Joint Treatment Effect Estimation from Incomplete Healthcare Data: Temporal Causal Normalizing Flows with LLM-driven Evolutionary MNAR Imputation

Olivia Jullian Parra*
Department of Mathematical Modeling and Machine Learning
University of Zürich
Winterthurerstrasse 190, Zürich 8057
olivia.jullianparra@uzh.ch

Sara Zoccheddu*
Department of Mathematical Modeling and Machine Learning
University of Zürich
Winterthurerstrasse 190, Zürich 8057
sara.zoccheddu@uzh.ch

David Catalan Cerezo
Department of Mathematics
ETH Zürich
Rämistrasse 101, Zürich 8092
dcatalan@ethz.ch

Tom Forzy
Epidemiology, Biostatistics & Prevention Institute (EBPI), University of Zürich
Hirschengraben 84, Zürich 8001
tom.forzy@uzh.ch

Franziska Ulrich
Department of Mathematics
ETH Zürich
Rämistrasse 101, Zürich 8092
ulricfr@ethz.ch

William Sutcliffe
Department of Physics
University of Zürich
Winterthurerstrasse 190, Zürich 8057
william.sutcliffe@physik.uzh.ch

Jakob Martin Burgstaller
Institute of Primary Care
University of Zürich
Sonneggstrasse 6, Zürich 8091
jakobmartin.burgstaller@usz.ch

Oliver Senn
Institute of Primary Care
University of Zürich
Sonneggstrasse 6, Zürich 8091
oliver.senn@usz.ch

Patrick Owen
Department of Physics
University of Zürich
Winterthurerstrasse 190, Zürich 8057
powen@physik.uzh.ch

Nicola Serra
Department of Physics &
Department of Mathematical Modeling and Machine Learning
University of Zürich
Winterthurerstrasse 190, Zürich 8057
nicola.serra@physik.uzh.ch

Abstract

Target trial emulation (TTE) provides a framework for answering causal questions using observational data when randomized controlled trials (RCTs) are infeasible. However, standard methods for treatment effect estimation have been developed in isolation, failing to jointly address the compounding challenges inherent to analyses of observational data such as electronic health records (EHRs). In particular, these challenges include time-varying confounding and missing-not-at-random (MNAR) missingness reaching 50%–80% for critical biomarkers. To address this gap, we propose a two-stage pipeline that jointly handles MNAR missingness

*Equal contribution.

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

and causal structure across time. CausalFlow-T adapts a Directed Acyclic Graph (DAG)-constrained normalizing flow conditioned on a Long Short-Term Memory (LSTM)-encoding of patient history and performs exact invertible counterfactual inference, eliminating the approximation errors and confounding biases where existing variational and adversarial methods can fail silently. Ablations on four synthetic and one semi-synthetic datasets with known counterfactuals confirm that its two core design choices address strictly non-overlapping failure modes (DAG constraints for confounding separation, exact inference for structural propagation) with neither compensating for the absence of the other. To handle the incomplete data CausalFlow-T receives as input, we propose an LLM-driven evolutionary imputer and evaluate it with three LLM backends, including an open-weight model. Across 30%–80% MNAR missingness, the imputer achieves the best pooled rank across biomarker and causal metrics, leading on pointwise accuracy and temporal extrapolation while maintaining a competitive average treatment effect (ATE) recovery where statistical methods progressively degrade. Applied to a cohort of adults with type 2 diabetes in Swiss primary care initiating a GLP-1 receptor agonist or SGLT-2 inhibitor, the pipeline recovers a per-protocol weight-loss difference of -0.93 kg [95% CI -1.26, -0.70] favoring GLP-1 receptor agonists, consistent with RCT evidence and estimated directly from realistically incomplete real-world EHR data.[1]

1 Introduction

Type 2 diabetes (T2D) is one of the most common chronic conditions affecting over half a billion adults worldwide [International Diabetes Federation, 2025]. To optimize glycemic control and reduce the risk of complications (e.g. cardiovascular disease), effective management requires a multifactorial approach [American Diabetes Association, 2025]. Recent therapeutic advances, particularly two drug classes, GLP-1 receptor agonists (GLP-1RAs; including semaglutide, marketed as Ozempic) and sodium-glucose co-transporter-2 inhibitors (SGLT-2is), have transformed T2D management by demonstrating cardiovascular, renal, and weight loss benefits beyond glycemic control in randomized controlled trials (RCTs) [American Diabetes Association, 2025]. However, RCTs are limited by high costs, restricted sample sizes, and strict eligibility criteria, making them insufficient to address the complexity of T2D treatment, where therapies are numerous, combined, and evolve over time [Schneeweiss and Patorno, 2021]. Observational data (e.g. electronic health records [EHRs]) can complement RCTs by providing real-world evidence across diverse patient populations and care settings [Schneeweiss and Patorno, 2021]. Yet, causal inference from EHRs remains challenging due to confounding, selection bias, and other structural biases arising from non-randomized treatment assignment. Target trial emulation (TTE) addresses this by explicitly framing observational analyses to mimic a hypothetical RCT, thereby improving causal validity [Hernán and Robins, 2016]. Despite this framework, existing methods for treatment effect estimation have been developed for individual challenges in isolation, and none jointly addresses the structural challenges of real EHR data [Shalit et al., 2017, Louizos et al., 2017, Bica et al., 2020, D'Amour et al., 2021]. First, treatment decisions are driven by the patient's evolving clinical profile, creating time-varying confounding that existing methods handle only under parametric assumptions or without explicit causal structure [Shalit et al., 2017, Louizos et al., 2017, Bica et al., 2020, D'Amour et al., 2021]. Second, EHR data exhibit substantial missingness (often exceeding 50%) that arises under missing-not-at-random (MNAR) mechanisms (clinical decisions drive data collection) while imputation strategies have typically been validated at far lower missingness rates [He et al., 2025, Mangussi et al., 2026]. These challenges interact: MNAR missingness distorts biomarker (e.g. blood pressure)–outcome relationships, compounding bias in downstream causal estimation.

We address this gap with a two-stage pipeline that jointly handles time-varying confounding and MNAR missingness within the TTE framework. We apply this pipeline to estimate the per-protocol effect of GLP-1RAs versus SGLT-2is on 1-year body weight in a Swiss primary care cohort of adults with T2D, extending existing RCT evidence from specific agents in selected populations to the full drug classes used in real-world clinical practice. Our contributions are as follows:

[1] Code for CausalFlow-T is available at `https://github.com/oliviajullian/CausalFlowT_final.git`; code for LLM-driven imputation at `https://github.com/sarazcc/LLMEvolutionaryImputation.git`.

CausalFlow-T: We adapt the causal masked autoregressive flow (CausalMAF) of Javaloy et al. [2023] to longitudinal data by conditioning it on a patient-history summary produced by an LSTM encoder [Hochreiter and Schmidhuber, 1997]. As a result, we model the joint distribution of covariates, treatment, and outcomes over time. Experiments on four synthetic datasets and one semi-synthetic evaluation dataset with known counterfactuals reveal two distinct failure modes: (i) DAG constraints are necessary for confounding separation (unconstrained models showed systematic low-effect bias on the semi-synthetic benchmark and unstable estimates under full mediation); (ii) exact inference is required for structural propagation (in our benchmarks, variational methods fail to recover the true treatment effect under complex mediation paths even with correctly specified causal graphs).

LLM-driven evolutionary imputation: A large language model (LLM)-driven evolutionary pipeline benchmarked against LOCF [Fitzmaurice et al., 2012], MissForest [Stekhoven and Bühlmann, 2012], and a DAG-aware flow matching (CFM) baseline (inspired by the m-graph framework [Mohan and Pearl, 2021]) across 30–80% MNAR missingness using three LLM backends (GPT-5.4 and GPT-OSS-120b from OpenAI, Qwen3.5-Plus from Alibaba Cloud). Evaluated via reconstruction, causal correlation structure preservation, and downstream ATE recovery, LLM-based methods achieve the strongest pooled performance, with GPT-5.4 ranking best overall and the other two LLM backends confirming robustness and reconstruction–causal tradeoffs.

Real-world application: When applied to a Swiss primary care T2D cohort initiating GLP-1RAs or SGLT-2is, our two-stage pipeline produces per-protocol estimates of 1-year body weight change (kg) from incomplete EHR data that are consistent with RCT evidence.

2 Related Work

Treatment effect estimation and time-varying confounding: Marginal structural models (MSMs) with inverse probability weighting (IPW) [Hernán and Robins, 2020] address time-varying confounding but rely on parametric assumptions that can break down in high-dimensional, nonlinear settings [D'Amour et al., 2021]. Deep learning (DL) approaches such as R-MSN [Lim et al., 2018], CRN [Bica et al., 2020], and the Causal Transformer [Melnychuk et al., 2022] improve flexibility but rely on approximate inference or lack explicit causal structure,and assume that the learned history representation preserves the information required for adjustment [Melnychuk et al., 2024]. Generative models such as CEVAE [Louizos et al., 2017] and GANITE [Yoon et al., 2018] introduce additional approximation gaps via variational or adversarial objectives. Many of these methods have been primarily evaluated on synthetic datasets with limited missingness and large sample sizes [Shalit et al., 2017], conditions rarely met in real-world clinical data [Sun et al., 2024, Barrett et al., 2020].

Normalizing flows for causal inference: Khemakhem et al. [2021] establish identifiability via nonlinear independent component analysis for autoregressive flows; Javaloy et al. [2023] extend this to general triangular mappings with a full do-operator formulation and introduce the CausalMAF normalizing flow that we use as our generative core. Chao et al. [2023] use diffusion models at the node level, but requiring a separate network per variable limits scalability for high-dimensional EHR data. To our knowledge, these methods assume static data. We extend CausalMAF to longitudinal settings using LSTM-encoded patient histories. Our contribution is their integration with MNAR imputation for longitudinal target trial emulation and the evaluation of failure modes invisible to factual prediction error.

Missing data handling in causal inference: Missing data in EHRs can bias treatment effect estimation even after imputation [Zhou et al., 2023]. Effective imputation must leverage temporal disease trajectories and account for the underlying missingness mechanism [Zhou et al., 2023]. Standard methods such as Multiple Imputation by Chained Equations (MICE) [Sterne et al., 2009] and missing forest (MissForest) [Stekhoven and Bühlmann, 2012] assume missing-at-random (MAR) and can amplify bias under MNAR, while LOCF distorts temporal correlation structure by construction. Selection models jointly model data and missingness [Diggle and Kenward, 1994], while pattern-mixture models impose identifying restrictions [Little, 1993]; both support MNAR-aware multiple imputation and sensitivity analysis [Carpenter et al., 2013]. M-graphs assess recoverability and inform our CausalCFM baseline [Mohan and Pearl, 2021]. Genetic algorithms have also been combined with regression-based imputation [Mohammed et al., 2022].

LLM-based imputation and program search: Recent LLM-based approaches show improved reconstruction, but have been validated mainly at regimes of 5–40% missingness [He et al., 2025,

Mangussi et al., 2026], below the high-rate MNAR missingness often encountered in real-world EHR data; in parallel, LLMs have been used as mutation operators in evolutionary program search for mathematical and algorithmic discovery [Romera-Paredes et al., 2024, Liu et al., 2024, Novikov et al., 2025, Lange et al., 2025]. We select executable longitudinal imputers, penalizing reconstruction error and biomarker–treatment and biomarker–outcome association distortion. This objective within causal estimation is our contribution.

3 Methodology

We propose a two-stage pipeline for treatment effect estimation from incomplete longitudinal observational data. CausalFlow-T (Section 3.2) addresses time-varying confounding and exact counterfactual inference, while the LLM-driven evolutionary imputer (Section 3.3) handles MNAR missingness and feeds completed data into CausalFlow-T. Sequential validation isolates the contribution of each stage, ensuring downstream performance differences reflect imputation quality rather than estimator failure.

3.1 Problem formulation and assumptions

Consider N patients over T discrete time steps. At each time t, patient i has covariates $x_{i,t} \in \mathbb{R}^d$ with components $x_{i,t,j}$, $j \in \{1, \dots, d\}$ (with $x_{i,t,-j}$ denoting all components except j), treatment $a_{i,t} \in \{0, 1\}$, outcome $y_{i,t} \in \mathbb{R}$, and missingness mask $m_{i,t} \in \{0, 1\}^d$, where $m_{i,t,j} = 1$ if $x_{i,t,j}$ is missing and 0 if observed; this convention is used throughout. Causal relationships between (x, a, y) are encoded in an expert-specified DAG $\mathcal{G} = (V, E)$. The complete dataset is $\mathcal{D} = \{(x_{i,t}, a_{i,t}, y_{i,t})\}_{i=1,t=1}^{N,T}$, and the incomplete longitudinal dataset is $\widetilde{\mathcal{D}} = \{(\tilde{x}_{i,t}, a_{i,t}, y_{i,t}, m_{i,t})\}_{i=1,t=1}^{N,T}$, where $\tilde{x}_{i,t,j} = x_{i,t,j}$ if $m_{i,t,j} = 0$ and $\tilde{x}_{i,t,j} = \texttt{NA}$ otherwise. An imputer g returns a completed dataset $\hat{\mathcal{D}} = g(\widetilde{\mathcal{D}})$ with entries $\hat{x}_{i,t,j}$, which equal $\tilde{x}_{i,t,j}$ wherever $m_{i,t,j} = 0$. Missingness is MNAR: the probability of a value being unobserved depends on the unobserved value itself,

$$P(m_{i,t,j} = 1 \mid x_{i,t,j}, x_{i,t,-j}, a_{i,t}, y_{i,t}) = f_j(x_{i,t,j}, x_{i,t,-j}, a_{i,t}, y_{i,t}), \tag{1}$$

where f_j depends on $x_{i,t,j}$ itself (the MNAR component), as well as on the remaining covariates $x_{i,t,-j}$ and on the outcome state $y_{i,t}$, mirroring the clinical pattern where tests ordered on suspicion produce structured, not random, absence. Let $v_{i,t} = (x_{i,t}, a_{i,t}, y_{i,t}) \in \mathbb{R}^p$, $p = d + 2$, denote the variables at time t, with nodes indexed by $k \in \{1, \dots, p\}$ (we reserve j for covariates). $\mathcal{G}$ encodes their contemporaneous causal relations; $\mathrm{pa}(k)$ denotes the parents of node k, and nodes are ordered by a topological sort of $\mathcal{G}$. The covariate history before time t, $H_{i,t} = (x_{i,1}, \dots, x_{i,t-1})$, is summarized by an LSTM encoding $h_{i,t} \in \mathbb{R}^H$ (Eq. (3)).

Finally, let $y_{i,t}(a')$ be the potential outcome at time t under $\mathrm{do}(a_{i,1} = \cdots = a_{i,t} = a')$ [Pearl, 2009], i.e., had the patient received treatment $a' \in \{0, 1\}$ from baseline through time t. With individual treatment effects $\tau_{i,t} = y_{i,t}(1) - y_{i,t}(0)$, the goal is to estimate their per-time average $\mathrm{ATE}_t = \mathbb{E}[\tau_{i,t}]$ and the overall $\mathrm{ATE} = T^{-1} \sum_{t=1}^{T} \mathrm{ATE}_t$ (Appendix D).

Assumptions. We require (i) *sequential ignorability* $y_{i,s}(a') \perp\!\!\!\perp a_{i,t} \mid h_{i,t}, x_{i,t}$ for $s \geq t$; (ii) *history sufficiency*, $p(v_{i,t} \mid H_{i,t}) = p(v_{i,t} \mid h_{i,t})$; (iii) *correct DAG specification*; and (iv) *positivity and consistency*. Then $\mathbb{E}[y_{i,t}(a')]$ is identified via the sequential g-formula [Robins, 1997, Gill and Robins, 2001], and abduction-action-prediction (AAP) recovers it given consistently estimated flow conditionals. These assumptions do not identify individual counterfactuals; exact inversion makes abduction exact only for the fitted model (Appendix M). Appendix I analyzes DAG misspecification.

Two-stage pipeline. We first validate CausalFlow-T on complete-data settings $\mathcal{D}$ with known counterfactuals, establishing its causal estimation properties independently of any imputation choices. Once validated, the LLM-driven evolutionary imputer produces $\hat{\mathcal{D}}$, enabling CausalFlow-T to operate on realistic MNAR data with the following two-stage pipeline (Figure 1):

$$\widetilde{\mathcal{D}} \xrightarrow{g^\star} \hat{\mathcal{D}} \xrightarrow{\text{CausalFlow-T}} \{\hat{y}_{i,t}(a'), \hat{\tau}_{i,t}\}$$

3.2 CausalFlow-T: Counterfactual Inference with Exact Inversion

CausalFlow-T estimates $\hat{y}_{i,t}(a')$ and $\hat{\tau}_{i,t}$ via a DAG-constrained normalizing flow conditioned on longitudinal patient history, validated first on complete data $\mathcal{D}$ where ground-truth counterfactuals are known.

DAG-Constrained CausalMAF. The core generative model is a Causal Masked Autoregressive Flow (CausalMAF) [Javaloy et al., 2023].Given the history encoding $h_{i,t}$, the flow learns a conditional bijection $f_\theta(\cdot\,; h_{i,t}) : \mathbb{R}^p \to \mathbb{R}^p$, $z_{i,t} = f_\theta(v_{i,t}; h_{i,t})$ with $z_{i,t} \sim \mathcal{N}(0, I_p)$, with autoregressive structure constrained to a topological sort of $\mathcal{G}$:

$$v_{i,t,k} = f_{\theta,k}^{-1}\big(z_{i,t,k};\, v_{i,t,\mathrm{pa}(k)}, h_{i,t}\big), \qquad k = 1, \ldots, p, \tag{2}$$

where $\mathrm{pa}(k)$ denotes the causal parents of node k in $\mathcal{G}$ and each $f_{\theta,k}$ is strictly increasing in $v_{i,t,k}$. Intervening on $a_{i,t}$ therefore propagates exclusively through causal descendants while non-descendants are held fixed, implementing $P(Y \mid \mathrm{do}(a'), X)$ directly. Without this constraint, the flow may propagate interventions through arbitrary learned dependencies, including anti-causal ones, producing systematically biased counterfactuals despite accurately fitting the factual distribution (Section 4.2).

The flow maximizes the exact log-likelihood $\log p_\theta(v_{i,t} \mid h_{i,t}) = \log p_z\big(f_\theta(v_{i,t}; h_{i,t})\big) + \log\big|\det \frac{\partial f_\theta(v_{i,t}; h_{i,t})}{\partial v_{i,t}}\big|$ whereas variational methods introduce an ELBO approximation gap relative to the exact likelihood. Although variational methods are powerful and can be consistent under certain conditions [Wang and Blei, 2019], in our experiments this gap manifested as behavior consistent with posterior collapse to a null effect on the fully mediated CVD Risk benchmark (HR $= 1.005$ vs true 0.831), suggesting that these conditions may not hold in our setting (Section 4.2, Appendix K).

Temporal encoding via LSTM. The CausalMAF is conditioned on an LSTM encoder:

$$h_{i,t} = \mathrm{LSTM}_\psi\big(h_{i,t-1},\, x_{i,t-1}\big), \qquad v_{i,t} \sim \mathrm{CausalMAF}_\theta(\cdot \mid h_{i,t}), \tag{3}$$

with $h_{i,1}$ initialized from baseline covariates. The hidden state $h_{i,t}$ summarizes the covariate history before time t, capturing disease progression and time-varying confounding. The DAG $\mathcal{G}$ governs the contemporaneous factorization $P(x_{i,t}, a_{i,t}, y_{i,t} \mid h_{i,t})$, enabling valid do-calculus [Pearl, 2009].

