat{\tau}^{(i)} = \hat{\mathrm{v}}_y^{(i)} - \hat{\mathrm{v}}_{y,\mathrm{ctrl}}^{(i)}. \tag{8}$$

We evaluate individual-level effect estimation using precision in estimation of heterogeneous effect (PEHE) (Hill, 2011) and population-level effect estimation using absolute error in the average treatment effect (ATE error):

$$\text{PEHE} = \sqrt{\frac{1}{N_q} \sum_{i=1}^{N_q} \left(\tau^{(i)} - \hat{\tau}^{(i)}\right)^2}, \qquad \text{ATE Error} = \left| \frac{1}{N_q} \sum_{i=1}^{N_q} \left(\tau^{(i)} - \hat{\tau}^{(i)}\right) \right|. \tag{9}$$

Both metrics are used in Sections 4.1 and 4.3. For the real-world perturbation datasets in Section 4.2, we use ATE error computed from differences in condition means because paired outcomes for the same unit under alternative interventions are unavailable, precluding direct evaluation of PEHE. Further details and supplementary ATE relative error (ATE-RE) are provided in Appendix D.

4 Experiments

We evaluate JOINT against TabICL, CausalPFN, and Do-PFN. Except for these causal and tabular foundation models, we additionally compare against JIL, which is fitted separately for each test task. Section 4.1 evaluates zero- and few-shot performance across intervention-set sizes, different graph sizes, and partially observed query variables. Section 4.2 assesses Sim2Real transfer to real genetic and chemical perturbation datasets, and Section 4.3 examines the contribution of dual-regime pre-training. Implementation details are provided in Appendix E, while sensitivity to the edge probability and initialization scale of the test data is examined in Appendix F.2.

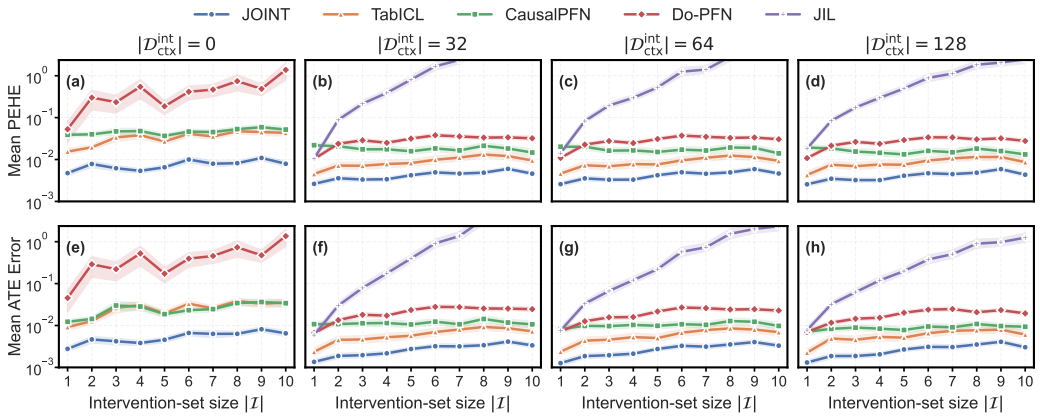

Figure 3: **Zero-shot (a, e) and few-shot (b–d, f–h) performance across intervention-set sizes.** Columns correspond to $|\mathcal{D}^{\text{int}}_{\text{ctx}}| \in \{0, 32, 64, 128\}$ interventional context samples, while keeping 1920 observational samples fixed. Rows report PEHE (top) and ATE error (bottom). Curves and shading show mean±SE over 100 independently sampled causal systems. JIL uses single-variable interventional samples with matched total sample counts and is evaluated only in the few-shot settings; vertical axes are logarithmic. JIL curves are clipped beyond the y-axis limits.

4.1 Performance

Zero-Shot and Few-Shot Performance across Intervention-Set Sizes. We evaluate intervention-set sizes $|\mathcal{I}| \in \{1, \ldots, 10\}$ under zero- and few-shot settings. Figure 3 shows that JOINT achieves the lowest mean PEHE and ATE error across all evaluated sizes and context settings. As $|\mathcal{I}|$ increases, JOINT maintains low errors with relatively gradual changes. TabICL and Do-PFN generally exhibit larger increases, particularly in the zero-shot setting, while CausalPFN remains comparatively stable but less accurate. Providing $|\mathcal{D}^{\text{int}}_{\text{ctx}}| \in \{32, 64, 128\}$ interventional samples improves all four models, yet JOINT maintains its lead, supporting prediction of joint intervention effects with or without interventional examples. JIL exhibits substantially increasing errors as the intervention set grows, although additional single-variable interventional samples reduce these errors. Notably, JOINT's zero-shot mean PEHE and mean ATE errors are already lower than all other methods (with the exception of TabICL) in their few-shot settings, despite JOINT using no interventional context samples. ATE-RE results in Appendix F.1 further support JOINT's advantage.

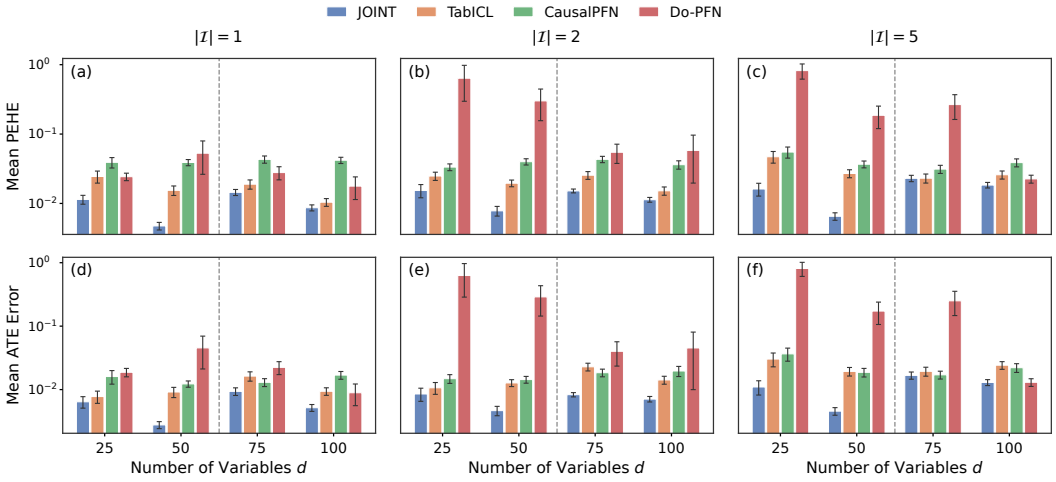

Figure 4: **Performance across graph sizes.** PEHE **(a–c)** and ATE error **(d–f)** under observational-only contexts, with columns corresponding to $|\mathcal{I}| \in \{1, 2, 5\}$. Expected average in-degree is kept approximately constant across graph sizes. Dashed lines separate the pre-training range ($d \leq 50$) from larger, unseen graphs. JOINT maintains strong performance beyond the pre-training range. Bars and error bars show mean±SE over 100 causal systems. vertical axes are logarithmic.

Performance on Graphs with Various Sizes. We evaluate zero-shot performance on graphs with $d \in \{25, 50, 75, 100\}$ variables and intervention-set sizes $|\mathcal{I}| \in \{1, 2, 5\}$. The corresponding edge probabilities are $p_{\text{edge}} \in \{0.10, 0.05, 0.033, 0.025\}$, keeping the expected average in-degree approximately constant at 1.2 (settings details are provided in Appendix E). Figure 4 shows that JOINT achieves the lowest mean PEHE and the lowest or comparable mean ATE error across all 12 evaluated settings. Although its errors increase moderately beyond the pre-training range ($d \leq 50$), JOINT maintains strong performance on unseen graphs with 75 and 100 variables, supporting generalization to larger causal systems.

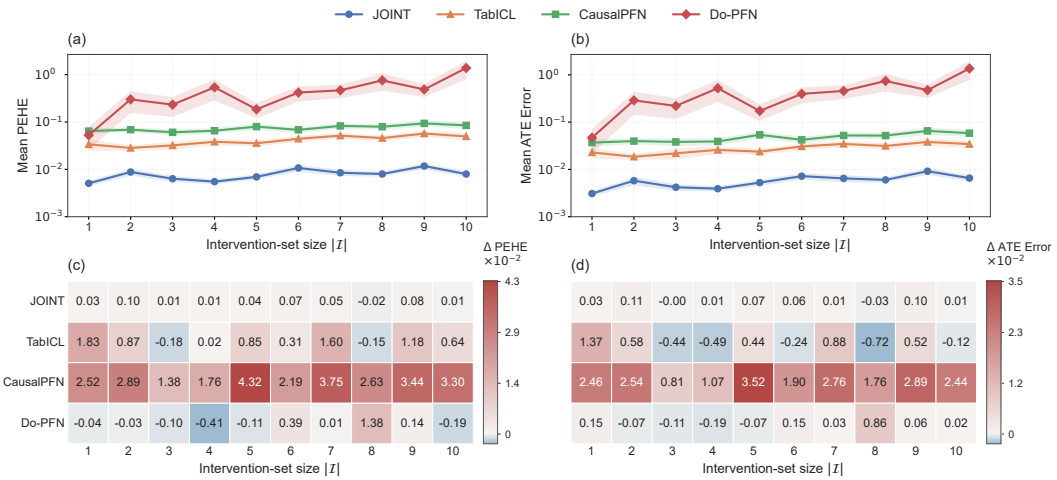

Figure 5: **Inference with partially observed variables.** We randomly mask 10%–90% of non-intervention, non-target query variables, using the same masks across all models. **(a–b)** PEHE and ATE error across intervention-set sizes using observational-only contexts. Curves and shading show mean±SE over 100 independently sampled causal systems; vertical axes are logarithmic. **(c–d)** Paired changes relative to fully observed inference, $\Delta\mathrm{e} = \mathrm{e}_{\text{partial}} - \mathrm{e}_{\text{full}}$, reported in units of 10^{-2}. Positive values indicate performance degradation.

Inference with Partially Observed Variables. In real-world systems, many variables other than the intervention and target variables may be unmeasured or inaccessible. We therefore randomly mask 10%–90% of non-intervention, non-target query variables to simulate these partially observed scenarios. All models use the same masks, with the context unchanged. Implementation details are provided in Appendix E. Figure 5 shows that JOINT achieves the lowest mean PEHE and ATE error across all evaluated intervention-set sizes. Compared with fully observed inputs, its largest mean error increases are only 1.0×10^{-3} and 1.1×10^{-3}, respectively, whereas CausalPFN exhibits increases of up to 4.32×10^{-2} and 3.52×10^{-2}. These results support JOINT's robustness to partially observed variables.

4.2 Validation on Real-World Scenarios

Table 1: ATE error on measured perturbation datasets (mean±SE over 100 realizations).

Method	Norman	ComboSciPlex
JOINT	$\mathbf{0.0161 \pm 0.0033}$	$\mathbf{0.0525 \pm 0.0083}$
TabICL	0.0188 ± 0.0034	0.0637 ± 0.0083
CausalPFN	0.0187 ± 0.0038	0.0603 ± 0.0095
Do-PFN	0.1931 ± 0.0456	0.2131 ± 0.0386

We assess Sim2Real transfer on Norman (Norman et al., 2019) and ComboSciPlex (Srivatsan et al., 2020), which measure responses to combinatorial CRISPRa and drug perturbations, respectively. For each scenario, we construct 100 evaluation datasets containing control, two single perturbations, and their combination. For each dataset, models predict a selected target gene's expression using the intervention indicators and 47 randomly selected gene-expression covariates, a small subset of the measured genes. Contexts contain control samples and samples with exactly one perturbation active, while the condition with both perturbations active is held out from the context. For evaluation, both perturbation indicators are jointly assigned to their control or active states in every query, yielding four configurations with a fixed intervention-set size of two. ATE error is averaged over the three non-control configurations relative to the all-control configuration. Implementation details are provided in Appendix E. Table 1 shows

that JOINT achieves the lowest mean ATE error on Norman (0.0161 ± 0.0033) and ComboSciPlex (0.0525 ± 0.0083), improving over the strongest baseline by approximately 13.9% and 12.9%, respectively. JOINT also achieves the lowest mean ATE-RE on both datasets (Appendix F.1). These results support Sim2Real transfer across genetic and chemical perturbations.

4.3 Ablation Study

To assess the effect of dual-regime pre-training, we compare JOINT with a variant pre-trained using observational-only contexts (Figure 6), replacing each interventional context sample with a new observational sample from the same causal system. Other settings remain unchanged. Both variants use identical test context–query splits under the settings in Appendix E, with 100 realizations, $|\mathcal{I}| \in \{1, 2, 5\}$, and $|\mathcal{D}_{\text{ctx}}^{\text{int}}| \in \{0, 32, 64, 128\}$. At inference time, the observational-only variant distinguishes observational and interventional samples using a binary regime-indicator column, while full JOINT uses its learned intervention embedding.

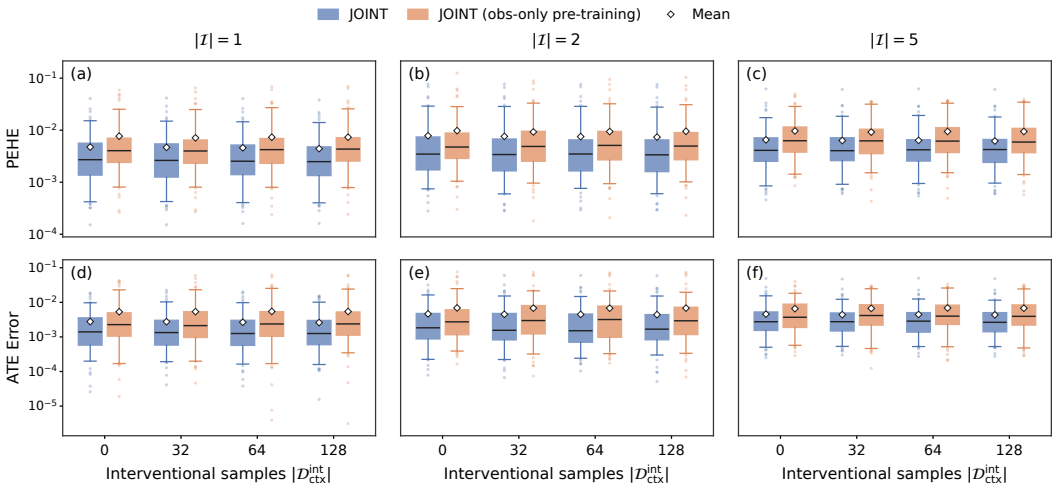

Figure 6: **Effect of dual-regime pre-training.** JOINT versus its observational-only pre-training variant across interventional context sizes. Columns correspond to $|\mathcal{I}| \in \{1, 2, 5\}$; rows show PEHE (top) and ATE error (bottom). Dual-regime pre-training consistently improves few-shot OJI prediction as interventional examples become available. Each box summarizes 100 realizations (25th–75th percentiles; whiskers: 5th–95th), with diamonds indicating means; vertical axes are logarithmic.