Counterfactual inference via AAP. Given the trained flow, counterfactuals are computed via Pearl's AAP procedure [Pearl, 2009]. *Abduction*: invert the flow exactly to recover the exogenous noise $z_{i,t} = f_\theta(v_{i,t}; h_{i,t})$, capturing all individual-specific factors not explained by observed covariates. *Action*: replace $a_{i,t} \leftarrow a'$; because the autoregressive ordering respects $\mathcal{G}$, the intervention propagates only to causal descendants of $a_{i,t}$; $z_{i,t}$ is held fixed across both arms, implementing the surgical intervention of the do-operator. *Prediction*: propagate through $f_\theta^{-1}(\cdot\,; h_{i,t})$ via the DAG ordering to obtain $v_{i,t}^{\mathrm{cf}}$ and $\hat{\tau}_{i,t} = \hat{y}_{i,t}(1) - \hat{y}_{i,t}(0)$. Because inversion is exact, both arms share the same $z_{i,t}$, coupling a patient's two predicted potential outcomes through their own abducted noise; the encoders of our CVAE baselines instead return a distribution over z. Under (i)–(iv) this yields consistent population effects, whereas individual predictions $\hat{\tau}_{i,t}$ are model-based rather than identified.

3.3 LLM-driven evolutionary imputation under MNAR

This stage constructs the completed dataset $\hat{\mathcal{D}}$ from the incomplete $\widetilde{\mathcal{D}}$ via evolutionary program search. At each iteration k, the LLM writes the source code of a Python imputation program $g^{(k)}$, which preserves observed entries and non-target columns; $g^{(k)}$ is executed, scored by the proxy below, and kept only if it beats the current best (Algorithm 1; example in Appendix L.3).

Self-supervised proxy score. Let $\Omega_{\mathrm{obs}} = \{(i,t,j) : m_{i,t,j} = 0\}$. We draw a proxy holdout $\Omega_p \subset \Omega_{\mathrm{obs}}$ by masking each observed cell independently with probability $\rho \in (0,1)$ (contiguous-run masking in real EHR; Appendix L), and let $\widetilde{\mathcal{D}}_-$ denote $\widetilde{\mathcal{D}}$ with these cells additionally masked. The composite proxy score penalizes both pointwise reconstruction error and distortion of covariate–treatment and covariate–outcome correlations:

$$s(g) = \underbrace{\sqrt{\frac{1}{|\Omega_p|}\sum_{(i,t,j)\in\Omega_p}\Big(\big[g(\widetilde{\mathcal{D}}_-)\big]_{i,t,j} - \tilde{x}_{i,t,j}\Big)^2}}_{\mathrm{RMSE}(g)} + \lambda_Y \Delta_Y(g) + \lambda_T \Delta_T(g), \tag{4}$$

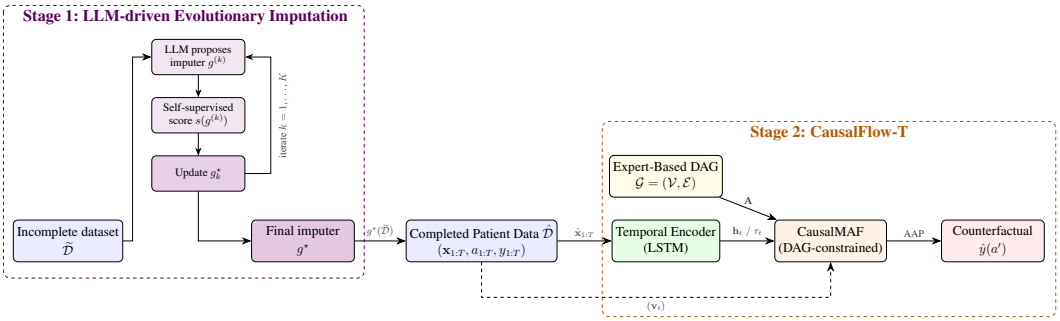


Figure 1: **Stage 1 (LLM-driven Evolutionary Imputation):** an LLM iteratively proposes candidate imputers $g^{(k)}$, scores them via a self-supervised proxy $s(g^{(k)})$, and updates the running best $g_k^\star$; after K rounds, $g^\star$ produces $\hat{\mathcal{D}}$. **Stage 2 (CausalFlow-T):** a temporal encoder conditions a DAG-constrained CausalMAF on patient history; counterfactual outcomes $\hat{y}(a')$ are obtained via AAP.

where, with $\hat{x}_j$ the j-th column of $g(\widetilde{\mathcal{D}}_-)$, $\Delta_{Y,j}(g) = |\text{Corr}(\hat{x}_j, y) - \text{Corr}(\tilde{x}_j, y)|$ and $\Delta_{T,j}(g) = |\text{Corr}(\hat{x}_j, a) - \text{Corr}(\tilde{x}_j, a)|$ measure how much imputation distorts biomarker–outcome and biomarker–treatment correlations, with column averages Δ_Y, Δ_T (weights λ_Y and λ_T reported in Table 6). Penalizing Δ_Y targets the correlation structure that CausalFlow-T's do-calculus depends on; penalizing Δ_T targets the propensity-related correlations that drive confounding adjustment.

Evolutionary loop. Starting from a deterministic seed imputer $g^{(0)}$, with $g_0^\star = g^{(0)}$, at each iteration k the LLM receives the current-best candidate $g_{k-1}^\star$, its score $s(g_{k-1}^\star)$, and a compact search history summary, then proposes a variation $g^{(k)}$. The update rule is

$$g_k^\star = \begin{cases} g^{(k)} & \text{if } s(g^{(k)}) < s(g_{k-1}^\star), \\ g_{k-1}^\star & \text{otherwise.} \end{cases} \tag{5}$$

After K iterations, $g^\star = g_K^\star$ is applied to $\widetilde{\mathcal{D}}$ to produce $\hat{\mathcal{D}}$. We use a single-parent evolutionary scheme, maintaining a single current-best imputer; implementation details are reported in Appendix L.1.

4 Experiments

We evaluate the pipeline in two steps: (i) CausalFlow-T on complete data with known counterfactuals, and (ii) four imputation strategies with CausalFlow-T fixed under 30%, 50%, and 80% MNAR missingness, including three LLM-based variants. Three findings emerge: (1) in our benchmarks CausalFlow-T is the only model satisfying all reliability criteria, revealing two distinct failure modes (Section 4.2); (2) LLM-driven imputation achieves the best pooled performance across biomarker and causal metrics, with GPT-5.4 strongest overall and the other LLMs demonstrating robustness and reconstruction–causal tradeoffs (Section 4.3); (3) the full pipeline recovers per-protocol treatment effects from incomplete real-world EHR data consistent with RCT evidence (Section 5). Experiments were run on a single HPC node (Appendix N).

4.1 Experimental setup and evaluation protocol

Datasets. We evaluate CausalFlow-T on four synthetic datasets with ground-truth counterfactuals of increasing structural complexity: **(1)** Simple 3-Node (N=10k, T=5; linear dependencies; positive control); **(2)** LDL Toy (N=10k, T=5; heterogeneous saturating effect, autoregressive AR(1) dynamics, selection-on-gain confounding; primary stress-test for subgroup calibration); **(3)** Cox Survival (N=30k, T=11; time-dependent hazard, bimodal covariate heterogeneity); and **(4)** CVD Risk Toy (N=50k, T=10; 17-node DAG, fully mediated treatment effect via systolic blood pressure (SBP) lags, HR ≈ 0.82; hardest benchmark). For imputation, we use a semi-synthetic dataset built on a real-world EHR backbone [Chmiel, 2011] with ten synthetic longitudinal biomarkers, a known benchmark ATE (-3.484), and MNAR missingness introduced at 30%, 50%, and 80%. All datasets are described in Appendix A.

Causal inference baselines. We compare against CVAE, GNN-CVAE, and TARNet, all sharing the same LSTM temporal encoder; discriminative sequence models (Causal Transformer, R-MSN, CRN) are excluded by construction as they fail the three jointly necessary criteria for distributional counterfactual evaluation (Appendix B.1).

Imputation baselines. Four imputation strategies are evaluated with CausalFlow-T held fixed: LOCF (carry-forward heuristic), MissForest (non-parametric random-forest), CausalCFM (a DAG-aware flow matching baseline, Appendix C), and LLM-driven evolutionary imputation with three backends (GPT-5.4, Qwen3.5-Plus, and GPT-OSS-120b).

Evaluation protocols. CausalFlow-T is validated on four metrics exposing failure modes invisible to predictive error: subgroup calibration $|\text{Bias}_{Q1}|/\text{MAE}_{Q1}$, where MAE is the mean absolute error (ratio ≈ 1 signals systematic confounding failure), arm reconstruction error Err_a, tail variance ratio VR_{Q4} (collapse < 1, overdispersion > 1), and hazard ratio recovery $|\text{HR}_{\text{true}} - \text{HR}_{\text{pred}}|$.These criteria build on PEHE [Hill, 2011, Shalit et al., 2017] and CATE calibration [Athey and Wager, 2019, Xu and Yadlowsky, 2022](Appendix D.1). Imputation is evaluated on two complementary layers: biomarker quality (MAE, root mean squared error RMSE, autocorrelation error $|\Delta\text{AC}|$, consecutive-step error CSE, and correlation inflation $|\text{inf}_Y|$) and downstream causal quality (Q1 MAE, $\bar{\text{Err}}_a$, and $|\widehat{\text{ATE}} - \text{ATE}^*|$), with CausalFlow-T held fixed so differences reflect imputation alone (formal definitions in Appendix D).

4.2 CausalFlow-T: Two non-overlapping failure modes

Table 1 summarizes per-metric ranks and binary reliability criteria across all four synthetic datasets; full numerical results per dataset are in Appendix G (Tables 8–9). All models perform comparably on the positive control (Simple 3-Node), confirming the DAG constraint does not degrade performance when confounding is absent; two non-overlapping failure modes emerge on harder benchmarks.

Table 1: Per-metric ranks (1 = best, 5 = worst) across four synthetic datasets and binary reliability criteria (✓/✗). Arm: mean per-arm error; VR: $|\text{VR}_{Q4} - 1|$. Reliability: *Bias* ratio < 0.5; *Tail*: mean $|\text{VR}_{Q4} - 1| < 0.5$ across all benchmarks; *HR*: correct direction; *Arm*: best on ≥ 2 benchmarks; *Stable*: no explosion or inversion. Metric definitions are in Appendix D.1.

	Simple		**LDL**				**Cox**			**CVD Risk**			**Mean rank**	**Reliability**				
Model	MAE	VR	MAE	Bias	VR	Arm	MAE	Arm	HR	MAE	VR	HR		Bias	Tail	HR	Arm	Stable
CausalFlow-T	4	**1.5**	2	2	**1**	3	3	**1.5**	2	**1**	2	2	**2.08**	✓	✓	✓	✓	✓
NF (no DAG)	2	**1.5**	3	**1**	2	4	**1**	**1.5**	4	2	4	**1**	2.25	✓	✗	✓	✗	✗
GNN-CVAE	**1**	3	5	5	3	5	4	5	3	3.5	4	4	3.79	✗	✗	✗	✗	✗
CVAE	3	5	4	4	4	2	2	4	**1**	3.5	4	3	3.29	✗	✗	✗	✗	✗
TARNet	5	4	**1**	3	5	**1**	5	3	5	5	**1**	5	3.58	✗	✓	✗	✗	✗

Failure mode 1: Confounding separation benefits from causal structure. On LDL Toy, GNN-CVAE and CVAE reach $|\text{Bias}/\text{MAE}|_{\text{Q1}} \approx 1.0$ (every low-effect error is systematic), while NF (no DAG) reduces this to 0.195 via exact inference alone, slightly below CausalFlow-T (0.270), both clearing the 0.5 threshold; the benefit of the DAG constraint for confounding separation appears on the semi-synthetic benchmark (0.316 vs 0.883) and as stability under mediation (CVD Risk). CausalFlow-T further achieves $\text{VR}_{\text{Q4}} = 1.049 \pm 0.033$ (rank 1), the only model preserving tail variance without collapse or inflation. TARNet's lowest absolute MAE (0.407 ± 0.117) has a $\text{VR}_{\text{Q4}} = 0.611$, illustrating why MAE rank alone is an insufficient criterion.

Failure mode 2: Structural propagation from exact inference. CVD Risk stress-tests DAG constraints because the treatment effect is entirely mediated, so models that do not respect the causal ordering are expected to collapse or diverge. GNN-CVAE and CVAE both rely on the ELBO and estimated HR ≈ 1.005 (a null effect), consistent with posterior collapse here; NF (no DAG) performs exact inference without structural constraint, routing the treatment signal through arbitrary pathways and yielding $\text{HR} = 0.834 \pm 0.324$ (clinically meaningless); TARNet, lacking both, inverts the effect entirely ($\text{HR} = 1.133 \pm 0.148$, predicting harm where there is benefit). In this benchmark, CausalFlow-T, combining DAG factorization with exact normalizing-flow inference, forces the treatment signal through the correct mediation pathways and recovers $\text{HR} = 0.786 \pm 0.051$ vs

true 0.831, with arm errors below competitors. On the other hand, dequantizing binary survival outcomes (Cox benchmark) introduces a calibration cost that explains CVAE's MAE advantage on that benchmark; on structurally meaningful metrics CausalFlow-T leads with best arm-1 error (0.014 ± 0.002), while recovering the HR reasonably closely (0.866 ± 0.011 vs. true 0.887); discrete flow extensions are a natural direction for future work (Section 6).

Semi-synthetic validation. CausalFlow-T and GNN-CVAE are the only models passing the bias threshold (0.316 and 0.351); CausalFlow-T is further the only model with near-perfect variance preservation ($\mathrm{VR}_{\mathrm{Q4}} = 1.006 \pm 0.009$) while GNN-CVAE collapses ($\mathrm{VR} = 0.616$), confirming that low systematic error alone is insufficient for distributional reliability. TARNet achieves the best arm-1 reconstruction (0.081 ± 0.012) but fails the bias threshold (0.708) and collapses variance ($\mathrm{VR} = 0.608$), reinforcing that factual accuracy does not imply counterfactual reliability. Removing the DAG constraint degrades the bias ratio to 0.883 with the largest ATE deviation (-3.382 ± 0.251), indicating that structural supervision was beneficial under this realistic clinical distribution. Full results can be found in Appendix F.

CausalFlow-T for causal inference. CausalFlow-T is the only model in our comparison that satisfies all five reliability criteria (Table 1; mean rank 2.08). Similar results are found when stress-testing in the real-world EHR semi-synthetic dataset. Competing models fail along at least one axis (trading bias for variance collapse, capturing trajectories with subgroup error, or achieving low MAE while reversing treatment direction) and none meets more than two criteria. We therefore fix CausalFlow-T in subsequent imputation experiments so that differences in ATE recovery reflect imputation quality alone.

4.3 Imputation under MNAR missingness

Full numerical results are in Appendix H (Tables 11–12) while Table 2 summarizes per-metric ranks across both evaluation layers.

Table 2: Per-metric ranks (1 = best, 6 = worst) across 30%, 50%, and 80% MNAR missingness, computed from point estimates. Ties receive average ranks. The ensemble ATE residual is computed from the mean ATE across training seeds ($\mathrm{ATE}^{*} = -3.484$). Metric definitions are in Appendix D.

		Biomarker quality					**Downstream causal**			**Mean rank**
Miss.	**Method**	MAE	RMSE	$\lvert\Delta\mathrm{AC}\rvert$	CSE	$\lvert\mathrm{inf}_Y\rvert$	Q1 MAE	$\overline{\mathrm{Err}}_a$	$\lvert\Delta\mathrm{ATE}\rvert$	
30%	LOCF	6	6	6	6	6	5	4	4	5.38
	MissForest	4	4	2	4	5	6	6	6	4.62
	CausalCFM	5	5	5	5	2	**1**	**1**	2	3.25
	GPT-5.4	**1**	**1**	4	**1**	4	4	3	**1**	**2.38**
	Qwen3.5-Plus	3	3	**1**	3	**1**	3	5	5	3.00
	GPT-OSS-120b	2	2	3	2	3	2	2	3	**2.38**
50%	LOCF	6	6	6	6	6	4	3	**1**	4.75
	MissForest	3	3	**1**	3	5	6	6	5	4.00
	CausalCFM	5	5	3	5	**1**	5	4	3	3.88
	GPT-5.4	**1**	**1**	2	**1**	3	3	2	4	**2.12**
	Qwen3.5-Plus	4	4	5	4	2	**1**	**1**	2	2.88
	GPT-OSS-120b	2	2	4	2	4	2	5	6	3.38
80%	LOCF	6	6	6	5	6	3	**1.5**	**1**	4.31
	MissForest	3	3	**1**	3	2	**1**	4	4	2.62
	CausalCFM	5	5	2	6	**1**	6	6	6	4.62
	GPT-5.4	**1**	**1**	5	**1**	5	4	3	3	2.88
	Qwen3.5-Plus	4	4	4	4	3	5	5	5	4.25
	GPT-OSS-120b	2	2	3	2	4	2	**1.5**	2	**2.31**

Method	**Mean rank**
LOCF	4.81
MissForest	3.75
CausalCFM	3.92
GPT-5.4	**2.46**
Qwen3.5-Plus	3.38
GPT-OSS-120b	2.69

Biomarker reconstruction. GPT-5.4 achieves the lowest pointwise MAE and RMSE at every missingness level with only modest degradation from 30% to 80% (MAE 1.914→2.059). Qwen3.5-Plus best preserves lag-1 autocorrelation at 30%, while MissForest is closest at 50% and 80%; CausalCFM best preserves biomarker–outcome correlation at 50% and 80% [$\lvert\mathrm{inf}_Y\rvert \approx 0.01$]. GPT-OSS-120b tracks GPT-5.4 closely on pointwise reconstruction at 50% and 80%, supporting robustness of the LLM-driven strategy beyond a single model.

Downstream causal recovery and the reconstruction: causal tradeoff. Pointwise reconstruction quality and causal correlation preservation are not aligned objectives, and Table 2 exposes this tradeoff directly. At 30%, GPT-5.4 gives the closest ATE recovery ($|\Delta\text{ATE}| = 0.003$) while CausalCFM leads on subgroup and arm reconstruction; at 50%, Qwen3.5-Plus gives the strongest causal reconstruction ($\bar{\text{Err}}_a = 0.069$) with near-best ATE residual (0.013 vs. 0.007 for LOCF); at 80%, GPT-5.4 retains competitive ATE recovery ($|\Delta\text{ATE}_{\text{ens}}| = 0.124$), although LOCF attains a smaller residual (0.052). Above 50% missingness (the regime of most real EHR data [Sun et al., 2024]), LLM imputation combines strong biomarker reconstruction with competitive ATE recovery.

LLM-driven evolutionary imputation. LLM-driven imputation achieves the strongest pooled performance across 30–80% MNAR (mean rank: GPT-5.4 2.46, GPT-OSS-120b 2.69, Qwen3.5-Plus 3.38; MissForest 3.75, CausalCFM 3.92, LOCF 4.81). GPT-5.4 is selected as the default; CausalCFM remains a useful comparator for causal-fidelity diagnostics at lower missingness. Appendix Figure 7 shows GPT-5.4 search progress, and Appendix Tables 11–12 compare the final imputer with the first executable GPT-5.4 proposal. Appendix Table 13 additionally benchmarks ATE recovery against a marginal structural model (MSM; IPTW/IPCW with weighted GEE) under LOCF and GPT-5.4 imputation. Search-configuration and bootstrap sensitivity are shown in Appendix J (Tables 15–17).

5 Real-world application: Causal inference from healthcare data

TTE in incomplete EHR data: On controlled semi-synthetic benchmarks, our pipeline recovers the ATE under MNAR missingness, as verified against ground-truth counterfactuals. Real-world EHR data, however, is collected for clinical and administrative rather than causal purposes: unlike in RCTs, treatment is not randomized, follow-up is irregular, treatment deviations are informative, and missingness is rarely MAR. We therefore apply the two-stage pipeline (validated LLM imputation followed by CausalFlow-T) to FIRE, Switzerland's most comprehensive primary care EHR database [Chmiel, 2011] (Appendix A.6). Within the TTE framework [Hernán and Robins, 2016], we estimate the per-protocol effect among GLP-1RA ($n = 3{,}722$) versus SGLT-2i ($n = 2{,}392$) initiators on body weight loss (kg) at 1 year in adults with T2D (Appendix A.6.2). Causal assumptions are encoded in an expert-based DAG and missing measurements are imputed via our validated LLM-based strategy (Appendix A.6.2). Privacy and regulatory constraints preclude public release of the dataset.

Per-protocol effect estimation and external consistency: Figure 2 shows both treatment arms exhibit monotonic weight loss from initiation to 1-year, estimated using the proposed pipeline (Appendix A.6.2). At 1-year, the model recovers $\widehat{\Delta W}^{\text{GLP-1RA}} = -3.95$ kg [95% CI -4.28, -3.62] and $\widehat{\Delta W}^{\text{SGLT-2i}} = -3.02$ kg [95% CI -3.33, -2.70], yielding a per-protocol effect of -0.93 kg [95% CI -1.26, -0.70] , favoring GLP-1RAs (bootstrap uncertainty analysis in Appendix Tables 16–17). This estimate is consistent with the SUSTAIN 8 trial (an RCT comparing a GLP-1RA [semaglutide] vs an SGLT-2i [canagliflozin]) in T2D, reporting 1-year body weight difference of -1.06 kg (95% CI -1.76, -0.36) [Lingvay et al., 2019]. Our study extends RCT evidence from specific agents in selected trial participants to the full GLP-1RA and SGLT-2i classes used in routine Swiss primary care, where mixed agents and $> 50\%$ MNAR missingness are expected to attenuate the contrast compared to the SUSTAIN 8 trial estimate [Lingvay et al., 2019]. While ground-truth counterfactuals are unavailable, consistency with RCT evidence supports the plausibility of the per-protocol effects estimated by the proposed pipeline from incomplete real-world EHR data.

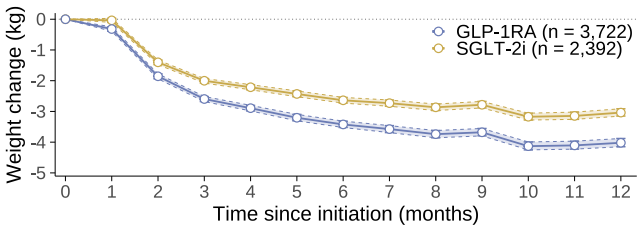


Figure 2: Weight change over 1-year for adults with type 2 diabetes initiating GLP-1 receptor agonists vs SGLT-2 inhibitors, estimated from real-world healthcare data using the two-stage pipeline.

6 Discussion

Our two-stage pipeline jointly handles two compounding challenges of observational data: time-varying confounding and high-rate MNAR missingness; validating CausalFlow-T, a DAG-constrained normalizing flow with exact inversion, on complete data before adding the LLM-driven evolutionary imputer isolates each component's impact.

A central finding is not that CausalFlow-T outperforms competing models, but that DAG constraints and exact inference addressed distinct failure modes, and neither alone was sufficient. Simpler benchmarks may overstate robustness, as strong factual prediction can mask failures in subgroup calibration, tail variance, and mediation-consistent effect propagation. Our five-criterion reliability framework, stricter than any single metric, could benchmark longitudinal causal models.

The imputation results reveal a complementary insight: pointwise reconstruction and causal correlation preservation are not aligned objectives. LLM-based methods, especially GPT-5.4, dominate reconstruction and temporal extrapolation, while CausalCFM is better on selected causal-fidelity diagnostics, notably biomarker–outcome correlation preservation. This tradeoff, invisible to standard imputation benchmarks, makes joint biomarker–causal evaluation necessary for causal pipelines under missingness; future methods could optimize it explicitly.

Applied to $6{,}114$ adults with T2D in Swiss primary care, the pipeline recovers a per-protocol effect of -0.93 kg [95% CI -1.26, -0.70] favoring GLP-1RAs over SGLT-2is, consistent with RCT evidence [Lingvay et al., 2019]. Although ground-truth counterfactuals are unavailable, this suggests the pipeline yields clinically plausible estimates from incomplete EHR data, extending causal evidence beyond RCTs' selected participants and conditions.

Several limitations suggest future directions. First, dequantizing binary outcomes incurs a calibration cost on survival endpoints, explaining CVAE's lower Cox Survival MAE (Appendix K); discrete or hybrid flows are a natural extension. Second, the DAG is assumed fixed and expert-specified; single-edge removals mainly increase bias or cross-seed instability, a possible signal for data-driven DAG refinement (Appendix I), while edge-direction errors and latent confounding remain open. Third, all three LLM backends share one evolutionary search protocol, on which performance may depend, and need API access during search, although the selected imputer is a frozen, offline module and the open-weight backend performs competitively; alternatives such as recombination and multi-objective acceptance merit exploration. Finally, like all observational studies, our real-world analysis assumes conditional exchangeability given the expert-specified adjustment set: prescribing preferences may leave residual confounding, no negative control was available in the current data extract, and our reliability criteria are validated only on semi-synthetic benchmarks. Independent replication, transfer to other EHR systems, and subgroup analyses (age, sex, comorbidity) remain future work.

7 Conclusion

We introduced a two-stage pipeline for treatment effect estimation from incomplete longitudinal EHR data, combining CausalFlow-T (a DAG-constrained normalizing flow with exact inversion) with an LLM-driven evolutionary imputer for MNAR missingness. Controlled ablations show that DAG constraints and exact inference address distinct failure modes, with CausalFlow-T being the only model that satisfies all five reliability criteria simultaneously. Joint biomarker–causal evaluation further reveals causal implications of imputation choices that reconstruction-only metrics overlook. Applied to real-world EHR data, the pipeline recovers a per-protocol effect consistent with RCT evidence despite substantial MNAR missingness. These findings show that jointly addressing causal structure and MNAR missingness is both necessary and feasible. They support ML-based treatment effect estimation from observational data as a practical complement to RCTs, extending causal inference to real-world populations and care settings where trials are infeasible or insufficient, and real-world evidence is needed to inform clinical and policy decision-making.

Acknowledgments and Disclosure of Funding

We thank the Institute of Primary Care (University of Zürich) for providing the dataset used in this study. This work was supported by SNSF under grant CRSII5_216636. The authors declare no competing interests.

References

American Diabetes Association. 10. Cardiovascular Disease and Risk Management: Standards of Care in Diabetes—2025. *Diabetes Care*, 48(Supplement_1):S207–S238, 1 2025. ISSN 0149-5992. doi: 10.2337/DC25-S010. URL https://dx.doi.org/10.2337/dc25-S010.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *Observational studies*, 5(2):37–51, 2019.

Donika Balaj, Jakob M Burgstaller, Audrey Wallnöfer, Katja Weiss, Oliver Senn, Thomas Rosemann, Thomas Grischott, Stefan Markun, FIRE research group, et al. Leveraging free-text diagnoses to identify patients with diabetes mellitus, obesity or dyslipidaemia–a cross-sectional study in a large swiss primary care database. *Swiss Medical Weekly*, 155(2):3360–3360, 2025.

James E Barrett, Aylin Cakiroglu, Catey Bunce, Anoop Shah, and Spiros Denaxas. Selective recruitment designs for improving observational studies using electronic health records. *Statistics in Medicine*, 39(19):2556–2567, 2020.

Ioana Bica, Ahmed M. Alaa, James Jordon, and Mihaela van der Schaar. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *International Conference on Learning Representations (ICLR)*, 2020.

James R. Carpenter, James H. Roger, and Michael G. Kenward. Analysis of longitudinal trials with protocol deviation: A framework for relevant, accessible assumptions, and inference via multiple imputation. *Journal of Biopharmaceutical Statistics*, 23(6):1352–1371, 2013. doi: 10.1080/10543406.2013.834911.

Patrick Chao, Patrick Blöbaum, Sapan Patel, and Shiva Prasad Kasiviswanathan. Modeling causal mechanisms with diffusion models for interventional and counterfactual queries. *arXiv preprint arXiv:2302.00860*, 2023.

Moshinsky Chmiel. The fire project. *Swiss medical weekly*, 141(0304):w13142–w13142, 2011.

P. Diggle and M. G. Kenward. Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 43(1):49–73, 1994. doi: 10.2307/2986113.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Garrett M Fitzmaurice, Nan M Laird, and James H Ware. *Applied longitudinal analysis*. John Wiley & Sons, 2012.

Richard D Gill and James M Robins. Causal inference for complex longitudinal data: the continuous case. *Annals of Statistics*, pages 1785–1811, 2001.

Xinrui He, Yikun Ban, Jiaru Zou, Tianxin Wei, Curtiss Cook, and Jingrui He. Llm-forest: Ensemble learning of llms with graph-augmented prompts for data imputation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6921–6936, 2025.

Miguel A Hernán and James M Robins. Practice of Epidemiology Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology*, 183 (8):758–764, April 2016. doi: 10.1093/aje/kwv254. URL https://academic.oup.com/aje/article/183/8/758/1739860.

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

Miguel A Hernán, Wei Wang, and David E Leaf. Target trial emulation: a framework for causal inference from observational data. *Jama*, 328(24):2446–2447, 2022.

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

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

International Diabetes Federation. IDF Diabetes Atlas, 11th edition, 2025. URL https://diabetesatlas.org/media/uploads/sites/3/2025/04/IDF_Atlas_11th_Edition_2025.pdf. Accessed 9 September 2025.

Adrián Javaloy, Pablo Sánchez-Martín, and Isabel Valera. Causal normalizing flows: from theory to practice. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Ilyes Khemakhem, Ricardo Pio Monti, Robert Leech, and Aapo Hyvärinen. Causal autoregressive flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3520–3528, 2021.

Robert Tjarko Lange, Yuki Imajuku, and Edoardo Cetin. ShinkaEvolve: Towards open-ended and sample-efficient program evolution. *arXiv preprint arXiv:2509.19349*, 2025.

Bryan Lim, Ahmed M. Alaa, and Mihaela van der Schaar. Forecasting treatment responses over time using recurrent marginal structural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

Ildiko Lingvay, Andrei-Mircea Catarig, Juan P Frias, Harish Kumar, Nanna L Lausvig, Carel W le Roux, Desirée Thielke, Adie Viljoen, and Rory J McCrimmon. Efficacy and safety of once-weekly semaglutide versus once-daily canagliflozin as add-on to metformin in patients with type 2 diabetes (SUSTAIN 8): a double-blind, phase 3b, randomised controlled trial. *The Lancet Diabetes & Endocrinology*, 7(11):834–844, 2019. doi: 10.1016/S2213-8587(19)30311-0.

Roderick J. A. Little. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*, 88(421):125–134, 1993. doi: 10.1080/01621459.1993.10594302.

Fei Liu, Xialiang Tong, Mingxuan Yuan, Xi Lin, Fu Luo, Zhenkun Wang, Zhichao Lu, and Qingfu Zhang. Evolution of heuristics: Towards efficient automatic algorithm design using large language model. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 32201–32223. PMLR, 2024.

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 (NeurIPS)*, volume 30, 2017.

Arthur Dantas Mangussi, Ricardo Cardoso Pereira, Ana Carolina Lorena, and Pedro Henriques Abreu. Large language models for missing data imputation: Understanding behavior, hallucination effects, and control mechanisms. *arXiv preprint arXiv:2603.22332*, 2026.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *International Conference on Machine Learning (ICML)*, pages 15293–15329, 2022.

Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Bounds on representation-induced confounding bias for treatment effect estimation. In *International Conference on Learning Representations*, 2024.

Y. Mohammed, H. Abdelkader, P. Pławiak, and M. Hammad. A novel model to optimize multiple imputation algorithm for missing data using evolution methods. *Biomedical Signal Processing and Control*, 76:103661, 2022. doi: 10.1016/j.bspc.2022.103661.

Karthika Mohan and Judea Pearl. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

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

James M Robins. Causal inference from complex longitudinal data. In *Latent variable modeling and applications to causality*, pages 69–117. Springer, 1997.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M. Pawan Kumar, Emilien Dupont, Francisco J. R. Ruiz, Jordan S. Ellenberg, Pengming Wang, Omar Fawzi, Pushmeet Kohli, and Alhussein Fawzi. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024. doi: 10.1038/s41586-023-06924-6.

Sebastian Schneeweiss and Elisabetta Patorno. Conducting Real-world Evidence Studies on the Clinical Outcomes of Diabetes Treatments. *Endocrine reviews*, 42(5):658–690, 10 2021. ISSN 1945-7189. doi: 10.1210/ENDREV/BNAB007. URL `https://pubmed.ncbi.nlm.nih.gov/33710268/`.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *International Conference on Machine Learning (ICML)*, 2017.

Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338, 2009.

Minghui Sun, Matthew M Engelhard, Armando D Bedoya, and Benjamin A Goldstein. Incorporating informatively collected laboratory data from ehr in clinical prediction models. *BMC Medical Informatics and Decision Making*, 24(1):206, 2024.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Yizhe Xu and Steve Yadlowsky. Calibration error for heterogeneous treatment effects. In *International Conference on Artificial Intelligence and Statistics*, pages 9280–9303. PMLR, 2022.

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.

Yizhao Zhou, Jiasheng Shi, Ronen Stein, Xiaokang Liu, Robert N Baldassano, Christopher B Forrest, Yong Chen, and Jing Huang. Missing data matter: an empirical evaluation of the impacts of missing ehr data in comparative effectiveness research. *Journal of the American Medical Informatics Association*, 30(7):1246–1256, 2023.

A Dataset Descriptions

All four benchmarks are fully synthetic, so both ground-truth outcomes $Y(0)$ and $Y(1)$ are available for every test patient, enabling exact evaluation of counterfactual predictions. Train/test splits of 80/20 are used throughout.

A.1 Simple 3-Node

The dataset contains N patients observed over T timesteps (default N=10,000, T=5). Three variables are simulated:

- **Confounder** $X_1^{(t)} \sim \mathcal{N}(t/4,\, 1)$: a single continuous covariate whose mean drifts linearly with time, inducing non-stationarity.
- **Treatment** $A \in \{0, 1\}$: a binary treatment assigned once at baseline and kept constant. The assignment probability is a nonlinear function of X_1:

$$P(A = 1 \mid X_1) = \begin{cases} 0.2 & \text{if } \bar{X}_1 > 2.5, \\ 0.8 & \text{otherwise,} \end{cases}$$

  where $\bar{X}_1 = \mathbb{E}_t[X_1^{(t)}]$ and the intermediate variable $X_1^2 - \sin(X_1) + \varepsilon$ ($\varepsilon \sim \mathcal{N}(0, 0.25)$) mediates the propensity. This creates nonlinear confounding with a treatment probability that flips between 20% and 80%.
- **Outcome** $Y^{(t)}$: a continuous outcome generated as

$$Y^{(t)} = \begin{cases} 3X_1^{(t)} + 0.25\,(t/T) + \varepsilon_0 & A = 0, \\ 3X_1^{(t)} - 0.50\,(t/T) + \varepsilon_1 & A = 1, \end{cases}$$

  where $\varepsilon_0, \varepsilon_1 \sim \mathcal{N}(0, 1)$. Treatment reduces the time trend by $0.75\,(t/T)$ relative to control, inducing a modest, time-growing ITE.

Causal graph. The DAG is $X_1 \to A \to Y$, $X_1 \to Y$, encoded as a 3×3 adjacency matrix.

Purpose. A positive control with a low confounded structure (linear dependencies mostly). Because the causal ordering matches the variable index order and confounding is low, all five models are expected to perform comparably. The dataset verifies that the framework performs well even in the simplest case.

A.2 LDL Toy

Generated dataset with N=10,000 patients, T=5 timesteps. The outcome Y represents a continuous biomarker (analogous to LDL cholesterol) subject to a heterogeneous drug response, log-saturation pharmacokinetics, autoregressive dynamics, and selection-on-gain confounding.

Specifically, the response magnitude at baseline is $\delta_i \sim \mathrm{Uniform}(0.20, 0.60)$, giving each patient an individual treatment-effect scale (20%–60% reduction). The dose-response saturates over time according to

$$\mathrm{sat}(t) = \frac{\log(1 + \lambda_{\mathrm{sat}} t)}{\log(1 + \lambda_{\mathrm{sat}} T)}, \qquad \lambda_{\mathrm{sat}} = 0.4,$$

so early timesteps carry most of the treatment signal. Temporal dependence is introduced via a first-order AR(1) process with coefficient $\rho_{\mathrm{AR}} = 0.75$, such that each observation exhibits strong dependence on its immediate predecessor.

Confounding is selection on gain: patients who respond more strongly (δ_i above median) are preferentially assigned to treatment, so the treated group has a higher true effect than the control group.

Purpose. Tests calibration of heterogeneous treatment effects. The combination of log-saturation (non-constant effect over time), AR(1) dynamics (temporal correlation), and selection-on-gain confounding makes this the primary benchmark for the structural-failure and variance-collapse diagnostics.

A.3 Cox Survival Toy

The dataset contains N=30,000 patients, T=11 timesteps (including t=0), d=5 covariates.

1. **Covariates.** Five covariates are drawn from bimodal Gaussian mixtures:
$$X_{ij} \mid Z_{ij} = \begin{cases} \mathcal{N}(\mu_1^{(j)},\, 0.01) & Z_{ij} = 1, \\ \mathcal{N}(\mu_2^{(j)},\, 0.04) & Z_{ij} = 0, \end{cases}$$
with $Z_{ij} \sim \mathrm{Bernoulli}(0.5)$ and $(\mu_1^{(1)}, \ldots) = (3, 1, 2, 5, 0)$, $(\mu_2^{(1)}, \ldots) = (0, 2, 4, 5, 5)$.
2. **Hazard function.** Per-patient log hazard ratios are
$$\log \mathrm{HR}_i = \boldsymbol{\beta}^\top X_i, \quad \boldsymbol{\beta} = (0.2,\, -0.4,\, -0.3,\, 0.3,\, -0.5),$$
giving $\mathrm{HR}_i = \exp(\boldsymbol{\beta}^\top X_i)$. The baseline hazard is $h_0 = 0.5$.
3. **Treatment assignment (confounding).** Patients with $\mathrm{HR}_i \geq 0.5$ (i.e., higher risk) are treated with probability 0.70; the remainder with probability 0.30.
4. **Treatment effect.** Treatment reduces the individual hazard ratio by 30%: $\mathrm{HR}_i^{\text{treated}} = 0.70\,\mathrm{HR}_i$.
5. **Event times.** Survival times are drawn from an exponential distribution, $\mathcal{E}(1/[h_0\,\mathrm{HR}_i])$, and converted to a binary event indicator at each of the 11 timesteps. Censoring is applied only at the end of the study (all weight is on the final timestep), so within-study censoring is minimal.
6. **Counterfactuals.** The smooth individual cumulative incidence function $F_i(t) = 1 - \exp(-h_0\,\mathrm{HR}_i\,t)$ is computed for both the always-treated and never-treated hazards and stored as ground-truth potential outcomes for evaluation.