5 Conclusion

We introduced JOINT, a causal foundation model for estimating outcomes of joint interventions from observational data while leveraging interventional examples when available. By combining next-token-style causal data synthesis with dual-regime pre-training, JOINT supports both observational-only and intervention-assisted inference within a unified framework. Across synthetic benchmarks, JOINT consistently improves OJI estimation, remains robust under partial observability, and generalizes to causal systems larger than those seen during pre-training. Its performance on genetic and chemical perturbation datasets further demonstrates that causal capabilities learned entirely from simulation can transfer to real-world intervention scenarios.

More broadly, our results suggest a practical paradigm for causal foundation models: scalable simulation can provide diverse intervention scenarios for pre-training, while flexible context conditioning enables these learned capabilities to adapt to different levels of interventional evidence at inference time. Extending the diversity of simulated mechanisms and incorporating domain-specific structural knowledge may further improve Sim2Real transfer and support downstream tasks such as selecting intervention combinations to optimize target outcomes.

AI USE STATEMENT

We used generative AI tools to aid and polish writing, retrieve and discover relevant literature, support research ideation and execution, and draft sections of the paper. Specifically, these uses included discussing methodological and experimental designs, assisting with code revision, and refining the organization and presentation of the manuscript. We did not use generative AI tools to generate synthetic datasets or prove mathematical claims. The authors take responsibility for the final content of this work, including its text, claims, code, and results.

REFERENCES

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021. doi: https://doi.org/10.3982/ECTA15732. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15732.

Vahid Balazadeh, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C. Cresswell, and Rahul Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RblaNJGx8C.

Alexander Balke and Judea Pearl. *Probabilistic Evaluation of Counterfactual Queries*, pp. 237–254. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022. ISBN 9781450395861. URL https://doi.org/10.1145/3501714.3501733.

Luis H. Braga, Forough Farrokhyar, M. İrfan Dönmez, Caleb P. Nelson, Bernhard Haid, Kathy Herbst, Massimo Garriboli, Salvatore Cascio, Anka Nieuwhof-Leppink, Martin Kaefer, Darius J. Bägli, Nicholas Kalfa, Christina Ching, Magdalena Fossum, and Luke Harper. Randomized controlled trials – the what, when, how and why. *Journal of Pediatric Urology*, 21(2):397–404, 2025. ISSN 1477-5131. doi: https://doi.org/10.1016/j.jpurol.2024.11.021. URL https://www.sciencedirect.com/science/article/pii/S147751312400617X.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162. URL https://doi.org/10.1198/jcgs.2010.08162.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025.

Armin Kekić, Sergio Hernan Garrido Mejia, and Bernhard Schölkopf. Learning joint interventional effects from single-variable interventions in additive models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=i0m3hToUwf.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. *arXiv preprint arXiv:2112.10510*, 2021.

Thomas M. Norman, Max A. Horlbeck, Joseph M. Replogle, Alex Y. Ge, Albert Xu, Marco Jost, Luke A. Gilbert, and Jonathan S. Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019. doi: 10.1126/science.aax4438. URL https://www.science.org/doi/abs/10.1126/science.aax4438.

Clair Null, Christine P Stewart, Amy J Pickering, Holly N Dentz, Benjamin F Arnold, Charles D Arnold, Jade Benjamin-Chung, Thomas Clasen, Kathryn G Dewey, Lia CH Fernald, et al. Effects of water quality, sanitation, handwashing, and nutritional interventions on diarrhoea and child growth in rural kenya: a cluster-randomised controlled trial. *The Lancet Global Health*, 6(3): e316–e329, 2018.

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

Judea Pearl and Elias Bareinboim. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4):579 – 595, 2014. doi: 10.1214/14-STS486. URL https://doi.org/10.1214/14-STS486.

Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICL: A tabular foundation model for in-context learning on large data. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=0VvD1PmNzM.

Jingang QU, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=SxsyLjIfWB.

Jake Robertson, Arik Reuter, Siyuan Guo, Noah Hollmann, Frank Hutter, and Bernhard Schölkopf. Do-PFN: In-context learning for causal effect estimation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=OaNbl9b56B.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature biotechnology*, 42(6):927–935, 2024.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

Sanjay R. Srivatsan, José L. McFaline-Figueroa, Vijay Ramani, Lauren Saunders, Junyue Cao, Jonathan Packer, Hannah A. Pliner, Dana L. Jackson, Riza M. Daza, Lena Christiansen, Fan Zhang, Frank Steemers, Jay Shendure, and Cole Trapnell. Massively multiplex chemical transcriptomics at single-cell resolution. *Science*, 367(6473):45–51, 2020. doi: 10.1126/science.aax6234. URL https://www.science.org/doi/abs/10.1126/science.aax6234.

Annika Stechemesser, Nicolas Koch, Ebba Mark, Elina Dilger, Patrick Klösel, Laura Menicacci, Daniel Nachtigall, Felix Pretis, Nolan Ritter, Moritz Schwarz, Helena Vossen, and Anna Wenzel. Climate policies that achieved major emission reductions: Global evidence from two decades. *Science*, 385(6711):884–892, 2024. doi: 10.1126/science.adl6547. URL https://www.science.org/doi/abs/10.1126/science.adl6547.

Christopher Stith, Hossein Rahmani, and Jesse C Cresswell. Causal foundation models. *arXiv preprint arXiv:2609.03003*, 2026.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

A NOTATION

	Causal Systems and Interventions
$\mathcal{V}$	Set of d endogenous variables in topological order.
$\mathrm{V}_i,\ \mathrm{v}_i$; $\mathbf{V},\ \mathbf{v}$	Random variable and its realization; corresponding random and realized vectors.
$\mathcal{G},\ \mathrm{PA}_i$	Causal graph and parent indices of V_i.
$\mathrm{U}_i,\ \mathrm{u}_i$; $\mathbf{U},\ \mathbf{u}$	Exogenous noise variable, its realization, and corresponding random and realized vectors.
$\mathcal{I},\ \mathbf{v}_{\mathcal{I}}$	Intervention index set and assigned values.
$\mathrm{do}(\mathbf{V}_{\mathcal{I}} = \mathbf{v}_{\mathcal{I}})$	Intervention fixing the selected variables to $\mathbf{v}_{\mathcal{I}}$.
V_y	Target variable whose outcome is predicted.
	Causal Data Generation
$g_{\boldsymbol{\Phi}},\ \boldsymbol{\Phi}$	Transformer generator and its weights, fixed within each causal system.
p_{edge}	Probability of including a preceding variable as a parent.
$\mathbf{m}_i$	Attention mask vector; 0 allows attention and $-\infty$ blocks it.
$\mathbf{h}_i^K,\ \mathbf{h}_i^V$	Key and value representations of variable V_i.
$\mathcal{KV}_{<i},\ \mathcal{KV}_{\leq i}$	Cache before and after appending the representations of V_i.
τ	Softmax temperature.
$s_{\mathrm{init}},\ \lambda_{\mathrm{noise}}$	Initialization scale and coefficient adjusting attention to exogenous noise.
	Samples, Contexts, and Queries
$\mathcal{D}^{\mathrm{obs}},\ \mathcal{D}^{\mathrm{twin}}$	Observational sample set and set of paired interventional samples.
$N,\ N^{\mathrm{obs}},\ N^{\mathrm{twin}}$	Number of table samples, observational samples, and twin pairs, respectively.
$\mathbf{v}'^{(p)}$, $\mathbf{v}''^{(p)}$	Twin samples sharing noise under complementary assignments.
$\mathcal{C},\ \mathcal{Q}$	Context and held-out query sets.
$\mathcal{D}_{\mathrm{ctx}}^{\mathrm{obs}},\ \mathcal{D}_{\mathrm{ctx}}^{\mathrm{int}}$	Observational and interventional context subsets.
$\mathbf{v}^{(q)}$	Full variable vector for query q, realized under its intervention.
$\mathbf{v}_{\setminus y}^{(q)},\ \mathrm{v}_y^{(q)}$	Query input values and held-out target value.
$\mathcal{I}^{(p)},\ \mathcal{I}^{(q)}$	Intervention index sets for sample p and query q.
	Model Representations and Prediction
$\mathbf{h}_i^{(p)},\ \mathbf{e}_{\mathrm{int}}$	Cell representation and intervention embedding.
d_{emb}	Dimension of a cell representation.
$\mathbf{e}_{\mathrm{out}}^{\mathrm{known}},\ \mathbf{e}_{\mathrm{out}}^{\mathrm{unknown}}$	Target-role embeddings for a known downstream outcome or an unknown role.
$\mathbf{e}_{\mathrm{out}}^{(q)}$	Target-role embedding selected for query q.
$\mathrm{TF}_{\mathrm{col}}$	Column-wise encoder producing cell representations.
$\mathrm{TF}_{\mathrm{row}}$	Row-wise aggregation module producing sample representations.
$\mathrm{TF}_{\mathrm{icl}}$	In-context Transformer predicting query outcomes using the context.
$\theta,\ p_\theta$	JOINT parameters and conditional predictive density.
$\psi,\ \mathcal{L}(\theta)$	Sampled generator configuration and pre-training objective.
	Effect Estimation and Evaluation
$\mathbf{0}_{\mathcal{I}},\ N_q$	Control assignment and number of query instances per assignment.
$\mathrm{v}_y^{(i)},\ \hat{\mathrm{v}}_y^{(i)}$	Realized and predicted outcomes under the current assignment; subscript ctrl denotes the all-zero control.
$\tau^{(i)},\ \hat{\tau}^{(i)}$	True and predicted individual effects relative to control.
$\mathrm{ATE},\ \widehat{\mathrm{ATE}}$	True and predicted average effects.
PEHE	Root mean squared error of individual-effect estimates.
ATE Error	Absolute error in the estimated average effect.
ATE-RE	ATE error divided by $\lvert\mathrm{ATE}\rvert + 10^{-6}$.
$\Delta\mathrm{e}$	Paired error change, $\mathrm{e}_{\mathrm{partial}} - \mathrm{e}_{\mathrm{full}}$.

B Extended Related Work

Tabular Foundation Models. The PFN framework approximates Bayesian prediction by pre-training a model on datasets sampled from a prior over data-generating mechanisms (Müller et al., 2021). Let $\boldsymbol{\phi} \sim p(\boldsymbol{\phi})$ parameterize a mechanism, with parameter space Φ, and let $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}$ contain context and query samples. For query covariates $\mathbf{x}_{\text{test}}$, the posterior predictive distribution marginalizes over mechanisms compatible with the available observations:

$$p(\mathrm{y}_{\text{test}} \mid \mathbf{x}_{\text{test}}, \mathcal{D}_{\text{train}}) = \int_{\Phi} p(\mathrm{y}_{\text{test}} \mid \mathbf{x}_{\text{test}}, \boldsymbol{\phi}) p(\boldsymbol{\phi} \mid \mathbf{x}_{\text{test}}, \mathcal{D}_{\text{train}}) \, \mathrm{d}\boldsymbol{\phi}. \tag{10}$$

A Transformer $q_{\boldsymbol{\theta}}$ learns to approximate this distribution by minimizing the expected negative log-likelihood of held-out query targets:

$$\mathcal{L}_{\text{PFN}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D} \sim p(\mathcal{D})} \left[- \sum_{(\mathbf{x}_{\text{test}}, \mathrm{y}_{\text{test}}) \in \mathcal{D}_{\text{test}}} \log q_{\boldsymbol{\theta}}(\mathrm{y}_{\text{test}} \mid \mathbf{x}_{\text{test}}, \mathcal{D}_{\text{train}}) \right]. \tag{11}$$

At inference time, the pre-trained model conditions on a new dataset's labeled context to predict query targets without updating its parameters. Its predictive behavior therefore depends on both the observed context and the prior learned during pre-training.

Causal In-Context Learning. Do-PFN uses an observational context $\mathcal{D}^{\text{obs}} = \{(\mathrm{t}_j^{\text{obs}}, \mathbf{x}_j^{\text{obs}}, \mathrm{y}_j^{\text{obs}})\}_{j=1}^{M^{\text{obs}}}$ to predict interventional query outcomes given pre-intervention covariates. Here, T is a scalar treatment, $\mathbf{X}$ denotes covariates, and Y is the outcome. Data-generating mechanisms are sampled from a prior $\psi \sim p(\psi)$ and used to generate observational contexts and interventional query sets $\mathcal{D}^{\text{int}}$. For a predictive distribution $q_{\boldsymbol{\theta}}$, the objective can be written schematically as

$$\mathcal{L}_{\text{Causal-ICL}}(\boldsymbol{\theta}) = \mathbb{E}_{\psi, \mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{int}}} \left[- \sum_{\substack{(\mathrm{t}^{\text{int}}, \mathbf{x}^{\text{pt}}, \mathrm{y}^{\text{int}}) \\ \in \mathcal{D}^{\text{int}}}} \log q_{\boldsymbol{\theta}}\left(\mathrm{y}^{\text{int}} \mid \mathrm{do}(\mathrm{T} = \mathrm{t}^{\text{int}}), \mathbf{x}^{\text{pt}}, \mathcal{D}^{\text{obs}}\right) \right]. \tag{12}$$

The expectation includes both the sampled mechanisms and the context–query data generated from them, with $\mathbf{x}^{\text{pt}}$ denoting pre-intervention covariates. The intervention notation specifies the prediction task. CausalPFN instead supervises predictions of conditional expected potential outcomes, so its predictive distribution concerns these conditional means rather than individual outcome realizations.