Purpose. Tests survival-model calibration under continuous confounding. The bimodal covariate structure creates two patient clusters. The correct arm reconstruction requires the model to disentangle hazard ratio heterogeneity from the selection-on-risk confounding.

A.4 CVD Risk Toy

The CVD Risk Toy is the most complex benchmark, designed to emulate a real-world antihypertensive target-trial setting. It is synthetic: the causal structure, covariate relationships, treatment dynamics, and outcome model are grounded in established cardiovascular epidemiology and RCT evidence, while outcomes are re-generated under a fully specified data-generating process to provide a known ground truth (including exact counterfactual trajectories and a precise hazard ratio ($\mathrm{HR} \approx 0.83$)). The dataset comprises N=50,000 patients over T=10 observation years, with 15 time-varying covariates and a 17-node DAG encoding the full mediation structure.

1. **Baseline demographics.** Age $\sim \mathcal{N}(55, 100)$; BMI from age with Gaussian noise; total cholesterol (TC) from a log-normal with age/BMI dependence; high-density lipoprotein (HDL) inversely related to age and BMI; smoking status $\sim \mathrm{Bernoulli}(0.6)$; diabetes probability proportional to BMI; race and sex drawn from fixed population proportions.
2. **Time-varying dynamics.** At each year t, age increments by 1; BMI, TC, and HDL evolve with Gaussian noise; hypertension (htn) transitions to 1 irreversibly with a probability driven by age, TC, and smoking; diabetes is absorbing once acquired. Systolic blood pressure (SBP) at time t is:
$$\mathrm{SBP}_{\text{final}}^{(t)} = \mathrm{SBP}^{(t)}\big(1 - [0.20\,A^{(t-1)} - 0.05\,A^{(t-2)} - 0.02\,A^{(t-3)}]\big),$$
clipped to $[80, 200]$.
3. **Treatment assignment (absorbing).** At t=0, patients with hypertension are treated with probability 0.70; non-hypertensive patients with probability 0.30. Treatment is absorbing: once assigned it persists. At subsequent time steps, untreated hypertensive patients initiate treatment with probability 0.01 per year.

4. **Mediated treatment effect.** The antihypertensive reduces SBP via lagged mediation:
$$\mathrm{SBP}_{\mathrm{final}}^{(t)} = \mathrm{SBP}^{(t)}\big(1 - [0.20\,T^{(t-1)} - 0.05\,T^{(t-2)} - 0.02\,T^{(t-3)}]\big),$$
with lags 1-3. This is the only pathway through which treatment affects the outcome; the complete causal chain is $A \to \mathrm{SBP}_{\mathrm{final}} \to Y$.
5. **CVD outcome.** Binary CVD event at each timestep from a logistic model with the following risk-factor coefficients:
$$\begin{aligned}\log\frac{p}{1-p} = {} & -10 + 0.005\,\mathrm{age} + 0.15\,\mathrm{sex} + 0.03\,\mathrm{BMI} + 0.015\,\mathrm{SBP}_{\mathrm{final}} \\ & - 0.01\,\mathrm{HDL} + 0.01\,\mathrm{TC} + 0.25\,\mathrm{smoker} + 0.30\,\mathrm{diabetes} \\ & + 0.20\,\mathrm{fam_hx} + 0.10\,\mathbb{1}[\mathrm{race} = \mathrm{Black}] - 0.05\,\mathbb{1}[\mathrm{race} = \mathrm{Asian}].\end{aligned}$$
The outcome is absorbing: rows after the first event are excluded. Both always-treated and never-treated probability trajectories are stored as counterfactual targets.

17-node DAG. The causal graph encodes the full mediation structure: $A \to \mathrm{SBP}_{\mathrm{final}} \to Y$, plus direct paths from all 15 covariates to Y and temporal self-edges.

Purpose. Strong confounding (hypertensive patients preferentially treated), a fully mediated treatment pathway (SBP lags 1-3), 15 correlated time-varying covariates, and an absorbing binary outcome combine to make this the hardest benchmark. Correctly recovering the true $\mathrm{HR} \approx 0.82$ requires disentangling the mediation chain from the confounded observational distribution (precisely the setting where the expert DAG prior is expected to matter most).

A.5 Summary and main purpose

The four benchmarks vary deliberately in complexity.

Simple 3-Node is a positive control (minimal confounding, three variables, known closed-form ITE) and is used to verify that all the models perform well in a simple, unrealistic case.

Cox Survival Toy introduces a survival endpoint with bimodal covariate heterogeneity and selection-on-risk confounding, testing whether models correctly recover per-arm cumulative incidence functions under censoring.

LDL Toy and *CVD Risk Toy* are the two hardest benchmarks and carry the most diagnostic weight in our evaluation. LDL Toy combines log-saturation pharmacokinetics, AR(1) temporal dynamics, and selection-on-gain confounding, making it the primary stress-test for subgroup calibration: a model that merely minimizes factual mean squared error (MSE) will exhibit structural bias (bias/MAE ≈ 1) and variance collapse ($\mathrm{VR}_{Q4} < 1$) under this design.

On the other hand, CVD Risk Toy is the most demanding benchmark overall: 15 time-varying covariates, an absorbing binary outcome, a purely mediated treatment effect through SBP (lags 1–3), and a 17-node expert DAG create a setting where correctly recovering the interventional distribution $p(Y \mid \mathrm{do}(A))$requires explicit encoding of the causal graph. It is the only benchmark where direction inversion ($\widehat{\mathrm{HR}} > 1$ for a protective treatment) is observed for models lacking structural constraints, making it the definitive test of causal identifiability.

Together, these benchmarks aim to cover the three main failure modes of longitudinal causal models (confounding, variance miscalibration, and mediation) while providing exact ground-truth potential outcomes for all evaluation metrics.

A.6 Real-World EHR Database Description and Use

The Family Medicine Research using Electronic Medical Records (FIRE) database (managed by the Institute of Primary Care at the University of Zurich, at the University Hospital Zurich, Switzerland) is the most comprehensive Swiss primary-care database of longitudinal (time-stamped) EHR from general practitioners across Switzerland [Balaj et al., 2025] The FIRE database contains individual-level information on demographics, consultations, medication prescriptions, laboratory values, vital signs, and diagnoses [Balaj et al., 2025, Chmiel, 2011]. FIRE is used in two ways throughout this work: as a clinical backbone for generating the semi-synthetic evaluation dataset (Section A.6.1), and to create the study population for the real-world application (Section A.6.2).

A.6.1 Semi-Synthetic MNAR Dataset

The semi-synthetic evaluation dataset is built on a patient-month panel structure from FIRE, keeping all observed real-world information on demographics, comorbidities, comedication, and calendar time.

Synthetic covariates: Ten continuous longitudinal variables are generated recursively on top of this backbone, indexed by patient and by month. The variables share latent factors and cross-variable dependencies, but in order to be realistic, the variables are designed to be non-smooth across time: they rely on near-memoryless shared latent drivers, external shocks, and noise.

The generation proceeds as follows:

1. A patient-month panel is constructed from selected demographic and clinical covariates.
2. Within-patient time indices are created.
3. Three patient-level latent traits are drawn once per patient.
4. Two weakly autocorrelated shared state variables drive autocorrelation across series.
5. External shocks introduce abrupt level changes.
6. Ten variable-specific noise terms complete the stochastic structure.
7. For each patient, a recursive simulation initializes each variable from covariates, latent traits, and noise, then iterates forward introducing autocorrelation, cross-variable dependence, treatment indicators, latent states, and shocks.

Synthetic outcome: The continuous outcome is a nonlinear function of the synthetic covariates, clinical covariates, treatment assignment, time gap, and lagged terms, with added individual noise and occasional outliers.

Ground-truth counterfactuals and benchmark average treatment effect (ATE): Potential outcomes under both treatment arms are computed directly from the structural outcome function, providing exact individual-level counterfactuals and a known benchmark ATE.

MNAR missingness mechanism: Missingness is introduced into the ten synthetic covariates at 30%, 50%, and 80% following the MNAR mechanism specified in Equation 1. This mirrors the clinical pattern where measurements are triggered by patient condition rather than scheduled, meaning the probability of observing a value depends on the underlying patient state. This controlled setup allows us to directly evaluate imputation quality and the subsequent robustness of causal estimators against a known ground truth.

Directed acyclic graph (DAG) and variable distributions: Figure 3 illustrates the DAG of the generative process. We also show the resulting distribution of the synthetic variables in Figure 4. Lagged covariates are included for realism: `lag(Synthetic Variable)` at time t is derived from the same variable at time $t-1$, so the causal direction is past value $\rightarrow$ lagged feature at the current step. Although the naming may seem counterintuitive, the causal direction is straightforward: the past value of a variable drives its own lagged feature at the current time step. Figure 4 shows the resulting variable distributions.

A.6.2 Target trial emulation using Real-World EHR data

Study design and patient cohort construction: Using the FIRE database, we design an active-comparator, new-user cohort study (emulating a target trial; see Table 3) to estimate the effect of initiating a GLP-1 receptor agonist (GLP-1RA) or an SGLT-2 inhibitor (SGLT-2i) on body weight change from treatment initiation over 1-year. Cohort entry (time-zero) is defined as the date of the first-ever prescription for a GLP-1RA or an SGLT-2i as first second-line glucose-lowering therapy between 01-Jan-2015 and 07-Dec-2024. To be eligible, individuals are aged 18 years or older at cohort entry and have a diagnosis of type 2 diabetes (T2D) before or at cohort entry (see detailed eligibility criteria in Table 3).

Treatment strategies and follow-up: We use a per-protocol analysis in which eligible individuals are followed from cohort entry until the earliest of: deviation from the assigned treatment strategy

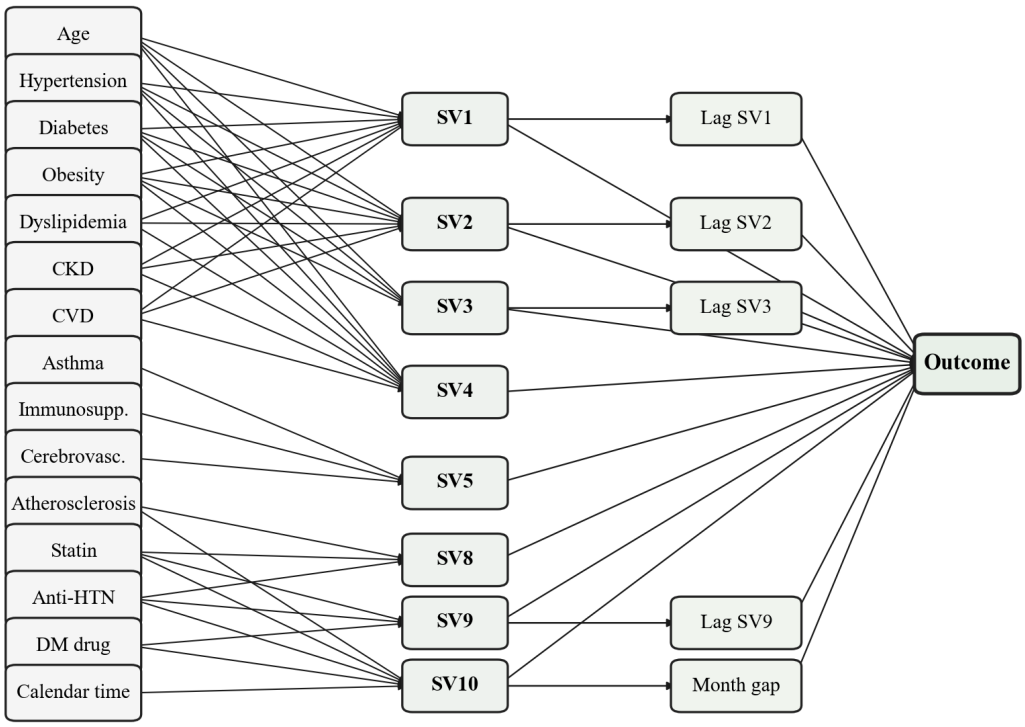


Figure 3: Directed acyclic graph of the generation of the synthetic variables and the outcome.

(defined as a gap between consecutive prescriptions exceeding 365 days, with a 30-day grace period); initiation of the comparator drug class; death; or end of the pre-specified follow-up period (390 days after cohort entry).

Outcome: The outcome is the absolute change in body weight (kg) from baseline over 1-year. The outcome is extracted in 30-day intervals throughout follow-up.

Covariates: Time-fixed baseline covariates measured at cohort entry include demographic characteristics (age, sex assigned at birth). Time-varying covariates updated from baseline every 30-day interval throughout follow-up include current comedications (antihypertensives, statins), comorbidities (cardiovascular disease, chronic kidney disease, hypertension, dyslipidemia), and laboratory measurements (Hemoglobin A_{1c}, eGFR, LDL cholesterol, systolic blood pressure).

Confounders and DAG: Using the covariates defined above, causal assumptions are encoded in an expert-specified DAG developed with pharmacoepidemiology experts (Figure 5), following standard practice in target trial emulation [Hernán et al., 2022]. The DAG identifies the minimal sufficient adjustment set required for conditional exchangeability between treatment arms.

Real-world application of the proposed ML pipeline: The final cohort comprised $6{,}114$ adults with T2D initiating a GLP-1RA ($n = 3{,}722$) or an SGLT-2i ($n = 2{,}392$) as their first second-line glucose-lowering therapy. Following our proposed pipeline (Section 3), missing covariates or outcome measurements are imputed using the validated LLM-based strategy across follow-up. CausalFlow-T was then fitted on the completed longitudinal panel.

B Baseline Architectures for Causal Inference Estimation

B.1 Baseline Selection Criteria

We characterize candidate baselines by three criteria:

Table 3: Specification and emulation of the target trial protocol aiming to estimate the per-protocol effect of GLP-1 receptor agonist (GLP-1RA) versus SGLT-2 inhibitor (SGLT-2i) initiation on body weight at one year in adults with type 2 diabetes (T2D).

Protocol component	Target trial specification	Target trial emulation
Eligibility criteria	Inclusion criteria: Adults ($\geq$18 yrs) initiating a GLP-1RA or an SGLT-2i; T2D diagnosis. Exclusion criteria: Prescription for the comparator drug class before treatment initiation; type 1 diabetes diagnosis; last recorded eGFR $\leq$30ml/min/1.73m2.	Inclusion criteria: Adults ($\geq$18 yrs) initiating a GLP-1RA or an SGLT-2i between 01-Jan-2015 and 07-Dec-2024 (cohort entry = first prescription date); T2D diagnosis before or at cohort entry; at least one weight measurement recorded within the 365 days before cohort entry. Exclusion criteria: Prescription for the comparator drug class before or at cohort entry; type 1 diabetes diagnosis before or at cohort entry; eGFR $\leq$30 before or at cohort entry. Missingness in variables needed to assess eligibility is imputed using the LLM-based strategy prior to eligibility assessment.
Treatment strategies	(1) Initiate and sustain a GLP-1RA receptor agonist; (2) initiate and sustain an SGLT-2i inhibitor; as add-on to first-line glucose-lowering therapy over 12 months.	Same strategies; patients classified by observed prescription at cohort entry.
Treatment assignment	Random assignment at baseline (open-label); participants not blinded to assigned strategy.	Non-random; conditional exchangeability emulated by adjusting for the minimal sufficient adjustment set identified from the expert-specified DAG (Figure 5).
Outcome	Change in body weight (kg) from baseline to 1-year (12 months), measured in 30-day intervals throughout follow-up.	Same. Missing outcome measurements are imputed using the LLM-based strategy.
Follow-up	From baseline until the earliest of: per-protocol deviation from assigned strategy; initiation of comparator drug class; death; or 390 days after cohort entry.	Same. Per-protocol deviation defined as a gap between consecutive prescriptions exceeding 365 days (and adding a 30-day grace period) or initiation of the comparator drug class.
Causal contrast	Per-protocol effect: treatment effect for all randomized individuals if all adhered to treatment and did not initiate the agent of the comparator drug class.	Observational analog of the per-protocol effect, estimated under the assumption of conditional exchangeability given the adjustment set identified from the DAG(Figure 5).
Statistical analysis	ATE estimation via marginal structural model with inverse probability of censoring weighting to account for informative censoring due to non-adherence.	Following our proposed pipeline (Section 3), missing covariates or outcome measurements are imputed using the validated LLM-based strategy across follow-up. CausalFlow-T was then fitted on the completed longitudinal panel.

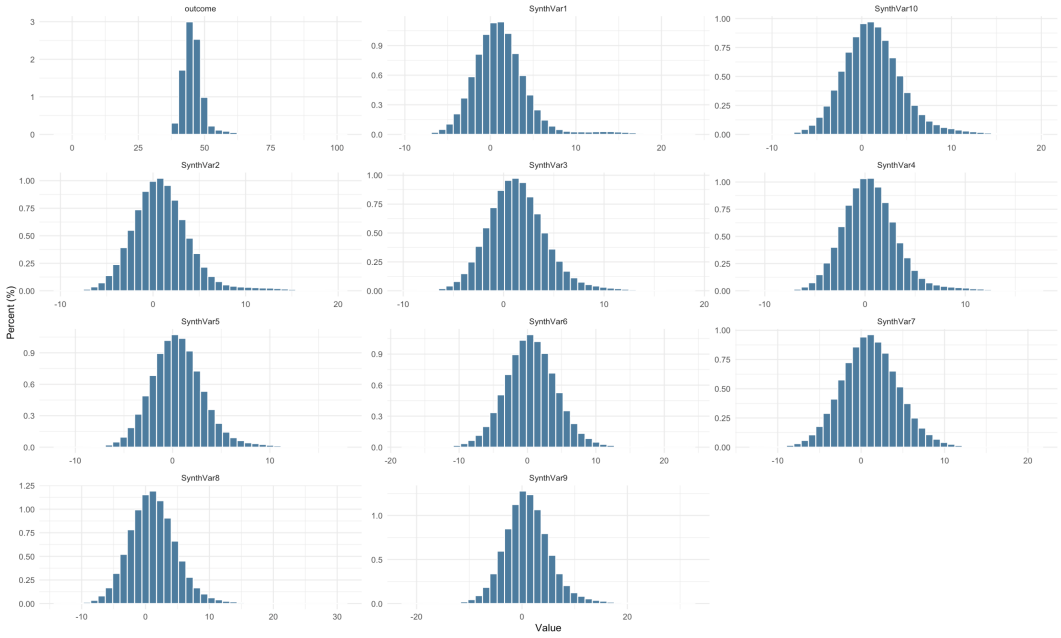


Figure 4: Distribution of the synthetic variables and outcome.