Joint Intervention Learning. Kekić et al. (2025) study the estimation of OJIs from observational and single-variable interventional data. Their formulation assumes an additive outcome mechanism $\mathrm{Y} = \sum_{i=1}^{|\mathcal{I}|} f_i(\mathrm{T}_i, \mathrm{C}_i) + \mathrm{U}_y$, where each component f_i depends on a treatment variable T_i and an unobserved confounder C_i shared with the outcome, while U_y is an independent exogenous noise term. Although each component can be nonlinear, the outcome is additive across treatment-specific components. Under their structural and data-support assumptions, the mean OJIs $\mathbb{E}[\mathrm{Y} \mid \mathrm{do}(\mathrm{T}_1 = \mathrm{t}_1, \ldots, \mathrm{T}_{|\mathcal{I}|} = \mathrm{t}_{|\mathcal{I}|})]$ is identifiable from observational data $\mathcal{D}^{\text{obs}}$ and single-variable interventional datasets $\{\mathcal{D}^{\text{int},i}\}_{i=1}^{|\mathcal{I}|}$. Their estimator combines component functions using binary regime indicators:

$$\hat{f}(\mathrm{t}_1, \ldots, \mathrm{t}_{|\mathcal{I}|}, \mathrm{r}_1, \ldots, \mathrm{r}_{|\mathcal{I}|}) = \sum_{i=1}^{|\mathcal{I}|} \hat{f}_i(\mathrm{t}_1, \ldots, \mathrm{t}_{|\mathcal{I}|}, \mathrm{r}_i), \tag{13}$$

where r_i indicates whether treatment T_i is intervened upon. Setting all indicators to one yields an estimate of the mean of OJIs.

This approach provides an identification-based route to joint intervention estimation under an additive outcome structure and with access to single-variable interventional data. It fits an estimator to the available data for a given causal system, rather than pre-training a foundation model for in-context inference across systems. Its identification result therefore does not directly establish recovery from observational data alone. These differences distinguish its setting from our goal of learning OJIs across heterogeneous systems with either observational-only or mixed observational and interventional contexts.

C Endogenous Variable Generation

This section expands on Algorithm 1, detailing how variables are generated, interventions are applied, and key–value representations are cached and reused.

Let $\mathcal{V}$ denote a set of d endogenous variables arranged in a sampled topological order. For each variable V_i, we sample an attention mask $\mathbf{m}_i \in \{0, -\infty\}^{i-1}$, where $\mathrm{m}_{i,j} = 0$ includes V_j as a parent and $\mathrm{m}_{i,j} = -\infty$ excludes it. Each entry is 0 with probability p_{edge}. The Transformer weights $\boldsymbol{\Phi}$ and masks are shared across samples from the same causal system.

We initialize the cache as $\mathcal{KV}_{<1} = \emptyset$. The first variable has no parents and is generated from exogenous noise $\mathrm{u}_1 \sim p(\mathrm{U}_1)$:

$$\mathrm{v}_1 = g_{\boldsymbol{\Phi}}(\mathrm{u}_1). \tag{14}$$

After any intervention overwrite, the final value v_1 is passed separately through $g_{\boldsymbol{\Phi}}$ to extract $(\mathbf{h}_1^K, \mathbf{h}_1^V) = \text{Extract-KV}(g_{\boldsymbol{\Phi}}(\mathrm{v}_1))$, initializing $\mathcal{KV}_{\leq 1} = \{(\mathbf{h}_1^K, \mathbf{h}_1^V)\}$.

For each subsequent variable V_i, we sample $\mathrm{u}_i \sim p(\mathrm{U}_i)$ and generate

$$\mathrm{v}_i = g_{\boldsymbol{\Phi}}(\mathrm{u}_i \mid \mathcal{KV}_{<i}, \mathbf{m}_i), \qquad \mathcal{KV}_{<i} = \{(\mathbf{h}_j^K, \mathbf{h}_j^V)\}_{j=1}^{i-1}. \tag{15}$$

The mask is added to the attention logits before softmax, restricting attention over preceding variables to parent representations while retaining the exogenous-noise token. Variables with no parents are generated from noise alone.

Under $\mathrm{do}(\mathbf{V}_{\mathcal{I}} = \mathbf{v}_{\mathcal{I}})$, each selected value is overwritten with its prescribed assignment. The final value v_i is then fed back into $g_{\boldsymbol{\Phi}}$ in a separate pass to obtain $(\mathbf{h}_i^K, \mathbf{h}_i^V)$, rather than reusing representations computed while generating v_i. The cache is updated as

$$\mathcal{KV}_{\leq i} = \mathcal{KV}_{<i} \cup \{(\mathbf{h}_i^K, \mathbf{h}_i^V)\}. \tag{16}$$

Subsequent variables use this updated cache, propagating changes in intervened values while leaving all other generation mechanisms unchanged. Cached representations are reused throughout generation to avoid repeated computation.

D Detailed Evaluation Metrics

For the simulated causal systems, we evaluate the same N_q query instances under each joint intervention assignment $\mathbf{v}_{\mathcal{I}} \in \{0, 1\}^{|\mathcal{I}|}$, holding the exogenous noise realization fixed across assignments for each instance. Let $\mathrm{v}_y^{(i)}$ and $\hat{\mathrm{v}}_y^{(i)}$ denote the realized and predicted target outcomes for instance i under this assignment, respectively. The all-zero assignment $\mathbf{0}_{\mathcal{I}}$ serves as the control intervention, with corresponding outcomes $\mathrm{v}_{y,\mathrm{ctrl}}^{(i)}$ and $\hat{\mathrm{v}}_{y,\mathrm{ctrl}}^{(i)}$. The dependence on the current assignment is omitted from the metric notation below.

The true and predicted individual effects follow Equation 8. The corresponding true and predicted average treatment effects are evaluated as

$$\mathrm{ATE} = \frac{1}{N_q} \sum_{i=1}^{N_q} \tau^{(i)}, \qquad \widehat{\mathrm{ATE}} = \frac{1}{N_q} \sum_{i=1}^{N_q} \hat{\tau}^{(i)}. \tag{17}$$

PEHE and ATE error are defined in Equation 9, with $\mathrm{ATE\ Error} = |\mathrm{ATE} - \widehat{\mathrm{ATE}}|$. As a supplementary metric, ATE relative error is defined as

$$\text{ATE-RE} = \frac{\text{ATE Error}}{|\mathrm{ATE}| + \epsilon}. \tag{18}$$

Here, $\epsilon = 10^{-6}$ prevents division by zero, but ATE-RE can still become large when the true ATE is close to zero. We therefore report ATE-RE as a supplementary metric. Lower values indicate better performance for all three metrics.

For the simulated causal systems, each metric is first computed for each non-control assignment and then averaged over these assignments within a realization. We report the mean and standard

error of these realization-level averages. For the measured perturbation datasets, paired outcomes for the same cell under alternative interventions are unavailable. Reference ATEs are therefore computed as differences in measured mean target expression between each perturbation condition and control; predicted ATEs use the corresponding differences in mean predicted outcomes. We apply the same ATE error and ATE-RE definitions to these condition-level effects and do not report PEHE. The main real-world results average over the two single perturbations and their combination, while supplementary results also report the combination separately.