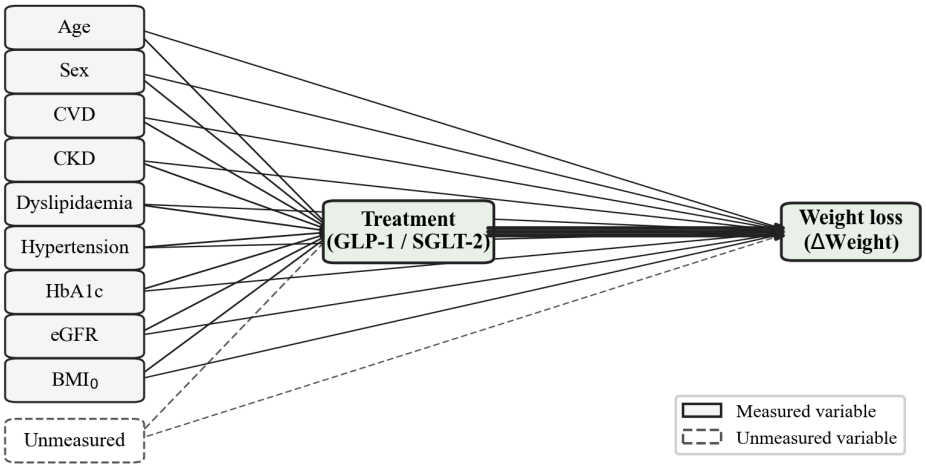


Figure 5: Expert-specified directed acyclic graph for the active-comparator, new-user cohort study assessing the effect of GLP-1 receptor agonist versus SGLT-2 inhibitor initiation on weight loss. The DAG is a simplified version with no differentiation between baseline and time-varying confounders.

1. **Generative mechanism.** The model must define a joint distribution $p(Y^{(0)}, Y^{(1)} \mid X)$ rather than only conditional means $\mathbb{E}[Y \mid X, A]$, as the latter is insufficient for individual-level counterfactual inference via the AAP procedure [Pearl, 2009].

2. **Invertible abduction.** The model must map each patient's observations to exogenous noise z and back exactly (a bijection), so that both potential outcomes of a patient are predicted from the same abducted noise; without this, $\hat{y}^{(0)}$ and $\hat{y}^{(1)}$ are predicted independently rather than coupled for the same patient.

3. **Distributional evaluability.** The model must produce outputs compatible with our five reliability criteria (subgroup calibration, tail variance ratio, arm reconstruction error, HR

recovery, and training stability) all of which require access to the full potential outcome distributions $p(\hat{y}^{(a)})$ rather than point predictions.

Table 4 summarizes how each candidate model maps onto these criteria and the resulting inclusion decision.

Table 4: Baseline selection against the three criteria. Only CausalFlow-T satisfies all three. CVAE and GNN-CVAE (approximate, non-invertible abduction) and TARNet (discriminative) are included as reference baselines along the design axes. The rest of the discriminative sequence models are excluded.

Model	Generative mechanism	Invertible abduction	Distributional evaluability	Included
CausalFlow-T (ours)	✓	✓	✓	✓
CVAE [Louizos et al., 2017]	✓	✗	✓	✓
GNN-CVAE	✓	✗	✓	✓
TARNet [Shalit et al., 2017]	✗	✗	✗	✓
Causal Transformer [Melnychuk et al., 2022]	✗	✗	✗	✗
R-MSN [Lim et al., 2018]	✗	✗	✗	✗
CRN [Bica et al., 2020]	✗	✗	✗	✗

Discriminative sequence models (including the Causal Transformer [Melnychuk et al., 2022], R-MSN [Lim et al., 2018], and CRN [Bica et al., 2020]) give conditional-mean factual forecasts under a treatment history rather than the paired per-unit potential outcomes the criteria need, so our reliability criteria cannot be computed on them; they are excluded for this reason, not for underperforming. TARNet, although discriminative and failing all three criteria, qualifies: its shared encoder reads only covariates, so its two heads give $\hat{y}_i^{(0)}(t)$ and $\hat{y}_i^{(1)}(t)$ for every patient and time step, and it is kept as a discriminative reference.

We select CVAE (a conditional VAE in the spirit of CEVAE [Louizos et al., 2017], not a reimplementation of it), GNN-CVAE, and TARNet [Shalit et al., 2017] to span the key design axes underlying the proposed criteria. CVAE provides a variational inference baseline without causal structure; GNN-CVAE augments this with DAG-structured encoding under the same variational objective, isolating the contribution of causal structure within approximate inference. TARNet, in contrast, serves as a purely discriminative reference point, failing all three criteria. Together, these baselines expose the two failure modes: confounding separation and structural propagation.

B.2 Shared Components

All models share: an outcome-reader LSTM (h_t^Y, used only in the encoder to prevent outcome leakage), a treatment-encoder LSTM (h_t^A, replaced from the intervention time t^* onwards at counterfactual generation time), and a covariate aggregator (linear projection to h_t^{cov}).

B.3 Conditional CVAE

Encoder. $q_\phi(z \mid x_{1:T}, y_{1:T}) = \mathcal{N}(\mu_\phi, \text{diag}(\sigma_\phi^2))$, where $[\mu_\phi, \log \sigma_\phi^2] = \text{MLP}_\phi([\text{flatten}(h_{1:T}^{\text{cov}}), \bar{h}^Y])$. Latent dimension $L = 64$.

Decoder. $\hat{y}_{1:T} = \text{MLP}_\theta([z, \text{flatten}(h_{1:T}^A)])$.

Objective. $\mathcal{L}_{\text{CVAE}} = -\mathbb{E}_{q_\phi}[\log p_\theta(y_{1:T} \mid z, h_{1:T}^A)] + \beta_t \, \text{KL}[q_\phi \| \mathcal{N}(0, I)]$, with linear β_t warm-up (0→1 over first 20 epochs).

Key limitation. The variational posterior is an approximation; the abducted noise is not the exact exogenous noise of the patient, and this approximation error can propagate into counterfactual predictions, depending on the variational family and training.

B.4 GNN-CVAE

Replaces the flat LSTM covariate encoder with a DAG-structured recurrent encoder performing message passing in topological order:

$$h_t^{(k)} = \mathrm{LayerNorm}\Big(\mathrm{GRU}\big(h_t^{(k),\mathrm{in}} + m_t^{(k)},\, h_{t-1}^{(k)}\big)\Big), \tag{6}$$

where $m_t^{(k)} = |\mathrm{pa}(k)|^{-1} \sum_{l \in \mathrm{pa}(k)} \mathrm{MLP}_{l\to k}\big([h_t^{(l)}, h_{t-1}^{(k)}]\big)$. The causal structure is enforced in the encoder only. However, the ELBO approximation gap in the abduction step remains, which may limit the benefit of the graph-structured encoder.

C Baseline model for MNAR imputation: Conditional DAG-aware Flow Matching

We implement a new Conditional DAG-aware Flow Matching (CFM) imputer as a baseline for MNAR imputation of the synthetic biomarker variables (SynthVars). CFM learns the conditional transport $p_{\mathrm{noise}} \to p(x^s_{\mathrm{miss}} \mid x^s_{\mathrm{obs}}, x^{ns}, A, Y, \delta)$ via flow matching, with a loss evaluated only at missing positions.

Architecture. A Transformer encoder (L=2, H=4) processes the always-observed non-SynthVar sequence $\mathbf{x}^{\mathrm{ns}}_{1:T}$ with a key-padding mask on unobserved timesteps, producing a leakage-free context vector e_t that interpolates from truly-observed neighbors on both sides of each gap.

A three-layer MLP vector field u_θ then maps the concatenation

$$c_t = \big[x^s_{t_f} \,\|\, x^{ns}_t \,\|\, A_t \,\|\, Y_t \,\|\, m_t \,\|\, x^s_{t-1} \,\|\, m_{t-1} \,\|\, e_t \,\|\, \delta_t \,\|\, t_f\big]$$

(δ_t: time since the previous visit) to a velocity $\hat{v} \in \mathbb{R}^{|\mathcal{S}|}$. Conditioning on Y_t and A_t ensures that downstream DAG evidence propagates into every imputed value. The architecture is illustrated in Figure 6.

Training. On fully-observed rows a random binary mask $\mathbf{m}^{\mathrm{miss}}$ is applied and a linear interpolant $\mathbf{x}_{t_f} = (1 - t_f)\mathbf{x}_0 + t_f \mathbf{x}_1$, $\mathbf{x}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is constructed, with target velocity $v^* = \mathbf{x}_1 - \mathbf{x}_0$. The objective is a masked mean-squared error at missing positions only, requiring no Jacobian:

$$\mathcal{L}(\theta) = \frac{1}{\|\mathbf{m}^{\mathrm{miss}}\|_1} \sum_i m_i^{\mathrm{miss}} \big(\hat{v}_i - v_i^*\big)^2.$$

Inference. The encoder runs once over the full dataset to produce $e_{1:T}$. Missing values are then recovered by Euler integration of the ordinary differential equation (ODE) $\mathrm{d}\mathbf{x}/\mathrm{d}t_f = u_\theta(\mathbf{x}_{t_f}, \mathbf{c}_t, t_f)$ from Gaussian noise to the imputed target, with observed dimensions held fixed throughout integration.

D Evaluation Protocol

In this appendix we write $y_i^{(a)}(t)$ for the potential outcome $y_{i,t}(a)$ of Section 3.1, $\hat{y}_i^{(a)}(t)$ for its estimate, and $x^*_{i,t,j}$ for the ground-truth value $x_{i,t,j}$ of a cell that is missing in $\tilde{\mathcal{D}}$.

D.1 CausalFlow-T Evaluation Protocol

(i) Subgroup calibration.

$$\mathrm{MAE}_q = \frac{1}{T} \sum_t \big|\widehat{\mathrm{ATE}}_q(t) - \mathrm{ATE}_q(t)\big|, \quad \mathrm{Bias}_q = \frac{1}{T} \sum_t \big(\widehat{\mathrm{ATE}}_q(t) - \mathrm{ATE}_q(t)\big),$$

where $\widehat{\mathrm{ATE}}_q(t) = \mathbb{E}_{i\in q}[\hat{y}_i^{(1)}(t)] - \mathbb{E}_{i\in q}[\hat{y}_i^{(0)}(t)]$ and analogously for the ground truth. Quartiles are defined by the per-patient mean true ITE, $\bar{\tau}_i = \frac{1}{T}\sum_t [y_i^{(1)}(t) - y_i^{(0)}(t)]$, ensuring identical patient subgroups across all models.

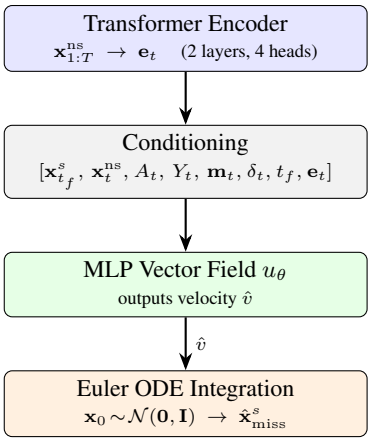


Figure 6: **CFM imputer.** The Transformer Encoder produces context $\mathbf{e}_t$ from always-observed $\mathbf{x}^{\text{ns}}_{1:T}$. The MLP u_θ is conditioned on $\mathbf{e}_t$ and DAG evidence (A_t, Y_t), trained with masked MSE on $v^*{=}\mathbf{x}_1{-}\mathbf{x}_0$. At inference, missing values are recovered by ODE integration from Gaussian noise.

(ii) Arm reconstruction error.

$$\text{Err}_a = \frac{1}{T}\sum_t \left|\mathbb{E}[\hat{y}^{(a)}(t)] - \mathbb{E}[y^{(a)}(t)]\right|, \quad a \in \{0,1\},$$

measured in original outcome units, matching the scale of the mean trajectory plots.

(iii) Variance calibration. $\text{VR}_{\text{Q4}} = \text{Var}(\hat{y}^{(a)}_{\text{Q4}})/\text{Var}(y^{(a)}_{\text{Q4}})$, averaged over both arms and all time steps within Q4. Values below 1 indicate distributional collapse (the model predicts near-constant outputs for high-effect patients); values above 1 indicate overdispersion.

(iv) PEHE and relation to established metrics. The precision in estimating heterogeneous effects [Hill, 2011, Shalit et al., 2017] is

$$\text{PEHE} = \Big(\frac{1}{NT}\sum_{i,t}\big(\hat{\tau}_{i,t} - \tau_{i,t}\big)^2\Big)^{1/2},$$

evaluated against ground truth. PEHE aggregates individual errors and does not distinguish random from systematic error. Our subgroup criteria decompose this aggregate by effect stratum, following the CATE-calibration literature: Athey and Wager [2019] assess calibration by grouping units by predicted effect, and Xu and Yadlowsky [2022] formalize a calibration error for heterogeneous effects. Because ground truth is available in our benchmarks, we stratify by the true effect, which isolates systematic bias in low-effect patients ($|\text{Bias}_{Q1}|/\text{MAE}_{Q1}$) and variance collapse in high-effect patients (VR_{Q4}). The HR criterion adds a direction check that PEHE lacks: a model can have moderate PEHE while predicting harm where there is benefit.

D.2 Imputer Evaluation Protocol

Let $\mathcal{M} = \{(i,t,j) : m_{i,t,j} = 1\}$ denote the set of MNAR-missing biomarker cells, $\hat{x}_{i,t,j}$ the imputed value, and $x^*_{i,t,j}$ the ground truth (available by construction on the FIRE oracle, Section A.6). Section 4.3 reports two layers: *biomarker quality* (i)–(v) on the reconstruction itself, and *downstream causal quality* (vi)–(ix) on the fixed CausalFlow-T estimator (Section 4.2). Throughout, $d{=}10$ is the number of synthetic biomarkers, N the cohort size, and T the number of timesteps.

D.2.1 Biomarker quality

(i) Pointwise MAE.

$$\text{MAE} = \frac{1}{|\mathcal{M}|}\sum_{(i,t,j)\in\mathcal{M}}\left|\hat{x}_{i,t,j} - x^*_{i,t,j}\right|,$$

in original biomarker units, restricted to masked positions.

(ii) Pointwise RMSE.

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\mathcal{M}|} \sum_{(i,t,j)\in\mathcal{M}} \big(\hat{x}_{i,t,j} - x^*_{i,t,j}\big)^2}.$$

Like MAE, RMSE is computed in original biomarker units and restricted to masked positions.

(iii) Lag-1 autocorrelation error. For biomarker j and patient i, let $r_{i,j}(z) = \mathrm{Corr}_t\big(z_{i,t,j},\, z_{i,t+1,j}\big)$ denote the Pearson correlation between consecutive timesteps of a series z. The mean lag-1 autocorrelations on the imputed and ground-truth datasets are,

$$\mathrm{AC} = \frac{1}{Nd}\sum_{i,j} r_{i,j}(\hat{x}), \qquad \mathrm{AC}^* = \frac{1}{Nd}\sum_{i,j} r_{i,j}(x^*),$$

and we report $|\mathrm{AC} - \mathrm{AC}^*|$, with AC^* =0.375. LOCF inflates AC toward 1; shrinkage-to-mean imputers deflate it toward 0.

(iv) Consecutive-step error. On observed-to-missing transitions $\mathcal{S} = \{(i,t,j) : m_{i,t,j} = 0,\, m_{i,t+1,j} = 1\}$,

$$\mathrm{CSE} = \frac{1}{|\mathcal{S}|} \sum_{(i,t,j)\in\mathcal{S}} \big|\hat{x}_{i,t+1,j} - x^*_{i,t+1,j}\big|,$$

isolating one-step extrapolation from an observed clinical anchor.

(v) Biomarker–outcome correlation inflation. For each biomarker j, let $\kappa_j(z) = \mathrm{Corr}\big(z_{\cdot,\cdot,j},\, y_{\cdot,\cdot}\big)$ be the Pearson correlation between the (imputed or true) biomarker series and the outcome, pooled over all (i,t) pairs. We define

$$\mathrm{inf}_Y = \frac{1}{d}\sum_{j=1}^{d} \big(\kappa_j(\hat{x}) - \kappa_j(x^*)\big),$$

with $\mathrm{inf}^*_Y = 0$. Positive values inflate the biomarker–outcome association, while negative values indicate deflation, in which a real signal is washed out.

D.2.2 Downstream causal quality

CausalFlow-T is held fixed (Section 4.2) so reported differences reflect imputation alone.

(vi) Subgroup calibration MAE. Following Appendix D (i), for each true-ITE quartile $q \in \{Q_1, Q_2, Q_3, Q_4\}$ we compute

$$\mathrm{MAE}_q = \frac{1}{T}\sum_{t=1}^{T} \big|\widehat{\mathrm{ATE}}_q(t) - \mathrm{ATE}_q(t)\big|.$$

We focus on the two tail quartiles, which expose complementary failure modes invisible in aggregate metrics: MAE_{Q_1} on low-effect patients, where confounding separation is hardest and systematic bias is most exposed; and MAE_{Q_4} on high-effect patients, where variance collapse manifests as flattening of predicted treatment effects. Table 2 ranks methods by MAE_{Q_1}; Appendix Table 12 reports both tail quartiles.

(vii) Mean per-arm reconstruction error $\bar{\mathrm{Err}}_a$. As in Appendix D (ii), averaged across the two arms, $\bar{\mathrm{Err}}_a = \frac{1}{2}\big(\mathrm{Err}_{a=0} + \mathrm{Err}_{a=1}\big)$, summarizing population-level trajectory accuracy under both treatment assignments.

(viii) Variance calibration. $\mathrm{VR}_{\mathrm{Q4}}$ as defined in Appendix D (iii).

(ix) Absolute ATE residual. The ATE is the population-level expected difference between the potential outcome under treatment and under control, averaged across patients and timesteps:

$$\widehat{\mathrm{ATE}} = \frac{1}{NT}\sum_{i,t}\big(\hat{y}_i^{(1)}(t) - \hat{y}_i^{(0)}(t)\big), \qquad \mathrm{ATE}^* = \frac{1}{NT}\sum_{i,t}\big(y_i^{(1)}(t) - y_i^{(0)}(t)\big),$$

with $\mathrm{ATE}^* = -3.484$. We report $|\widehat{\mathrm{ATE}} - \mathrm{ATE}^*|$.

E Training and Hyperparameter Details

Tables 5 and 6 define the hyperparameter details for all the experimental parts.

Table 5: Hyperparameters across datasets

Component	Hyperparameter	Synthetic	FIRE semi	FIRE real
Shared training	Optimizer	Adam	Adam	Adam
	Learning rate	10^{-3}	10^{-3}	10^{-4}
	Batch size	512	256	64
	Data split (train(train/val)/test)	80(80/20)/20	80(80/20)/20	80(80/20)/20
LSTM encoder	Hidden dimension H	128	256	256
	Dropout	-	-	0.3
CausalMAF	Coupling layers	1	1	1
	Hidden dimension	128	128	128
TARNet	Hidden dimension	128	128	Not Used
	MLP depth	2	1	Not Used
GNN-CVAE	Encoder layers	2	2	Not Used
	Latent dimension	64	64	Not Used
CVAE	Architecture	No graph encoder	Not graph encoder	Not Used
Normalization	Covariates & treatments	Joint standardization	Joint standardization	Per-timestep
	Outcomes	Per-timestep	Per-timestep	Per-timestep

Table 6: Main hyperparameters of the LLM-driven evolutionary imputation search.

Hyperparameter	Value
LLM model	`gpt-5.4, qwen3.5-plus, gpt-oss-120b`
Iteration budget K	20
Proxy holdout fraction ρ	0.10
Holdout base seed	fixed across candidate evaluations
Composite-score weight λ_Y	2.0
Composite-score weight λ_T	0.5
Selection rule	minimize $s = \mathrm{RMSE} + \lambda_Y \Delta_Y + \lambda_T \Delta_T$
History window W	3
Per-dataset timeout $t_{\max}$	180 s
Memory cap M_{run}	6144 MB

F FIRE Semi-Synthetic Oracle Results

Table 7 reports full results on the FIRE semi-synthetic oracle (no missingness), used as a controlled bridge between synthetic benchmarks and the imputation experiments (Section 4.3). True ATE ≈ -3.48; multiple seeds throughout.

Table 7: FIRE semi-synthetic oracle: systematic-error ratio $|\mathrm{Bias}_{\mathrm{Q1}}|/\mathrm{MAE}_{\mathrm{Q1}}$ ($\downarrow$; <0.5 = predominantly random errors), arm reconstruction, variance ratio, and ATE recovery (multiple seeds). $\mathrm{VR}_{\mathrm{Q4}} \to 1$. True ATE ≈ -3.48. **Bold**: best per metric.

Model	$\lvert\mathrm{Bias}_{\mathrm{Q1}}\rvert/\mathrm{MAE}_{\mathrm{Q1}}\downarrow$	$\mathrm{Err}_{a=1}$	$\mathrm{Err}_{a=0}$	$\mathrm{VR}_{\mathrm{Q4}}$	ATE
CausalFlow-T **(ours)**	**0.316**	0.088±0.032	**0.063±0.022**	**1.006±0.009**	−3.568±0.112
NF (no DAG)	0.883	0.161±0.073	0.084±0.044	1.016±0.010	−3.382±0.251
TARNet	0.708	**0.081±0.012**	0.078±0.003	0.608±0.003	−3.415±0.015
GNN-CVAE (DAG)	0.351	0.216±0.048	0.202±0.038	0.616±0.021	−3.395±0.113
CVAE	0.938	0.134±0.045	0.127±0.047	0.482±0.055	**−3.451±0.075**

CausalFlow-T (0.316) and GNN-CVAE (0.351) are the only two models passing the <0.5 reliability threshold, indicating that low-responder errors are predominantly random rather than systematic; NF (no DAG) (0.883), TARNet (0.708), and CVAE (0.938) all fail, confirming that explicit causal structure is necessary for confounding separation under realistic clinical covariate distributions. CausalFlow-T is further distinguished as the only model with near-perfect variance preservation ($\mathrm{VR} = 1.006 \pm 0.009$), while GNN-CVAE collapses ($\mathrm{VR} = 0.616$) and CVAE collapses more severely ($\mathrm{VR} = 0.482$), confirming that passing the bias threshold alone is insufficient for distributional reliability. TARNet achieves the best arm-1 reconstruction (0.081) and a good ATE point estimate (-3.415) but fails the ratio threshold (0.708) and collapses variance ($\mathrm{VR} = 0.608$), reinforcing that factual accuracy does not imply counterfactual reliability. Removing the DAG constraint (NF no DAG) degrades the bias ratio to 0.883 and produces the largest ATE deviation (-3.382, $\Delta = 0.102$) with substantial cross-seed instability (± 0.251), underscoring that structural supervision is necessary for stable estimation under clinical covariate distributions.

G Full Numerical Results

Tables 8 and 9 report complete per-model, per-benchmark results for the four synthetic benchmarks; they are the primary evidence base for Table 1 in the main body. FIRE semi-synthetic oracle results are in Table 7 (Appendix F).

- **Simple 3-Node.** On *Simple 3-Node* (linear, no confounding), all models show $|\mathrm{Bias}_q| \approx \mathrm{MAE}_q$ with near-zero standard deviations throughout. This is expected: in a confounding-free linear setting, residual errors are purely systematic rather than variance-driven. Seed consistency confirms training stability across all architectures, so differences on harder benchmarks reflect genuine architectural sensitivity. GNN-CVAE achieves the lowest Q1 MAE (0.425) in this regime, illustrating that explicit graph encoding can reduce absolute error when confounding is absent (but this advantage disappears under the more complex generating processes below).
- **LDL Toy.** TARNet achieves the lowest absolute MAE across all the quartiles (0.407–1.142), but a $\mathrm{VR}_{\mathrm{Q4}} = 0.611 \pm 0.003$. CausalFlow-T achieves $\mathrm{VR}_{\mathrm{Q4}} = 1.049 \pm 0.033$ (closest to 1 across all models), confirming that tail variance is faithfully preserved rather than collapsed or inflated, while maintaining sign-consistent positive bias ($+0.547$–$+0.963$ across quartiles) indicating mild but coherent overestimation rather than directionally inconsistent errors. NF (no DAG) systematically underestimates across all quartiles (-0.402 to -1.004), confirming that causal factorization is necessary for directionally correct inference. GNN-CVAE and CVAE show variance collapse ($\mathrm{VR}_{\mathrm{Q4}} \leq 0.85$), revealing distributional narrowing in the high-effect tail.
- **CVD Risk.** Ground-truth $\mathrm{HR} = 0.831$ (protective treatment). CausalFlow-T recovers $\mathrm{HR} = 0.786 \pm 0.051$ ($\Delta = 0.045$), the only model within clinically meaningful range with arm errors below all competitors. GNN-CVAE and CVAE estimate HR $\approx 1.005 \pm 0.030$ across all seeds (predicting a null effect despite a true 17% hazard reduction), consistent with posterior collapse in our implementations. TARNet predicts $\mathrm{HR} = 1.133 \pm 0.148$, inverting the effect direction on average. NF (no DAG) yields $\mathrm{HR} = 0.834 \pm 0.324$: the large standard deviation (related to the explosion in the variance inf Q4) reveals that without the DAG constraint the flow is fundamentally unstable under strong mediation (sometimes approximately correct, sometimes catastrophically wrong), making it unreliable for clinical deployment regardless of average performance.
- **Cox Survival.** CVAE achieves lower Q1–Q4 MAE on Cox (0.023 ± 0.001–0.020 ± 0.006) because variational models handle binary outcomes without continuous relaxation. However, CausalFlow-T achieves the best arm-1 reconstruction error (0.014 ± 0.002) and a close HR recovery (0.866 ± 0.011 vs. true 0.887), while CVAE and GNN-CVAE show arm errors 4–6$\times$ higher. TARNet achieves a good arm-0 error (0.015 ± 0.001) but overshoots the HR (0.966 ± 0.005), indicating factual accuracy without causal calibration. The Cox MAE advantage of CVAE does not generalize to more complex DAGs, as the CVD Risk results confirm.

CVAE vs. GNN-CVAE. Across benchmarks, CVAE consistently outperforms GNN-CVAE despite the latter incorporating explicit graph structure in the encoder. This contrasts with the normalizing

Table 8: Subgroup calibration and tail variance ratio across all benchmarks. GNN-CVAE and CVAE on LDL and Cox use the full simulation variant. **Bold**: best value(s) per dataset per metric (lowest MAE, bias closest to 0, VR closest to 1). "expl." = numerical explosion.

Dataset	Model	Quartile MAE				Quartile Bias				VR Q4
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	
Simple 3-Node	CausalFlow-T	0.533±0.002	0.127±0.001	0.151±0.005	0.553±0.001	0.433±0.002	0.110±0.001	−0.148±0.001	**−0.353±0.001**	**0.991±0.001**
	NF (no DAG)	0.499±0.000	**0.093±0.000**	0.182±0.000	0.586±0.000	0.499±0.000	0.093±0.000	−0.182±0.000	−0.586±0.000	**0.991±0.000**
	GNN-CVAE	**0.425±0.000**	0.110±0.000	0.248±0.000	0.659±0.000	**0.425±0.000**	**0.028±0.000**	−0.248±0.000	−0.659±0.000	0.852±0.000
	CVAE	0.514±0.000	0.121±0.000	0.156±0.000	0.566±0.000	0.514±0.000	0.121±0.000	−0.156±0.000	−0.566±0.000	0.763±0.000
	TARNet	0.549±0.000	0.147±0.000	**0.133±0.000**	**0.546±0.000**	0.549±0.000	0.147±0.000	**−0.133±0.000**	−0.546±0.000	0.810±0.000
LDL	CausalFlow-T	2.029±0.201	2.326±0.289	2.393±0.449	3.564±0.626	0.547±0.639	**0.677±0.569**	0.781±0.657	**0.963±1.020**	**1.049±0.033**
	NF (no DAG)	2.057±0.320	2.449±0.427	2.752±0.442	3.769±0.453	−0.402±0.356	−0.713±0.734	−0.773±0.641	−1.004±0.650	1.098±0.025
	GNN-CVAE†	23.056±1.133	13.111±0.833	7.563±0.658	1.729±0.565	23.035±1.170	13.071±0.854	7.516±0.653	1.677±0.568	0.846±0.033
	CVAE†	7.005±0.554	1.540±0.242	1.477±0.418	4.216±0.499	6.809±0.543	1.296±0.311	−1.477±0.418	−4.212±0.494	0.698±0.026
	TARNet	**0.407±0.117**	**1.072±0.051**	**0.868±0.062**	**1.142±0.069**	**0.212±0.109**	0.879±0.048	**0.723±0.061**	1.142±0.069	0.611±0.003
Cox Survival	CausalFlow-T	0.025±0.008	0.025±0.003	0.027±0.003	0.034±0.008	**−0.005±0.018**	−0.010±0.008	−0.011±0.007	−0.021±0.013	1.528±0.016
	NF (no DAG)	**0.022±0.005**	0.024±0.006	0.028±0.009	0.038±0.009	−0.011±0.011	−0.012±0.013	−0.016±0.015	−0.029±0.012	1.522±0.009
	GNN-CVAE†	0.026±0.012	0.020±0.007	0.015±0.005	**0.018±0.002**	−0.014±0.017	−0.006±0.012	**−0.002±0.007**	**0.000±0.003**	**0.916±0.035**
	CVAE†	0.023±0.001	**0.014±0.007**	**0.014±0.007**	0.020±0.006	0.011±0.018	**0.002±0.011**	**−0.002±0.008**	−0.002±0.006	0.849±0.038
	TARNet	0.047±0.003	0.047±0.002	0.042±0.004	0.029±0.003	0.047±0.003	0.047±0.002	0.042±0.004	0.029±0.003	0.855±0.058
CVD Risk Toy	CausalFlow-T	**0.010±0.005**	**0.007±0.004**	**0.004±0.002**	**0.003±0.001**	**−0.007±0.008**	**−0.005±0.005**	−0.003±0.003	**0.000±0.002**	1.036±0.001
	NF (no DAG)	0.023±0.010	0.009±0.004	0.006±0.003	0.557±1.750	0.021±0.011	0.007±0.007	0.003±0.007	0.555±1.751	expl.
	GNN-CVAE	0.027±0.001	0.011±0.000	0.008±0.000	0.004±0.000	0.027±0.001	0.011±0.000	0.008±0.000	0.004±0.000	expl.
	CVAE	0.027±0.000	0.012±0.000	0.008±0.000	0.004±0.000	0.027±0.000	0.012±0.000	0.008±0.000	0.004±0.000	expl.
	TARNet	0.080±0.011	0.022±0.011	0.011±0.007	0.013±0.010	0.080±0.011	0.021±0.012	**0.001±0.013**	−0.007±0.014	**1.032±0.051**

† GNN-CVAE and CVAE evaluated on the full LDL / full Cox simulation variant.

Table 9: Arm reconstruction errors and ATE/HR recovery. **Bold**: best per dataset per metric. † GNN-CVAE and CVAE on LDL/Cox use the full simulation variant.

Dataset	Model	$\mathrm{Err}_{a=1}$	$\mathrm{Err}_{a=0}$	ATE or HR (true / pred)
Simple 3-Node	CausalFlow-T	**0.025±0.004**	**0.019±0.003**	−0.942 / −0.950±0.001
	NF (no DAG)	0.028±0.000	0.022±0.000	−0.942 / −0.984±0.000
	GNN-CVAE	0.203±0.000	0.239±0.000	−0.942 / −1.055±0.000
	CVAE	0.200±0.000	0.204±0.000	−0.942 / −0.964±0.000
	TARNet	0.081±0.000	0.064±0.000	**−0.942 / −0.938±0.000**
LDL	CausalFlow-T	0.963±0.119	1.549±0.240	−28.351 / −27.609±0.579
	NF (no DAG)	1.035±0.154	1.697±0.218	**−28.351 / −28.874±0.459**
	GNN-CVAE†	4.854±0.451	6.671±0.773	−28.351 / −17.027±0.789
	CVAE†	0.535±0.245	**0.786±0.372**	−28.351 / −27.747±0.337
	TARNet	**0.267±0.044**	1.010±0.042	−28.351 / −27.613±0.053
Cox Survival	CausalFlow-T	**0.014±0.002**	**0.013±0.002**	0.887 / 0.866±0.011
	NF (no DAG)	0.015±0.005	**0.012±0.002**	0.887 / 0.857±0.021
	GNN-CVAE†	0.076±0.016	0.071±0.009	0.887 / 0.865±0.020
	CVAE†	0.072±0.017	0.074±0.007	**0.887 / 0.879±0.022**
	TARNet	0.040±0.002	0.015±0.001	0.887 / 0.966±0.005
CVD Risk Toy	CausalFlow-T	**0.003±0.001**	**0.002±0.001**	**0.831 / 0.786±0.051**
	NF (no DAG)	0.006±0.002	0.143±0.438	0.831 / 0.834±0.324
	GNN-CVAE	0.053±0.000	0.065±0.000	0.831 / 1.006±0.030
	CVAE	0.052±0.000	0.065±0.000	0.831 / 1.005±0.007
	TARNet	0.040±0.005	0.017±0.008	0.831 / 1.133±0.148

flow setting, where the DAG constraint yields consistent improvement. Both CVAE variants rely on variational inference; in our implementations, the ELBO approximation gap may limit the benefit of improved structural inductive bias. The GNN encoder does not translate into better counterfactual estimates and introduces additional optimization complexity, evidenced by GNN-CVAE's substantially higher LDL arm-1 error (4.854 ± 0.451) relative to plain CVAE (0.535 ± 0.245). Normalizing flows perform exact likelihood-based inference, so architectural improvements such as DAG factorization can directly improve abduction quality. This is consistent with the pattern CVAE $>$ GNN-CVAE, whereas CausalFlow-T $>$ NF (no DAG).

PEHE versus reliability criteria. Table 10 reports PEHE on the positive control and the two primary diagnostic benchmarks. On Simple 3-Node, all models tie (PEHE ≈ 0.43). On LDL, TARNet and CVAE have the best PEHE (6.23 and 6.51 vs. 6.81 for CausalFlow-T), yet CVAE shows a bias ratio of ≈ 1.0. Both TARNet and CVAE show also a tail-variance collapse ($\mathrm{VR}_{Q4} = 0.611$ and 0.698): the best aggregate scores coexist with systematic miscalibration. On CVD Risk, CVAE and GNN-CVAE have PEHE close to CausalFlow-T (0.019 and 0.022 vs. 0.015) despite estimating a null effect (HR ≈ 1.005–1.006 vs. true 0.831; Table 9), and TARNet's fivefold higher PEHE (0.075) does not reveal that it inverts the effect (HR $= 1.133$). A small absolute PEHE can thus coexist with a vanished or reversed effect. PEHE alone therefore masks the designed failure modes that the subgroup and HR criteria detect.

Table 10: PEHE (mean over seeds) on the positive control and the two primary diagnostic benchmarks. Lower is better.

Model	Simple 3-Node	LDL Toy	CVD Risk Toy
CausalFlow-T (ours)	≈ 0.43	6.81	$\mathbf{0.015}$
NF (no DAG)	≈ 0.43	7.05	expl.
GNN-CVAE	≈ 0.43	$14.25^{\dagger}$	0.022
CVAE	≈ 0.43	$6.51^{\dagger}$	0.019
TARNet	≈ 0.43	$\mathbf{6.23}$	0.075

†Full LDL simulation variant, as in Tables 8–9.

H Full Imputation Results

Tables 11 and 12 report complete per-method, per-missingness-level results underlying Table 2. Figure 7 summarizes the normalized proxy-score trajectories during the selected GPT-5.4 evolutionary search, showing rapid early improvement followed by smaller accepted refinements.

To separate the effect of the evolutionary loop from a single executable LLM proposal, the tables also report "GPT-5.4 first-valid", defined as the first GPT-5.4 candidate that passed static and runtime checks and was evaluated. This row is diagnostic only and is not included in the pooled ranks used for model selection. The first-valid rows show that executable LLM proposals are not sufficient on their own, especially at 50%–80% missingness, where the final selected imputer substantially improves biomarker reconstruction and ATE recovery.

Biomarker-level metrics are computed once over all imputed cells across the ten semi-synthetic biomarkers and are therefore reported without seed variability; downstream causal metrics are averaged over multiple training seeds.

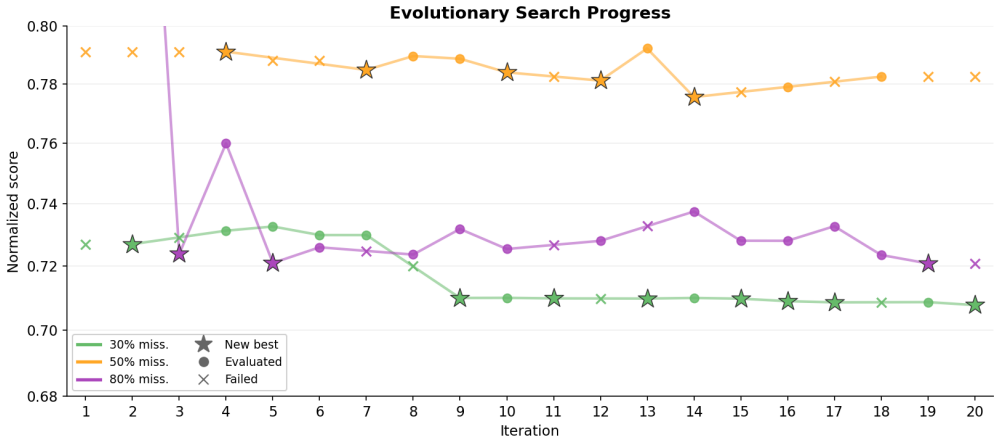


Figure 7: Evolutionary search progress for the LLM-driven imputer across 30%, 50%, and 80% MNAR missingness, shown as normalized proxy scores over the 20-iteration budget.

Table 13 compares ATE recovery of the proposed pipeline with a marginal structural model (MSM). Under GPT-5.4 imputation, CausalFlow-T achieves a lower ensemble ATE residual at 30% missingness (0.003 vs. 0.219). At 50% and 80%, the gaps (0.027 and 0.072) are small relative to CausalFlow-T's across-seed variability (Table 12), so we do not interpret them as clear evidence of superiority. GPT-5.4 imputation also modestly reduces the MSM residual relative to LOCF at 50% and 80%, suggesting benefits beyond the proposed causal estimator. MSM is deterministic conditional on the completed dataset; its patient-sampling uncertainty is distinct from the training-seed variability reported for CausalFlow-T.