E Implementation Details

Data Synthesis Setup. We synthesize pre-training datasets using the proposed endogenous variable generation procedure (Algorithm 1). For each realization, the number of variables is sampled uniformly as $d \sim \mathcal{U}\{2, \ldots, 50\}$. Transformer configurations, including the number of layers, hidden dimensions, attention heads, and parameter initialization scales, are sampled to induce diverse marginal distributions and functional dependencies. The softmax temperature τ is sampled log-uniformly from $[0.1, 10.0]$, and the edge probability for each node is sampled from $\mathcal{U}(0, 1)$. Each synthesized dataset contains $N = 4096$ samples, with the context fraction sampled uniformly from $[0.2, 0.8]$. For each dataset, we select the mixed-context regime with probability 0.5 and sample the interventional proportion of the context uniformly from $[0, 0.2]$. Otherwise, the context contains only observational samples. The interventional context samples are drawn from the generated twin-network pairs. Exogenous noise is sampled using randomly selected Gaussian, uniform, or mixed schemes, with the mixed scheme allowing different variables to follow Gaussian, categorical, Zipf, or uniform distributions. We additionally sample an adjustment coefficient for attention to exogenous noise, $\lambda_{\text{noise}} \sim \mathcal{U}(-1, 5)$. When more than three keys are present, we subtract λ_{noise} times the standard deviation of the current query's temperature-scaled attention logits from the logit corresponding to its exogenous-noise token. This adjustment is applied before softmax: positive coefficients reduce attention to exogenous noise, while negative coefficients increase it.

JOINT Details. JOINT follows the architecture described in Section 3.2 and contains 28M learnable parameters. The column-wise and row-wise modules each use 3 blocks with representation dimension 128, while the in-context Transformer uses 12 blocks with dimension 512, obtained by concatenating four CLS representations. We train the model on the synthesized datasets for 5,000 optimization steps using AdamW with a batch size of 512 and a learning rate of 10^{-4}. We use a cosine learning-rate schedule with a warmup proportion of 2%.

Baseline Details. We compare JOINT with two causal foundation models, CausalPFN and Do-PFN, and the general-purpose tabular foundation model TabICL. We use the TabICLv2 regressor (QU et al., 2026), denoted as TabICL throughout the experiments. All baselines use pre-trained checkpoints without gradient-based test-time parameter updates. To accommodate multiple intervention variables through the scalar-treatment interfaces of CausalPFN and Do-PFN, we designate the first intervention variable as the treatment and include the remaining intervention variables among the covariates. TabICL treats all intervention variables as ordinary predictors. When observational and interventional samples need to be distinguished explicitly, we append a binary regime-indicator column to the baseline inputs, taking value one for observational context samples and zero for interventional context and query samples. JOINT receives the number of intervention variables and the number of interventional context samples through its native interface to construct the corresponding intervention annotations. Each model predicts all joint assignments using the same context within each realization. All models receive the same realizations, context–query splits, and preprocessing, while query outcomes are withheld from prediction and used only for evaluation. Do-PFN was originally pre-trained with pre-intervention query covariates; here, all models receive the available non-target measurements realized under the queried intervention. Its results therefore reflect transfer to both the joint-intervention and query-conditioning settings evaluated in this work.

Joint Intervention Learning Baseline. We additionally compare against an adaptation of the estimator proposed by Kekić et al. (2025), referred to here as the Joint Intervention Learning (JIL), using the authors' default regression settings. For each test realization, JIL uses the same observational context and query samples as the other methods, but replaces the joint-interventional context with supplementary single-variable interventional samples generated from the same causal system.

The total number of interventional samples is matched across methods at 32, 64, or 128, with JIL's samples distributed approximately evenly across the intervention variables. All available non-target variables are included as predictor inputs. This covariate extension is an empirical adaptation beyond the original identifiability result. JIL fits separate observational and single-variable interventional predictors for each realization and combines their predictions to estimate OJIs, without using query labels for fitting. Because it requires single-variable interventional data, JIL is excluded from the observational-only zero-shot setting. We report both PEHE and ATE error for this adaptation, with supplementary ATE-RE results provided in Appendix F.1.

Comparison with TabICLv2. Our TabICL baseline uses the TabICLv2 regressor (QU et al., 2026), while JOINT builds on the original TabICL backbone (QU et al., 2025). TabICLv2 incorporates architectural and pre-training improvements and is trained on approximately 35.2 million generated datasets, compared with 2.56 million for JOINT. Despite using fewer pre-training datasets, JOINT achieves lower errors on the evaluated joint-intervention tasks. Although this comparison does not isolate the effects of architecture, training scale, and data synthesis, it supports the value of pre-training specifically for OJI estimation. The dual-regime ablation further supports incorporating interventional examples into pre-training contexts.

Test Settings. We synthesize 100 independent test dataset realizations using the proposed endogenous variable generation procedure (Algorithm 1). Each realization contains $d = 50$ endogenous variables generated by a randomly initialized 2-layer Transformer with 4 attention heads and hidden dimension 20. The softmax temperature τ is sampled log-uniformly from $[0.1, 10.0]$, with edge probability $p_{\text{edge}} = 0.05$ and initialization scale $s_{\text{init}} = 1.0$. Additional graph-density settings, $p_{\text{edge}} \in \{0.025, 0.10\}$, are evaluated in Appendix F.2. Across methods and shot settings, each realization uses the same $|\mathcal{D}_{\text{ctx}}^{\text{obs}}| = 1920$ observational samples and $|\mathcal{D}_{\text{ctx}}^{\text{int}}| \in \{0, 32, 64, 128\}$ interventional samples, yielding total context sizes of $\{1920, 1952, 1984, 2048\}$, respectively. Zero and positive interventional counts define the zero-shot and few-shot settings. Interventional context samples are generated under randomly sampled intervention sets and assignments and are not paired with the observational context samples. For fair comparison, JOINT uses the unknown target-role embedding $\mathbf{e}_{\text{out}}^{\text{unknown}}$ during testing. We evaluate intervention-set sizes $|\mathcal{I}| \in \{1, \ldots, 10\}$. For each realization, the query set contains $N_q = 32$ samples for each of the $2^{|\mathcal{I}|}$ binary assignments, yielding $32 \times 2^{|\mathcal{I}|}$ query samples in total. The same 32 exogenous noise vectors are reused across assignments to obtain paired query instances. PEHE and ATE error are computed relative to the all-zero query assignment and averaged over the non-control assignments within each realization. We report their means and standard errors over the 100 realizations. Supplementary ATE-RE results are provided in Appendix F.1. For the partially observed setting, mask values are sampled from the empirical distributions of the corresponding context columns. For each query instance, the same masks are reused across intervention assignments and models.

Real-World Data Settings. Each realization uses two binary intervention indicators and 47 randomly sampled gene-expression covariates, with a randomly selected outcome gene. For Norman, we select perturbation pairs with measured control, both single perturbations, and their combination. Covariates are sampled without replacement from eligible highly variable genes, and the outcome is sampled from non-target genes. Each context contains 1920 control cells and 64 cells from each single perturbation. For ComboSciPlex, each context contains 896 control cells and 64 cells from each single-drug treatment. Models predict all four assignments using the same context, with the joint condition excluded from context and query outcomes withheld from prediction.

F Supplementary Experiments

F.1 Supplementary Effect Estimation Results

ATE Relative Error across Intervention-Set Sizes. We report supplementary ATE-RE results using the definition in Appendix D, with 10^{-6} added to the denominator. Figure 7 shows that JOINT achieves the lowest mean ATE-RE across all evaluated intervention-set sizes and context settings. Interventional context improves all four foundation models relative to inference with observational-only contexts, while JOINT maintains its advantage. JIL, evaluated only with interventional data,

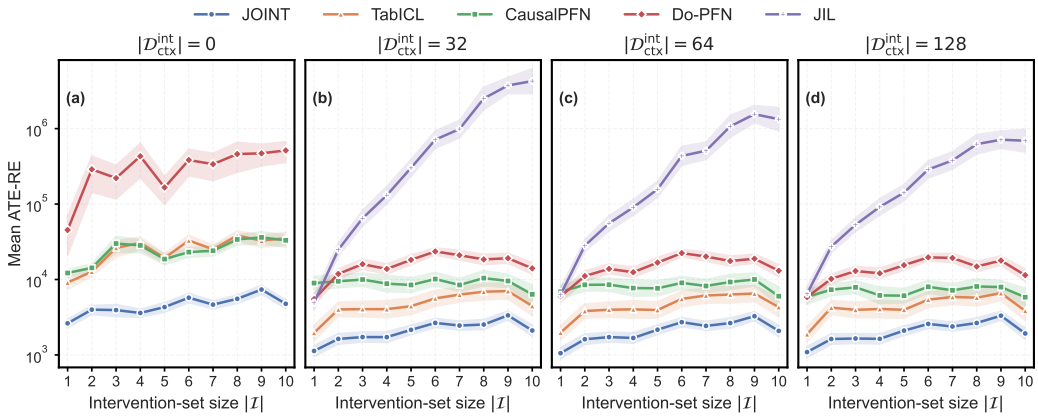

Figure 7: **ATE relative error across intervention-set sizes.** Panels correspond to $|\mathcal{D}^{\text{int}}_{\text{ctx}}| \in \{0, 32, 64, 128\}$ interventional samples with the same 1920 observational context samples. Curves and shading show means and $\pm$ one standard error over 100 independent causal systems. Vertical axes are logarithmic.

exhibits increasing ATE-RE as the intervention set grows, although more single-variable interventional samples generally reduce its errors. The large absolute values reflect the sensitivity of ATE-RE to near-zero reference effects.

Norman and ComboSciPlex. Table 2 reports supplementary ATE-RE averaged over all non-control assignments, together with ATE error and ATE-RE for simultaneous activation of both perturbations ($\mathrm{v}_j = 1,\ for\ all\ j \in \mathcal{I}$) relative to the all-zero control. JOINT achieves the lowest mean errors across all reported metrics on both datasets, including the simultaneous activation of both perturbations, which is withheld from the context. The large ATE-RE magnitudes reflect its sensitivity to zero or near-zero reference effects, even with the denominator stabilized by 10^{-6}.

Table 2: **Supplementary effect estimation results on measured perturbation datasets.** Results are mean $\pm$ standard error over 100 realizations. "All" averages over all non-control assignments; $\mathrm{v}_j = 1\ (for\ all\ j \in \mathcal{I})$ denotes simultaneous activation of both perturbations. All effects are relative to the all-zero control $\mathbf{0}_{\mathcal{I}}$. ATE-RE uses $\epsilon = 10^{-6}$ and is reported in units of 10^3.

	Norman			ComboSciPlex		
Method	ATE-RE ($\times 10^3$) All	ATE error $\mathrm{v}_j = 1$	ATE-RE ($\times 10^3$) $\mathrm{v}_j = 1$	ATE-RE ($\times 10^3$) All	ATE error $\mathrm{v}_j = 1$	ATE-RE ($\times 10^3$) $\mathrm{v}_j = 1$
JOINT	$\mathbf{0.343 \pm 0.060}$	$\mathbf{0.0185 \pm 0.0043}$	$\mathbf{0.469 \pm 0.075}$	$\mathbf{0.642 \pm 0.129}$	$\mathbf{0.0591 \pm 0.0100}$	$\mathbf{0.996 \pm 0.220}$
TabICL	1.293 ± 0.416	0.0199 ± 0.0042	1.368 ± 0.533	7.702 ± 2.401	0.0668 ± 0.0094	5.905 ± 1.844
CausalPFN	0.824 ± 0.233	0.0228 ± 0.0064	0.934 ± 0.316	2.032 ± 0.538	0.0640 ± 0.0119	2.793 ± 0.979
Do-PFN	104.801 ± 43.142	0.2591 ± 0.0886	144.509 ± 86.729	13.429 ± 5.592	0.2122 ± 0.0346	8.139 ± 5.398

F.2 Sensitivity to Edge Probability and Initialization Scale

We assess sensitivity to the edge probability p_{edge} and initialization scale s_{init} used to generate the test causal systems. We vary one parameter at a time while keeping the pre-trained models and all other test settings fixed as described in Appendix E. Both experiments consider intervention-set sizes $|\mathcal{I}| \in \{1, 2, 5\}$, with 100 realizations per setting. The default configuration, $p_{\text{edge}} = 0.05$ and $s_{\text{init}} = 1.0$, serves as a shared reference, and its results are repeated in both figures.

Edge Probability. We evaluate $p_{\text{edge}} \in \{0.025, 0.05, 0.10\}$ with $d = 50$ variables, covering sparser and denser causal graphs relative to the default setting. Figure 8 shows that JOINT achieves the lowest mean PEHE and mean absolute ATE error across all evaluated edge probabilities and intervention-set sizes. For $|\mathcal{I}| = 5$, increasing p_{edge} to 0.10 raises the errors of all evaluated methods, but JOINT retains its advantage. These results demonstrate that its comparative performance is not restricted to the default graph density.

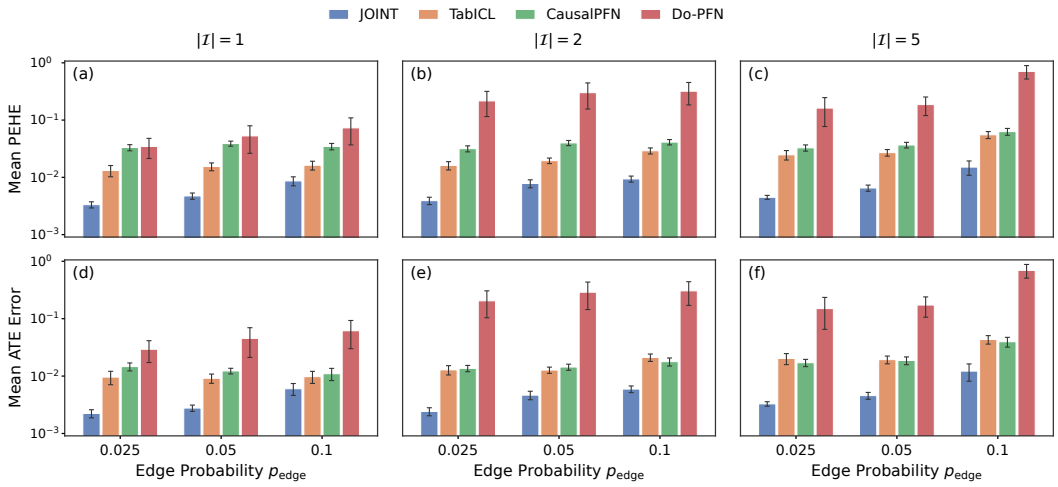

Figure 8: **Sensitivity to the edge probability of test causal graphs.** PEHE (top row) and absolute ATE error (bottom row) across $p_{\text{edge}} \in \{0.025, 0.05, 0.10\}$, with columns corresponding to $|\mathcal{I}| = 1$, 2, and 5. Bars and error bars denote the mean and standard error, respectively, over 100 realizations. The default edge probability is 0.05; all other test settings are held fixed. The vertical axes use logarithmic scales, and lower values indicate better performance.