I DAG Sensitivity Analysis

We evaluate CausalFlow-T under two partial-graph misspecifications on the FIRE semi-synthetic dataset (true ATE $= -3.484$; multiple seeds). Three baselines serve as reference points: NF (no DAG) shares CausalFlow-T's exact normalizing flow inference but uses no causal structure, isolating the contribution of graph specification within the exact-inference family; CVAE and TARNet use neither exact inference nor graph structure and are fully DAG-invariant, producing identical results across all graph variants by construction. All covariate-to-outcome edges are preserved across variants, isolating confounding path misspecification from outcome model misspecification.

We evaluate two misspecification types that expose complementary robustness properties: (i) a unique-pathway removal (Asthma), where the removed covariate has no correlated substitute in the adjustment set, producing maximal bias ratio deterioration; and (ii) a redundant-pathway removal (Hypertension), where correlated covariates partially absorb the misspecification, producing instability rather than systematic bias as the primary signal. Results can be found in Table 14.

Table 11: Biomarker-level imputation quality across missingness levels, averaged over the semi-synthetic biomarkers. Pointwise MAE and RMSE are computed at missing positions only in original biomarker units. Lag-1 autocorrelation AC is compared against the ground-truth value $\mathrm{AC}^* = 0.375$; CSE denotes consecutive-step error; and inf_Y is biomarker–outcome correlation inflation (ground truth: 0)."GPT-5.4 first-valid" denotes the first executable GPT-5.4 proposal evaluated during the evolutionary search.

Missing	**Method**	MAE ↓	RMSE ↓	AC (→ 0.375)	CSE ↓	inf_Y (→ 0)
30%	LOCF	3.314	4.399	0.473	2.781	−0.113
	MissForest	2.090	2.674	0.384	2.073	+0.036
	CausalCFM	2.755	3.523	0.350	2.708	−0.007
	GPT-5.4	**1.914**	**2.461**	0.392	**1.893**	+0.034
	Qwen3.5-Plus	2.055	2.659	**0.375**	1.956	+0.003
	GPT-OSS-120b	1.972	2.548	0.386	1.914	+0.026
	GPT-5.4 first-valid	2.026	2.598	0.389	2.009	+0.036
50%	LOCF	3.378	4.463	0.554	2.773	−0.197
	MissForest	2.219	2.839	**0.390**	2.185	+0.059
	CausalCFM	2.833	3.623	0.334	2.758	−0.009
	GPT-5.4	**2.052**	**2.628**	0.413	**2.003**	+0.052
	Qwen3.5-Plus	2.461	3.163	0.314	2.349	−0.013
	GPT-OSS-120b	2.059	2.646	0.423	2.011	+0.054
	GPT-5.4 first-valid	2.805	3.741	0.619	2.343	−0.138
80%	LOCF	3.407	4.509	0.713	2.760	−0.347
	MissForest	2.504	3.187	**0.380**	2.468	+0.072
	CausalCFM	2.875	3.679	0.348	2.798	−0.011
	GPT-5.4	**2.059**	**2.634**	0.596	**2.035**	+0.159
	Qwen3.5-Plus	2.788	3.645	0.208	2.612	−0.080
	GPT-OSS-120b	2.147	2.745	0.514	2.109	+0.129
	GPT-5.4 first-valid	3.146	4.193	0.729	2.442	−0.318

The two misspecification types evaluated here are structurally motivated to expose complementary failure modes. Peripheral misspecification (unique pathway removal) represents the worst case for bias: the removed covariate has no correlated substitute, so confounding separation degrades maximally. Central misspecification (redundant pathway removal) represents the worst case for stability: correlated covariates partially absorb the missing path, producing cross-seed instability rather than systematic bias as the primary signal. Together these two cases bound the space of single-edge removals along the bias-instability tradeoff axis. While multi-edge removals, edge direction errors, and latent confounders represent additional misspecification regimes not evaluated here, the instability diagnostic identified in the central case (detectable through cross-seed variance without access to ground-truth counterfactuals) generalizes as a practical signal across misspecification types, since any structural misspecification that disrupts the adjustment set will manifest as estimation instability across random initializations.

Interpretation. *A partially correct DAG is better than no DAG.* Comparing CausalFlow-T against NF (no DAG) within the exact-inference family isolates the contribution of causal structure from inference quality. Under central misspecification, CausalFlow-T's bias ratio (0.517) still improves on NF (no DAG)'s DAG-invariant 0.883, demonstrating that even an incomplete causal graph provides meaningful confounding separation beyond what exact inference alone achieves.

Distributional reliability is preserved under misspecification. CausalFlow-T maintains near-perfect variance calibration across both removals (VRQ4 = 0.996 and 1.005), while CVAE (0.482) and

Table 12: Downstream causal inference quality. All methods use the same fixed CausalFlow-T estimator (Section 4.2); only the imputation strategy varies. Q1 MAE and Q4 MAE (lowest- and highest-ITE quartiles), mean per-arm reconstruction error $\bar{\mathrm{Err}}_a = \frac{1}{2}(\mathrm{Err}_{a=0} + \mathrm{Err}_{a=1})$, tail variance ratio $\mathrm{VR}_{\mathrm{Q4}}$ (target value 1), and absolute ATE residual against $\mathrm{ATE}^* = -3.484$. Means $\pm$ standard deviations across multiple seeds; $|\Delta\mathrm{ATE}|$ is computed from the mean ATE across seeds, with $\pm$ the across-seed SD of the per-seed ATE estimates. † Oracle reference: CausalFlow-T on fully observed data, $\bar{\mathrm{Err}}_a = 0.075$ (Table 7). "GPT-5.4 first-valid" denotes the first executable GPT-5.4 proposal evaluated during the evolutionary search.

Missing	Method	Q1 MAE ↓	Q4 MAE ↓	$\bar{\mathrm{Err}}_a$ ↓	$\mathrm{VR}_{\mathrm{Q4}}$	$\lvert\Delta\mathrm{ATE}\rvert$ ↓
30%	LOCF	0.224 ± 0.123	0.153 ± 0.022	0.079 ± 0.023	1.005 ± 0.006	0.090 ± 0.113
	MissForest	0.416 ± 0.137	0.401 ± 0.144	0.177 ± 0.093	0.998 ± 0.003	0.331 ± 0.223
	CausalCFM	$\mathbf{0.161 \pm 0.073}^{\dagger}$	$\mathbf{0.118 \pm 0.015}$	$\mathbf{0.059 \pm 0.006}$	1.008 ± 0.005	0.027 ± 0.066
	GPT-5.4	0.223 ± 0.120	0.159 ± 0.036	0.071 ± 0.016	1.007 ± 0.011	0.003 ± 0.118
	Qwen3.5-Plus	0.201 ± 0.107	0.220 ± 0.054	0.086 ± 0.035	1.004 ± 0.004	0.145 ± 0.092
	GPT-OSS-120b	0.180 ± 0.040	0.181 ± 0.034	0.065 ± 0.010	1.005 ± 0.007	0.075 ± 0.053
	GPT-5.4 first-valid	0.363 ± 0.239	0.556 ± 0.159	0.233 ± 0.073	1.008 ± 0.009	0.464 ± 0.199
50%	LOCF	0.224 ± 0.074	0.146 ± 0.049	0.084 ± 0.031	1.008 ± 0.008	0.007 ± 0.174
	MissForest	0.305 ± 0.143	0.342 ± 0.192	0.155 ± 0.084	0.999 ± 0.006	0.184 ± 0.306
	CausalCFM	0.232 ± 0.105	0.206 ± 0.105	0.096 ± 0.048	1.007 ± 0.006	0.048 ± 0.198
	GPT-5.4	0.219 ± 0.067	0.199 ± 0.043	0.083 ± 0.029	1.012 ± 0.007	0.178 ± 0.151
	Qwen3.5-Plus	$\mathbf{0.187 \pm 0.099}$	$\mathbf{0.132 \pm 0.034}$	$\mathbf{0.069 \pm 0.022}$	1.008 ± 0.006	0.013 ± 0.114
	GPT-OSS-120b	0.194 ± 0.095	0.283 ± 0.120	0.116 ± 0.056	1.010 ± 0.006	0.199 ± 0.149
	GPT-5.4 first-valid	0.263 ± 0.146	0.285 ± 0.191	0.129 ± 0.063	1.005 ± 0.008	0.216 ± 0.201
80%	LOCF	0.235 ± 0.045	$\mathbf{0.176 \pm 0.050}$	$\mathbf{0.088 \pm 0.031}$	1.009 ± 0.005	0.052 ± 0.163
	MissForest	$\mathbf{0.175 \pm 0.044}$	0.253 ± 0.169	0.102 ± 0.070	1.001 ± 0.002	0.145 ± 0.188
	CausalCFM	0.548 ± 0.358	0.440 ± 0.244	0.226 ± 0.132	1.037 ± 0.024	0.408 ± 0.332
	GPT-5.4	0.311 ± 0.133	0.183 ± 0.086	0.093 ± 0.050	1.016 ± 0.005	0.124 ± 0.090
	Qwen3.5-Plus	0.370 ± 0.220	0.219 ± 0.080	0.133 ± 0.063	1.032 ± 0.007	0.190 ± 0.185
	GPT-OSS-120b	0.222 ± 0.105	0.211 ± 0.153	$\mathbf{0.088 \pm 0.044}$	1.019 ± 0.008	0.072 ± 0.172
	GPT-5.4 first-valid	0.283 ± 0.150	0.327 ± 0.182	0.140 ± 0.064	1.010 ± 0.006	0.201 ± 0.205

Table 13: ATE recovery on the semi-synthetic benchmark ($\mathrm{ATE}^* = -3.484$). MSM denotes IPTW/IPCW with weighted GEE; IPCW weights are unity on this benchmark because there is no dropout. CausalFlow-T residuals are computed from the ensemble mean across multiple training seeds; the corresponding seed standard deviations are reported in Table 12. MSM is deterministic conditional on the completed dataset.

	$\lvert\Delta\mathrm{ATE}\rvert$ ↓		
Missing	MSM + LOCF	MSM + GPT-5.4	CausalFlow-T + GPT-5.4 ensemble
30%	0.214	0.219	0.003
50%	0.213	0.205	0.178
80%	0.213	0.196	0.124

TARNet (0.608) collapse regardless of graph quality. NF (no DAG) also preserves VRQ4 (1.016) but at the cost of a consistently high bias ratio (0.883), confirming that variance calibration and confounding separation are non-redundant properties requiring both exact inference and causal structure jointly.

Misspecification is detectable through instability. The central removal produces a $4.5\times$ increase in CausalFlow-T's cross-seed standard deviation relative to the full DAG (0.504 vs. 0.112). DAG-invariant models show no such signal (TARNet's seed SD remains 0.015 and NF (no DAG)'s 0.251 across all conditions) meaning they fail silently under the same misspecification that CausalFlow-T flags through detectable instability. This diagnostic property is practically valuable: a practitioner monitoring cross-seed variance can detect central misspecification without access to ground-truth counterfactuals.

Table 14: Sensitivity to DAG misspecification on FIRE semi-synthetic (true ATE $=-3.484$, multiple seeds). NF (no DAG), CVAE, and TARNet are DAG-invariant and produce identical results across all graph variants by construction (†). VRQ4 $\to 1$; |Bias|/MAE < 0.5 indicates predominantly random errors.

DAG variant	Model	$\lvert\text{Bias}\rvert/\text{MAE}_{\text{Q1}} \downarrow$	$\text{VR}_{\text{Q4}} \to 1$	ATE	Seed SD
Full DAG	CausalFlow-T	**0.316**	**1.006**	-3.568	0.112
	NF (no DAG)†	0.883	1.016	-3.382	0.251
	CVAE†	0.938	0.482	-3.451	0.075
	TARNet†	0.708	0.608	-3.415	0.015
Peripheral (Asthma removed)	CausalFlow-T	0.955	**0.996**	-3.852	0.145
	NF (no DAG)†	0.883	1.016	-3.382	0.251
	CVAE†	0.938	0.482	-3.451	0.075
	TARNet†	**0.708**	0.608	-3.415	0.015
Central (HTN removed)	CausalFlow-T	**0.517**	**1.005**	-3.896	0.504
	NF (no DAG)†	0.883	1.016	-3.382	0.251
	CVAE†	0.938	0.482	-3.451	0.075
	TARNet†	0.708	0.608	-3.415	0.015

Bias is directionally consistent. Both misspecifications shift the estimate away from zero, overstating the protective effect (peripheral: -3.852; central: -3.896; true: -3.484). The direction is consistent across both removals in this benchmark.

J Sensitivity Analysis for the LLM Imputer

J.1 Search-configuration sensitivity

We vary the iteration budget K, the proxy weights (λ_Y, λ_T), and the seed imputer $g^{(0)}$ one at a time (all other settings as in Table 6) and report the proxy score of the selected imputer together with the downstream $|\Delta\text{ATE}|$ (Table 15). The search converges well within the default budget: the selected imputer no longer changes after $K{=}10$ at 30% and $K{=}15$ at 50% and 80%. Across all settings, downstream $|\Delta\text{ATE}|$ shows no significant differences (one-way ANOVA, $p \geq 0.12$) and varies within seed-level variability. Setting $\lambda_Y{=}0$ lowers the proxy only because it removes the Δ_Y penalty, without downstream gain; as this test has limited power, we keep $\lambda_Y > 0$ by design, since Δ_Y targets the correlation structure CausalFlow-T relies on.

J.2 Bootstrap uncertainty analysis

Table 16 compares uncertainty analyses of the real-world per-protocol effect with the SUSTAIN 8 estimate. The conditional bootstrap resamples patients while holding the imputed dataset and the fitted CausalFlow-T fixed, and therefore captures between-patient sampling variation only. The bootstrap-then-impute analysis instead resamples the incomplete panel, re-imputes each resample with the selected imputer $g^\star$, and re-fits CausalFlow-T end to end over multiple resamples, so that imputation and model-fitting uncertainty propagate into the interval; we report it in the main text. Its interval is about $11\times$ wider than the conditional one, indicating that imputation, not sampling, dominates uncertainty, and the point estimate shifts slightly ($-0.98 \to -0.93$) as resampling perturbs cohort composition under missingness. The propagated estimate agrees with the implemented IPTW–IPCW marginal structural model, and all intervals lie within the SUSTAIN 8 interval.

Table 17 reports bootstrap sensitivity intervals for the IPTW (semi-synthetic) and IPTW–IPCW (real cohort) estimates obtained after applying the selected LLM imputer. After imputation, patients were resampled with replacement. For each bootstrap replicate, the weighting models were re-estimated and the ATE was recomputed. We used 500 bootstrap replicates and percentile-based 95% confidence intervals, applied to the completed imputed datasets; these intervals therefore quantify sampling and downstream weighting uncertainty conditional on the completed imputed dataset.

Table 15: Search-configuration sensitivity of the GPT-5.4 imputer (semi-synthetic benchmark). Proxy: $s(g^\star)$ at 30/50/80% (Eq. 4). $|\Delta\text{ATE}|$: computed per seed and then averaged (mean $\pm$ SD over multiple seeds). Panel (a) pairs logged proxy scores from the original search with downstream errors from rerun searches; panels (b) and (c) are fully rerun. Independent runs can yield different default results across panels and are not comparable with the seed-ensemble estimates of Table 12.

Configuration	Proxy (30/50/80)	$\lvert\Delta\text{ATE}\rvert$ 30%	$\lvert\Delta\text{ATE}\rvert$ 50%	$\lvert\Delta\text{ATE}\rvert$ 80%
(a) Iteration budget K				
K=5	2.83/3.02/3.22	0.105 ± 0.079	0.361 ± 0.204	0.239 ± 0.125
K=10	2.80/2.99/3.21	0.140 ± 0.102	0.219 ± 0.093	0.192 ± 0.136
K=15	2.80/2.96/3.20	0.140 ± 0.102	0.298 ± 0.181	0.204 ± 0.143
K=20 (default)	2.80/2.96/3.20	0.140 ± 0.102	0.298 ± 0.181	0.204 ± 0.143
(b) Proxy weights (λ_Y, λ_T)				
$(2.0, 0.5)$ (default)	3.54/2.98/3.16	0.226 ± 0.185	0.288 ± 0.131	0.298 ± 0.137
$(0.0, 0.5)$	2.89/2.87/2.75	0.236 ± 0.098	0.264 ± 0.162	0.254 ± 0.105
$(4.0, 0.5)$	2.97/3.75/3.23	0.254 ± 0.169	0.271 ± 0.172	0.248 ± 0.206
$(2.0, 0.0)$	2.79/3.08/3.13	0.175 ± 0.108	0.286 ± 0.177	0.273 ± 0.209
$(2.0, 1.0)$	2.82/3.07/3.20	0.220 ± 0.099	0.153 ± 0.136	0.233 ± 0.180
(c) Seed imputer $g^{(0)}$				
Deterministic (default)	3.54/2.98/3.16	0.140 ± 0.102	0.298 ± 0.181	0.204 ± 0.143
MissForest	2.78/2.94/3.19	0.229 ± 0.132	0.249 ± 0.140	0.185 ± 0.099
Column mean	2.88/3.18/3.20	0.285 ± 0.184	0.149 ± 0.137	0.266 ± 0.130

Table 16: Real-world per-protocol effect of GLP-1RA vs. SGLT-2i initiation on 1-year body weight (kg) under different uncertainty analyses, compared with the SUSTAIN 8 RCT.

Analysis	ATE (kg)	95% CI
Conditional bootstrap (imputation and model fixed)	-0.98	$[-1.01, -0.96]$
Bootstrap-then-impute (propagated)	-0.93	$[-1.26, -0.70]$
IPTW–IPCW MSM (Table 17)	-0.94	$[-1.34, -0.55]$
SUSTAIN 8 RCT [Lingvay et al., 2019]	-1.06	$[-1.76, -0.36]$

K Theoretical Aspects of CausalFlow–T

K.1 Exact Likelihood vs. ELBO: When the Gap Matters

CVAEs maximize $\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi}[\log p_\theta(v \mid z)] - \text{KL}[q_\phi(z \mid v) \,\|\, p(z)] \leq \log p_\theta(v)$. The gap $\log p_\theta(v) - \mathcal{L}_{\text{ELBO}} \geq 0$ is the variational approximation error, and it equals $\text{KL}[q_\phi(z \mid v) \,\|\, p_\theta(z \mid v)]$. It therefore vanishes when the variational family contains the true posterior, and variational estimators can be consistent under suitable conditions [Wang and Blei, 2019]. In practice, its size depends on the variational family, the encoder, and training. Posterior collapse, in which $q_\phi(z \mid v) \approx p(z)$ and the decoder ignores z, is a known failure mode of variational models, and the null effects observed in our benchmarks are consistent with it.

Normalizing flows compute the exact log-likelihood via the change-of-variables formula: $\log p_\theta(v_t \mid h_t) = \log p_z(f_\theta(v_t; h_t)) + \log |\det \partial f_\theta(v_t; h_t)/\partial v_t|$. For MAF-type flows the Jacobian is triangular and computable in $\mathcal{O}(p)$. The AAP procedure is therefore: (1) *Abduction:* $z_t = f_\theta(v_t; h_t)$ (exact and deterministic for the fitted model), (2) *Action:* $a_t \leftarrow a'$, (3) *Prediction:* decode all descendants of A in $\mathcal{G}$, holding non-descendants fixed. Because inversion is exact, the same z_t is used under both arms; this couples a patient's two predicted potential outcomes.

Interpretation of our baseline results. On CVD Risk, both CVAE-based baselines we implemented estimated HR ≈ 1.005 (true 0.831), consistent with posterior collapse (Section 4.2). We report this as an empirical outcome for these implementations and hyperparameters (Table 5), not as evidence that variational methods cannot recover mediated effects.