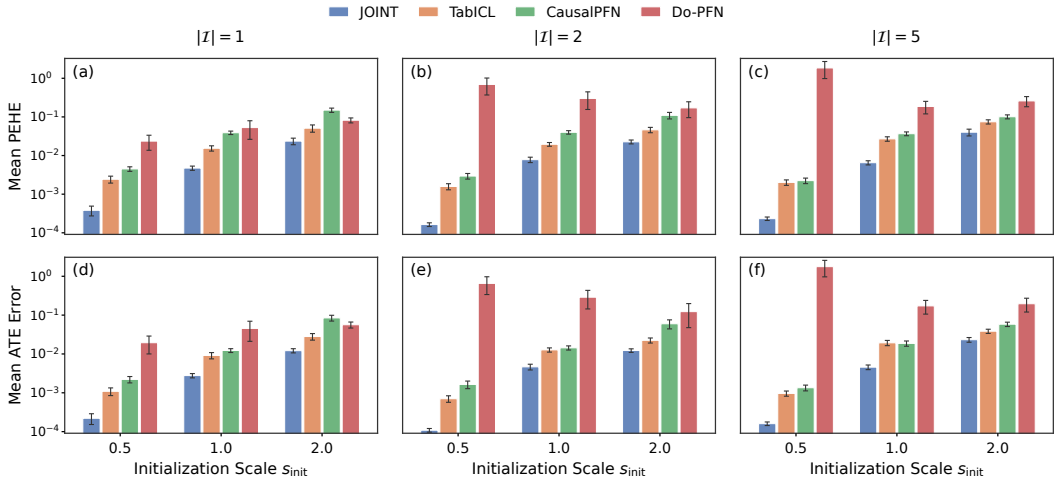

Figure 9: **Sensitivity to the initialization scale used for test-data generation.** PEHE (top row) and absolute ATE error (bottom row) across $s_{\text{init}} \in \{0.5, 1.0, 2.0\}$, with columns corresponding to $|\mathcal{I}| = 1$, 2, and 5. Bars and error bars denote the mean and standard error, respectively, over 100 realizations. The default initialization scale is 1.0; all other test settings are held fixed. The vertical axes use logarithmic scales, and lower values indicate better performance.

Initialization Scale. We evaluate $s_{\text{init}} \in \{0.5, 1.0, 2.0\}$ for the Transformer used to generate the test causal systems. This parameter scales the initialization standard deviations of the input-mapping and feed-forward weights, while leaving the QKV projection initialization unchanged. Figure 9 shows that JOINT attains the lowest mean PEHE and mean absolute ATE error across all evaluated scales and intervention-set sizes. Increasing the initialization scale generally raises the errors of JOINT, TabICL, and CausalPFN, whereas Do-PFN exhibits a different, non-monotonic response for joint interventions. Despite these differences, JOINT consistently maintains its advantage, indicating that its comparative performance extends beyond the default initialization scale.

F.3 Case Study on Semi-Synthetic Benchmarks

To assess transfer from joint-intervention learning to conventional single-treatment tasks, we evaluate JOINT on established semi-synthetic benchmarks. JOINT is pre-trained exclusively on our simulated causal systems; benchmark-specific observations are provided only as inference-time context, without benchmark-specific pre-training or fine-tuning. Table 3 reports PEHE and ATE relative error, with baseline results taken from Balazadeh et al. (2025). JOINT follows the same evaluation protocol, including dataset versions, data splits, preprocessing, number of repetitions, and metric aggregation. JOINT achieves competitive results on Lalonde CPS and PSID, with PEHE close to that of CausalPFN, the lowest mean ATE relative error on CPS, and an ATE relative error matching CausalPFN on PSID at the reported precision. On IHDP, JOINT also achieves competitive ATE relative error (0.19 ± 0.03). Performance on ACIC 2016 is weaker across both metrics. These results illustrate transfer to several conventional single-treatment settings without benchmark-specific parameter updates, while revealing variation across datasets and evaluation metrics.

Table 3: **Performance on semi-synthetic benchmarks.** Results are reported as mean $\pm$ standard error. Baseline results are taken from Balazadeh et al. (2025). PEHE values for Lalonde CPS and PSID are reported in thousands. Dashes indicate metrics not reported.

	PEHE				ATE Relative Error			
Method	IHDP	ACIC 2016	Lalonde CPS ($\times 10^3$)	Lalonde PSID ($\times 10^3$)	IHDP	ACIC 2016	Lalonde CPS	Lalonde PSID
Do-PFN	6.07±0.89	4.11±0.52	12.01±0.03	20.91±0.14	0.57±0.10	0.67±0.04	0.87±0.01	0.92±0.01
CausalPFN	**0.58±0.07**	0.92±0.11	**8.96±0.02**	**14.40±0.20**	0.20±0.04	0.05±0.01	0.13±0.01	0.22±0.02
T-Learner	1.73±0.30	0.76±0.07	9.22±0.04	15.16±0.46	0.21±0.04	**0.03±0.01**	0.24±0.02	0.16±0.03
DA-Learner	2.07±0.36	0.72±0.08	9.39±0.06	14.55±0.24	0.23±0.04	**0.03±0.01**	0.27±0.02	0.20±0.03
DragonNet	2.16±0.25	2.11±0.19	10.93±0.15	16.45±0.29	0.20±0.04	0.06±0.02	0.55±0.03	0.47±0.03
IPW	–	–	–	–	0.24±0.04	0.21±0.05	0.17±0.01	**0.10±0.01**
RA-Net	2.35±0.19	2.35±0.25	11.74±0.09	18.33±0.43	0.20±0.04	0.07±0.03	0.74±0.02	0.50±0.04
X-Learner	3.31±0.51	**0.60±0.08**	12.15±0.15	20.28±0.49	0.16±0.04	**0.03±0.01**	0.84±0.03	0.72±0.03
TarNet	1.82±0.14	2.20±0.21	12.88±0.02	19.19±0.18	0.20±0.04	0.05±0.02	1.00±0.00	0.78±0.01
S-Learner	2.57±0.41	0.85±0.13	12.66±0.05	21.80±0.18	0.20±0.04	**0.03±0.01**	0.97±0.01	0.90±0.02
BART	2.50±0.39	0.68±0.11	12.81±0.05	21.36±0.16	0.44±0.09	0.04±0.01	0.99±0.01	0.86±0.01
GRF	3.67±0.61	1.32±0.30	12.33±0.06	22.91±0.17	0.18±0.03	0.07±0.02	0.82±0.02	0.85±0.02
Forest DML	4.53±0.73	1.48±0.31	12.95±0.04	22.99±0.15	**0.08±0.01**	0.05±0.01	1.03±0.01	1.05±0.01
Forest DR Learner	4.02±0.67	1.34±0.29	15.98±0.68	22.78±0.54	0.17±0.03	0.04±0.02	1.20±0.23	3.64±2.78
JOINT	3.73±0.57	3.26±0.55	9.03±0.02	14.90±0.17	0.19±0.03	0.26±0.06	**0.12±0.01**	0.22±0.01