Table 17: Bootstrap sensitivity intervals for selected LLM-imputed primary endpoint estimates. These use a weighted difference-in-means estimator and therefore differ from the weighted-GEE MSM of Table 13. IPTW denotes inverse probability of treatment weighting; IPCW denotes inverse probability of censoring weighting.

Miss.	Estimator	Point ATE	95% CI
30%	IPTW	-3.230	$[-3.354, -3.101]$
50%	IPTW	-3.228	$[-3.354, -3.095]$
80%	IPTW	-3.172	$[-3.323, -3.011]$
Real	IPTW–IPCW	-0.936	$[-1.335, -0.553]$

K.2 The Role of the DAG in Counterfactual Propagation

For an unconstrained autoregressive flow with ordering (A, X, Y), replacing $A \leftarrow a'$ causes the decoder to adjust X along the learned autoregressive path, traversing the $A \to X$ direction which does not exist in the causal graph. This anti-causal propagation produces systematically biased counterfactuals even when the factual distribution is perfectly fitted. CausalFlow-T fixes the autoregressive ordering to a topological sort of $\mathcal{G}$, so the intervention propagates only through causal descendants, implementing do-calculus exactly.

K.3 Limitations of Continuous Relaxation for Binary Outcomes

Normalizing flows require dequantization of binary targets, effectively relaxing the discrete problem into a continuous approximation. This enables exact likelihood-based inference but introduces a modeling approximation that can affect calibration for survival outcomes. On the Cox benchmark, this explains why variational approaches can match or exceed flows in MAE and $\mathrm{VR}_{\mathrm{Q4}}$, while CausalFlow-T recovers an HR close to the true value and maintains an advantage in arm reconstruction (the structural benefit persists at a calibration cost). Future work on discrete or hybrid flow-based models for survival outcomes will address this limitation.

L Theoretical Aspects of LLM-driven Evolutionary Imputation

L.1 Search Algorithm

The search maintains a single current-best candidate and accepts a new proposal only if it strictly improves the composite score $s(g)$ defined in Eq. (4). Starting from a deterministic seed imputer $g^{(0)}$, Algorithm 1 repeats for K iterations. The same proxy holdout Ω_p is used at every iteration so that all candidates are scored on identical hidden cells.

Real-world proxy holdout. For the semi-synthetic missingness benchmarks, the proxy holdout Ω_p is sampled independently at the observed-cell level. In the real FIRE application, the imputation target set includes all incomplete analysis variables, including the outcome. Because raw clinical measurements are expanded onto a 30-day longitudinal target grid, the same observed measurement can appear across adjacent rows; a cellwise proxy holdout would therefore allow trivial reconstruction from neighboring rows via forward/backward copying, artificially favoring LOCF-like candidates. We therefore sample Ω_p at the level of contiguous within-patient observed-value runs: for each patient and target variable, consecutive rows with the same observed value are treated as one segment, and selected segments are masked in full with probability ρ. The selected segment-level holdout is fixed across candidate evaluations.

Single-parent scheme. We deliberately avoid population-based search. The evaluation cost is dominated by the proxy fit on $\widetilde{\mathcal{D}}_-$, which is cheap relative to the cost of an LLM call. A single-parent scheme keeps the prompt focused on a single, well-characterized current best, which we found empirically to converge faster than larger populations under a fixed call budget. The strict-improvement rule is equivalent to a deterministic acceptance step in a single-parent evolution strategy without recombination.

Algorithm 1 LLM-driven evolutionary imputation search (high-level view).

Require: Incomplete datasets $\{\widetilde{\mathcal{D}}^{(r)}\}_{r\in\{30,50,80\}}$ (missingness rate $r\%$); seed imputer $g^{(0)}$; holdout fraction ρ; correlation weights $\lambda_Y, \lambda_T \geq 0$; budget K; history window W.

Ensure: One imputer $g^{\star(r)}$ per missing-rate dataset.

1: **for** $r \in \{30, 50, 80\}$ **do** ▷ three independent runs

2: Sample proxy holdout $\Omega_p^{(r)}$ by masking each *observed* target cell of $\widetilde{\mathcal{D}}^{(r)}$ independently with probability ρ (fixed seed), forming $\widetilde{\mathcal{D}}_-^{(r)}$

3: $g^{\star(r)} \leftarrow g^{(0)}$; $s^{\star(r)} \leftarrow \textsc{Score}(g^{(0)}, r)$; $\mathcal{H} \leftarrow \emptyset$

4: **for** $k = 1, \ldots, K$ **do**

5: $g^{(k)} \leftarrow \mathrm{LLM}\big(\textsc{Prompt}\big(g^{\star(r)},\, s^{\star(r)},\, \mathcal{H}_{[-W:]}\big)\big)$

6: **if** $g^{(k)}$ fails static check or runtime guard **then**

7: $\mathcal{H} \leftarrow \mathcal{H} \cup \{(k,\, g^{(k)},\, \texttt{failed})\}$; **continue**

8: **end if**

9: $s_k \leftarrow \textsc{Score}(g^{(k)}, r)$

10: **if** $s_k < s^{\star(r)}$ **then** ▷ minimize scalar s

11: $g^{\star(r)} \leftarrow g^{(k)}$; $s^{\star(r)} \leftarrow s_k$ ▷ accept

12: **end if**

13: $\mathcal{H} \leftarrow \mathcal{H} \cup \{(k,\, g^{(k)},\, s_k)\}$

14: **end for**

15: **end for**

16: **return** $\{g^{\star(r)}\}_{r\in\{30,50,80\}}$

17: **function** $\textsc{Score}(g,\, r)$

18: From $g(\widetilde{\mathcal{D}}_-^{(r)})$, compute $\mathrm{RMSE}(g)$ on $\Omega_p^{(r)}$ and $\Delta_Y(g)$, $\Delta_T(g)$ (Eq. (4)) ▷ plus $\mathrm{MAE}(g)$ as diagnostic for $\mathcal{H}$

19: **return** $s \leftarrow \mathrm{RMSE}(g) + \lambda_Y\, \Delta_Y(g) + \lambda_T\, \Delta_T(g)$

20: **end function**

L.2 Prompting

At iteration k, the prompt assembled in line 5 of Algorithm 1 contains three blocks, in order:

- **Current-best source code.** The full Python module implementing $g_{k-1}^{\star}$, including any helper functions. The LLM is asked to return a complete replacement module rather than local edits; this avoids whole classes of patch-application errors and makes acceptance a binary check on the new module.
- **Proxy summary of $g_{k-1}^{\star}$.** A short text block stating the selection rule (minimize $s = \mathrm{RMSE} + \lambda_Y\, \Delta_Y + \lambda_T\, \Delta_T$) and the current-best values of RMSE and MAE on the proxy holdout and of Δ_Y and Δ_T. Each missingness regime is searched independently, so the summary is regime-specific.
- **Search history $\mathcal{H}$.** A short log of the last W proposals from the same run, summarizing for each one whether it was accepted, rejected without improvement, or failed. Accepted and rejected proposals contribute their proxy values RMSE, Δ_Y, and Δ_T, with MAE included only as diagnostic feedback; failed proposals contribute a short reason (e.g. banned import, timeout, malformed output). This lets the LLM see which directions have already been tried and which kinds of mistakes to avoid in the next proposal.

L.3 Example of an LLM-proposed imputation program

The following excerpt shows the main function of the selected LLM-proposed imputation program, including its calls to helper functions defined in the same module.

```
def impute(masked_df):
    work = prepare_working_frame(masked_df)
    out = initial_target_fill(work)
    for target in TARGET_COLS:
```

```
        obs = work[target].notna()
        X = _build_target_features(work, out, target)
        pred = _weighted_blend([
            _fit_predict_trees(
                X[obs], work[target][obs], X[~obs]),
            _fit_predict_ridge(
                X[obs], work[target][obs], X[~obs])
        ])
        out.loc[~obs, target] = _conservative_calibration(
            pred, work[target][obs])
    return build_result_frame(masked_df, work, out)
```

L.4 Constraints and Failure Handling

Each candidate $g^{(k)}$ is screened in two stages before its proxy score is computed.

Static checks. Before execution, the candidate source is parsed and rejected if it tries to access the file system, the network, or any form of dynamic code execution (e.g. `eval`, `exec`, disk I/O via `pandas`).

Runtime guard. Candidates that pass the static check run in an isolated child process with a per-dataset time and memory budget (Table 6). Any crash, timeout, or attempt to overwrite observed values or non-target columns aborts the candidate; the failure is recorded in $\mathcal{H}$ and the search moves on.

M Assumptions and Identification

In this appendix, $\bar{x}_{i,t} = (x_{i,1}, \ldots, x_{i,t})$ denotes the covariate history, $H_{i,t} = \bar{x}_{i,t-1}$ the history before time t, and $h_{i,t}$ its LSTM encoding (Eq. (3)). Treatment strategies are sustained from baseline, $a' \in \{0, 1\}$, and $\bar{a}_{i,t} = a'$ is shorthand for $a_{i,1} = \cdots = a_{i,t} = a'$.

M.1 Formal statement for Assumptions made

Here we present a detailed description of assumptions (i)–(iv) from Section 3.1:

(i) *Sequential ignorability:* $y_{i,s}(a') \perp\!\!\!\perp a_{i,t} \mid h_{i,t}, x_{i,t}, \bar{a}_{i,t-1} = a'$ for all $s \geq t$. Because the encoder uses covariates only, it implicitly assumes that past outcomes do not confound treatment beyond what the covariate history captures. For the per-protocol analysis (Section 5), we additionally assume exchangeability for censoring due to protocol deviation given the same variables. Section 3.1 states (i) without the conditioning on $\bar{a}_{i,t-1} = a'$, which is implicit there because treatment strategies are sustained from baseline.

(ii) *History sufficiency:* $p(v_{i,t} \mid H_{i,t}) = p(v_{i,t} \mid h_{i,t})$ for all t.

(iii) *Correct DAG:* given $h_{i,t}$, the distribution of $v_{i,t}$ factorizes as $\prod_{k=1}^{p} p(v_{i,t,k} \mid v_{i,t,\mathrm{pa}(k)}, h_{i,t})$ according to $\mathcal{G}$, each conditional remains unchanged when other nodes are intervened on, and there are no unmeasured common causes among the nodes of $v_{i,t}$.

(iv) *Positivity:* $P(a_{i,t} = a' \mid h_{i,t}, x_{i,t}, \bar{a}_{i,t-1} = a') > 0$ wherever the conditioning event has positive density. *Consistency:* if $\bar{a}_{i,t} = a'$, then $y_{i,t} = y_{i,t}(a')$ and $x_{i,t+1} = x_{i,t+1}(a')$, with no interference between patients.

The assumptions play different roles. (i) and (iv) are the standard conditions under which the population effect is identified from observational data; (ii) and (iii) ensure that the CausalFlow-T conditionals, which depend on the history only through $h_{i,t}$ and factorize according to $\mathcal{G}$, can represent the identified quantity. Proposition 1 states the resulting guarantee.

Proposition 1. *Under (i) and (iv),*

$$\mathbb{E}[y_{i,t}(a')] = \int \mathbb{E}\big[y_{i,t} \mid \bar{x}_{i,t}, \bar{a}_{i,t} = a'\big] \prod_{s=1}^{t} p\big(x_{i,s} \mid \bar{x}_{i,s-1}, \bar{a}_{i,s-1} = a'\big)\, d\bar{x}_{i,t}. \tag{7}$$

Under (ii)–(iii), each factor can be written conditionally on $h_{i,s}$, and the within-slice distribution under $\mathrm{do}(a_{i,s} = a')$ is the truncated factorization of $\mathcal{G}$. If the flow conditionals are consistently estimated, the CausalFlow-T plug-in estimator, which sets $a_{i,s} = a'$ and generates the descendants of $a_{i,s}$ in topological order, is consistent for $\mathbb{E}[y_{i,t}(a')]$ and hence for ATE_t.

Derivation. Equation (7) is the sequential g-formula [Robins, 1997, Gill and Robins, 2001, Hernán and Robins, 2020]: by iterated expectations, $\mathbb{E}[y_{i,t}(a')]$ is written as the conditional expectations over $x_{i,1}, \ldots, x_{i,t}$; at each step, sequential ignorability and positivity allow conditioning on $a_{i,s} = a'$ without changing the expectation, and consistency replaces the potential outcome by the observed one in the final step.

Finally, these results are related to the complete data. Since CausalFlow-T is fitted to imputed data and MNAR distributions are not identified without further assumptions [Mohan and Pearl, 2021], the imputation stage is validated empirically (Section 4.3).

M.2 Scope of the history-sufficiency assumption

Assumption (ii) may not be guaranteed by LSTM training: maximizing the likelihood encourages $h_{i,t}$ to be predictive of $v_{i,t}$, but finite capacity, finite samples, and optimization error can leave confounding information unencoded, and LSTM representations are not identifiable in general. Whether a learned representation is a valid adjustment set is a known open issue, studied as representation-induced confounding bias [Melnychuk et al., 2024]. Therefore state (ii) is an assumption. It cannot be verified from observational data alone, and an encoder of the observed history cannot recover unmeasured confounders. Instead, we provide three complementary sources of supporting evidence: (1) on the semi-synthetic benchmark, where ground truth is available, the estimated ATE is close to the true ATE (Table 7); (2) cross-seed instability, computable without ground truth, increases markedly under DAG misspecification (Appendix I), although it diagnoses structural misspecification rather than testing (ii) directly; and (3) on real data, the estimate agrees with a conventional IPTW–IPCW analysis and with RCT evidence (Section 5 and Table 16), within an active-comparator, new-user, per-protocol design.

N Compute Resources

All experiments were run on a single node of an HPC cluster, equipped with 4× NVIDIA L40S GPUs (46 GB VRAM each, 350 W TDP), CUDA 12.8, and driver version 570.172.08. In practice, all experiments ran on a single GPU, with peak memory usage well below 2 GB, confirming that the pipeline is lightweight relative to the available hardware. Synthetic and semi-synthetic benchmark experiments (Sections 4.2–4.3) require on the order of a few minutes per run; the LLM evolutionary imputation search ($K = 20$ iterations) takes approximately 57 minutes per missingness level; and the real-world TTE experiment (Section 5) requires approximately 58 minutes end-to-end. Each fully imputed dataset with GPT-5.4 incurred an estimated API cost of approximately $2–3. All models are implemented in PyTorch. Total compute across all experiments, including multiple seed repetitions through the full pipeline, amounts to approximately 10 GPU-hours on a single L40S.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the two main contributions (CausalFlow-T and LLM-driven evolutionary imputation), the experimental setup validating them, and the real-world application. CausalFlow-T builds on existing components (CausalMAF and an LSTM encoder), which are explicitly explained; the contribution is their integration for longitudinal target trial emulation, the evaluation protocol, and the observed failure modes. Claims are supported by controlled ablations on synthetic and semi-synthetic benchmarks with known counterfactuals and are appropriately scoped. The real-world result is framed as consistency with RCT evidence rather than ground-truth validation.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Section 6 discusses limitations.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification:The identification assumptions (sequential ignorability, history sufficiency, correct DAG specification, positivity, and consistency) are stated in Section 3.1 and formally in Appendix M, together with Proposition 1 (identification via the sequential g-formula and consistency of the plug-in estimator) and a derivation based on standard results [Robins, 1997, Gill and Robins, 2001, Hernán and Robins, 2020]. Theoretical aspects of exact likelihood vs ELBO and the role of the DAG in counterfactual propagation are discussed in Appendix K. Additional data imputation improvements with evolutionary LLM are explained in Section 3.3 and evaluated empirically in Section 4. The paper does not claim novel theoretical results beyond these; the contributions are primarily methodological and mostly empirical.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Full hyperparameter details are provided in Tables 5 and 6 for all models across all datasets. Dataset generation procedures are specified in Appendix A including structural equations, DAG adjacency matrices, MNAR mechanism parameters, and train/test splits. The evolutionary imputer algorithm is given in full in Algorithm 1 with prompting strategy described in Appendix L.2. The real-world FIRE dataset cannot be released due to privacy constraints but the TTE protocol is fully specified in Table 3.

   Guidelines:

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

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code added as supplementary material. The real-world FIRE database cannot be publicly released due to privacy and regulatory constraints. However, all four synthetic datasets are fully specified by their structural equations and can be reconstructed from the descriptions in Appendix A. The anonymized code is released.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Training details including optimizer, learning rate, batch size, data splits, and architecture hyperparameters are reported in Table 5 for all models across synthetic, semi-synthetic, and real settings. Evolutionary imputer hyperparameters including iteration budget, proxy holdout fraction, composite score weights, and timeout constraints are reported in Table 6. Compute resources are detailed in Appendix N.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: All causal inference results are reported as means ± standard deviations over multiple random seeds. Bootstrap 95% confidence intervals for the real-world ATE estimate are reported in Tables 16 and 17. Biomarker-level imputation metrics are computed over all imputed cells and reported without seed variability as they are deterministic given the imputer; this is explicitly noted in Appendix H.

   Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Full compute details are provided in Appendix N. Runtime estimates are given per experiment type as well as the total complete estimation. API costs for LLMs are also detailed.

   Guidelines:

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

9. **Code of ethics**

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

Answer: [Yes]

Justification: The study uses retrospective EHR data from the FIRE database under appropriate data governance. No patient-identifiable information is released. The research aims to improve causal inference from incomplete clinical data, with direct benefit to pharmacoepidemiology in settings where RCTs are infeasible. No deceptive practices, dual-use risks, or privacy violations are involved.

Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes]

Justification: The positive societal impact is improving real-world evidence generation for drug comparative effectiveness when RCTs are infeasible, directly relevant to clinical decision-making for conditions like type 2 diabetes affecting over half a billion people. A potential negative impact is that overconfident deployment of ML-based causal estimates in clinical settings without appropriate uncertainty quantification could mislead treatment decisions; the paper mitigates this by reporting bootstrap confidence intervals and explicitly noting the untestable conditional exchangeability assumption.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The paper does not release pretrained models, scraped datasets, or generative models with high misuse risk. The evolutionary imputer relies on API access to existing LLMs and does not introduce new model weights. The FIRE database is not released.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All baseline methods are properly cited. The FIRE database is cited and described. LLM APIs (GPT-5.4, Qwen3.5-Plus, GPT-OSS-120B) are used under their respective commercial and open-source terms.

Guidelines:

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

13. **New assets**

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

Answer: [Yes]

Justification: The paper introduces the two-stage pipeline, some baselines ML models and the four synthetic benchmark datasets as new assets. Full specifications are provided in Appendices A and C. Code is released. No new datasets involving human subjects are released; the semi-synthetic FIRE dataset uses real demographic structure but synthetic outcomes.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or prospective research with human subjects. The real-world analysis uses retrospective EHR data collected as part of routine clinical care under existing data governance agreements.

    Guidelines:

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

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

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

    Answer: [Yes]

    Justification: The FIRE database is managed by the Institute of Primary Care at the University Hospital Zurich under appropriate ethical and regulatory oversight for secondary use of routine clinical data. Approval has been granted without sharing sensitive data.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [Yes]

    Justification: LLMs (GPT-5.4, Qwen3.5-Plus, GPT-OSS-120B) are a core and original component of the methodology, used as mutation operators in the evolutionary imputation search rather than for writing or editing. Their role, prompting strategy, and constraints are fully described in Section 3.3 and Appendix L.

Guidelines:

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